High-throughput Genetic Clustering of Type 2 Diabetes Loci Reveals Heterogeneous Mechanistic Pathways of Metabolic Disease

Hyunkyung Kim,[1,2,3] Kenneth E. Westerman,[2,4] Kirk Smith,[1,2,3] Joshua Chiou,[5] Joanne B. Cole,[2,3,6,7] Timothy Majarian,[2] Marcin von Grotthuss,[8] Josep M. Mercader,[1,2,3,6] Soo Heon Kwak,[9] Jaegil Kim,[2,10] Jose C. Florez,[1,2,3,6] Kyle Gaulton,[5] Alisa K. Manning[2,4,6] and Miriam S. Udler[1,2,3,6] *

[1]Diabetes Unit, Massachusetts General Hospital, Boston, MA 02114, USA;
[2]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA;
[3]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA;
[4]Clinical and Translational Epidemiology Unit, Massachusetts General Hospital, Boston, MA 02114, USA;
[5]Department of Pediatrics, University of California San Diego, San Diego, CA 92161, USA;
[6]Department of Medicine, Harvard Medical School, Boston, MA 02115, USA;
[7]Division of Endocrinology and Center for Basic and Translational Obesity Research, Boston Children's Hospital, Boston, MA 02115, USA;
[8]Takeda Pharmaceuticals, Cambridge, MA 02139, USA;
[9]Department of Internal Medicine, Seoul National University Hospital, Seoul 03080, Republic of Korea;
[10]Current address: GlaxoSmithKline, Cambridge, MA 02140, USA;

Abstract

Aims/hypothesis

Type 2 diabetes (T2D) is highly polygenic and influenced by multiple biological pathways. Rapid expansion in the number of T2D loci can be leveraged to identify such pathways, thus facilitating improved disease management.

Methods

We developed a high-throughput pipeline to enable clustering of T2D loci based on variant-trait associations. Our pipeline extracted summary statistics from genome-wide association studies (GWAS) for T2D and related traits to generate a matrix of 324 variant x 64 trait associations and applied Bayesian Non-negative Factorization (bNMF) to identify genetic components of T2D. We generated cluster-specific polygenic scores and performed regression analysis in an independent cohort (N=25,419) to assess for clinical relevance.

Results

We identified ten clusters, replicating the five from our prior analysis as well as novel clusters related to beta-cell dysfunction, pronounced insulin secretion, and levels of alkaline phosphatase, lipoprotein-A, and sex hormone-binding globulin. Four clusters related to mechanisms of insulin deficiency, five to insulin resistance, and one had an unclear mechanism. The clusters displayed tissue-specific epigenomic enrichment, notably with the two beta-cell clusters differentially enriched in functional and stressed pancreatic beta-cell states. Additionally, cluster-specific polygenic scores were differentially associated with patient clinical characteristics and outcomes. The pipeline was applied to coronary artery disease and chronic kidney disease, identifying multiple shared genetic pathways with T2D.

Conclusions/interpretation

Our approach stratifies T2D loci into physiologically meaningful genetic clusters associated with distinct tissues and clinical outcomes. The pipeline allows for efficient updating as additional GWAS become available and can be readily applied to other conditions, facilitating clinical translation of GWAS findings. Software to perform this clustering pipeline is freely available.

Introduction

Diabetes is one of the most common complex diseases, afflicting approximately 460 million adults worldwide, with prevalence estimated to more than double by 2045. Type 2 diabetes (T2D) accounts for more than 90% of diabetes [1], with variable contributions of insulin resistance and beta cell dysfunction, and influenced by multiple risk factors, including genetics [2]. Hence, untangling the heterogeneity of T2D may be fundamental to improving patient management and facilitating precision medicine.

Hundreds of loci associated with T2D have been identified in large-scale genetic studies, including genome-wide association studies (GWAS), accounting for approximately 20% of the variance in individual predisposition to T2D [3,4,5]; however, the translation of these established T2D genetic loci into improved understanding of disease pathophysiology and clinical relevance has been challenging, due in large part to the non-coding nature and the small effect sizes of the genetic variants. Recent studies have leveraged a growing number of available GWAS datasets to connect genetic loci to mechanistic pathways by clustering loci based on shared patterns of associations across multiple traits. In our previous work [6], clustering was performed on 94 T2D variants identified by manual curation of published T2D GWAS manuscripts. Soft clustering analysis with Bayesian Non-negative Matrix Factorization (bNMF) of the associations of these 94 T2D variants with 47 diabetes-related traits identified five distinct clusters, recognizable as biological pathways of T2D. A similar set of five clusters of T2D loci were independently identified by Mahajan *et al.*, along with a sixth cluster of “undetermined” physiological impact [3,7]. Of these five shared clusters, two related to beta-cell dysfunction, and the other three clusters represented different mechanisms of insulin resistance: obesity-mediated, abnormal lipodystrophy-like fat distribution, and altered hepatic lipid metabolism [3].

With new T2D loci continuously being discovered and additional GWAS trait summary statistics becoming publicly available, we sought to expand our prior work which involved manual curation of a smaller set of T2D loci. We developed a high-throughput pipeline to enable extraction of hundreds of genetic variants and traits from multiple GWAS to be used for cluster analysis in order to identify new genetic pathways of disease.

Material and methods

Pipeline for input variant-trait association matrix for clustering

An overview of preprocessing steps for variants and traits used for generating the input matrix for variant-trait association clustering analysis is illustrated in **Figure S1**.

Variant selection

To obtain a comprehensive set of independent genetic variants associated with T2D, we gathered T2D GWAS, exome-chip, and whole-genome sequencing summary statistic datasets deposited in the AMP-Common Metabolic Disease Knowledge Portal (CMDKP) [9]. We required the GWAS to have sample sizes larger than 10,000 to reduce false positive associations and focused on studies of predominantly European ancestry to minimize heterogeneity across studies and reduce artifactual clustering results. Thirteen GWAS datasets were included as input for identifying T2D genetic loci for clustering (**Table S1**).

From the selected thirteen T2D GWAS datasets, we first extracted 21,666 variants reaching genome-wide significance ($P < 5\times10^{-8}$) and then ensured variant signals replicated in the largest of these studies at a Bonferroni significance level (P-value<0.05/21,666 tests). Next we identified the set of T2D variants (N=16,074) that were multi-allelic, ambiguous (A/T or C/G), or represented in less than 80% of trait GWAS datasets, and found proxy variants in linkage

disequilibrium (LD) using HaploReg v4.1 [10,11]. One proxy variant was selected to represent each locus based on prioritization of 1) non-ambiguity, 2) trait GWAS representation and 3) strong LD ($r^2 \geq 0.8$) with the initial variant. Ambiguous variants were avoided to reduce the chance of using the incorrect allele when extracting data from multiple summary statistic files.

This expanded set of T2D-associated variants was then LD-pruned based on LD data of European ancestry (CEU) using the LDlinkR package in R by LD Link [11,12,13,14]. We performed stringent LD-pruning of variants using $r^2 < 0.1$; when performing LD-pruning, variants with lower *P*-values for T2D were preferentially retained. Following LD-pruning, the final LD-pruned list of T2D variants consisted of 324 variants (**Table S2**). For each variant, we identified the T2D risk allele with odds ratio (OR) > 1 using the largest available T2D GWAS and used this allele in assessing associations with traits.

A total of six GWAS and exome-chip datasets [15,16,17] were used as input for identifying coronary artery disease (CAD) genetic loci. For chronic kidney disease (CKD), 39 GWAS summary statistics results were queried, which included CKD GWAS as well as diabetic kidney disease (DKD), end-stage renal disease (ESRD), estimated glomerular filtration rate (eGFR) and cystatin C [18,19,20,21,22,23]. The additional kidney GWAS were included to expand the number of loci; variants were included only if at least one of their associations with CKDGen, DNCRI, and SUMMIT CKD GWAS [18,19,20] was significant at Bonferroni *P*-value cutoff of 0.05/N_initial_variants (N=2,269) in order to ensure correct risk allele alignment. The aforementioned variant selection procedure was applied to CAD and CKD variants identified from the input datasets, selecting 219 variants for CAD and 70 for CKD.

Trait selection

For trait selection, we utilized summary statistics available for GWAS of glycemic traits, anthropometric traits, vital signs, and additional laboratory measures in the AMP-CMDKP. We restricted the analysis to GWAS of continuous traits with sample sizes larger than 5,000 to

reduce false positive associations and selected studies from predominantly European populations to minimize heterogeneity, as noted above. Additionally, to incorporate additional potentially relevant biomarkers, we included GWAS summary statistics of serum biomarkers available from the UK Biobank [23]. Our goal was to let the genetics guide which traits were included in the clustering analysis, and thus traits were used only if the minimum *P*-value across the final set of variants was lower than a Bonferroni *P*-value cutoff of 0.05/N_final_variants (N=324); therefore all the traits included in the clustering analysis had robust associations with at least one T2D variant. If two or more traits were highly correlated ($|r| \geq 0.85$), we kept the one with the most significant *P*-value across the selected variants and discarded the rest. Among an initial set of 75 trait GWAS datasets, two traits were dropped during the minimum *P*-value filtering process and nine were dropped due to high correlation with other traits for the selected T2D variant set, leaving 64 traits (**Table S3**).

For the selected lists of variants and traits, we utilized the GWAS summary statistics to generate a matrix of standardized Z-scores, choosing the T2D risk-increasing allele for each variant and dividing the estimated regression coefficient beta by the standard error. To account for the differences in sample size across trait GWAS studies, we scaled the standardized Z-scores in a two-step process: each value was divided by the square root of the sample size for each variant in each trait GWAS, then all elements were multiplied by the mean of square root of median sample size across all SNPs in each GWAS.

bNMF clustering

The variant-trait association matrix Z (m by n, m: # of variants, n: # of traits) was constructed as above. We then generated a non-negative input matrix X (2m by n) by concatenating two separate modifications of the original Z matrix: one containing all positive standardized Z-scores (zero otherwise) and the other all negative standardized Z-scores multiplied by -1.

The bNMF procedure factorizes X into two matrices, W (2m by K) and H^T (n by K), as X ~ WH with an optimal rank K, corresponding to the association matrix of variants and traits to the number of clusters, respectively. While conventional nonnegative matrix factorization (NMF) requires the desired model order K as an input, bNMF determines an optimal K which best balances between an error measure $||X-WH||^2$ and a penalty for model complexity derived from a nonnegative half-normal prior for W and H [24,25,26]. Furthermore, bNMF iteratively regresses out irrelevant components in representing X with an automatic relevance determination technique, which enables an optimal inference for the number of clusters K. The key features for each cluster are determined by the most strongly associated traits, a natural output of the bNMF approach. bNMF algorithm was performed in R Studio for 1,000 iterations with maximum number of cluster K set to 20, and the maximum posterior solution at the most probable K was selected for downstream analyses. The output of this clustering consists of matrices of cluster-specific weights for each variant (W) and trait (H) [6].

To define a set of strongest-weighted variants in each cluster and maximize the signal to noise ratio of weights, we developed a method to determine a cluster weight cutoff for the clusters (**Figure S2**). This involved aggregating the weights from all the clusters and plotting them in descending order. We fitted a line to the top 1% of the weights and another line to the bottom 80% of the weights. We identified the point where the distance to the first line became shorter than to the second line, and made this the cutoff for cluster weights. For T2D, cluster weight cutoff was thus set at 0.832.

Cluster associations with relevant phenotypes using GWAS summary statistics

To better characterize each cluster, particularly with regard to the associations of the loci with glycemic traits included in the clustering process, we generated GWAS-partitioned polygenic scores (GWAS pPS) for each cluster, utilizing inverse-variance weighted fixed effects

meta-analysis of GWAS summary statistics. For these analyses, the set of strongest-weighted variants above the weight cutoff for each cluster (above the weight cutoff, described above) were included in the model. We performed these meta-analyses with the dmetar package in R [27] using GWAS summary statistics.

Additionally, we applied this same approach to test cluster associations with relevant cardiometabolic outcomes studied in GWAS. As opposed to above, these outcomes were independent of the traits included in the bNMF clustering analysis. We tested associations between each cluster and seven relevant cardiometabolic disease outcomes; coronary artery disease (CAD), chronic kidney disease (CKD), estimated glomerular filtration rate (eGFR), hypertension, ischemic stroke, diabetic retinopathy, and diabetic neuropathy (**Table S4**). The significance threshold was set to 0.05/(7×K), representing a Bonferroni correction for 7 outcomes and K clusters.

Functional annotation and enrichment analysis

At each locus, we calculated approximate Bayes Factors (aBF) for all variants 500 kb upstream and downstream with $r^2 \geq 0.1$, with the index variant (100% credible set) from effect size estimates and standard errors, using the approach of Wakefield [28]. We then calculated a posterior probability for each variant by dividing the aBF by the sum of all aBF in the credible set.

We obtained previously published 13-state ChromHMM [29] chromatin state calls for 28 cell types, excluding cancer cell lines [30]. For each cell type, we extracted chromatin state annotations for enhancer (Active Enhancer 1, Active Enhancer 2, Weak Enhancer, Genic Enhancer) and promoter (Active Promoter) elements. We also compiled candidate cis-regulatory elements (cCREs) for 14 cell types and subtypes from published single cell chromatin accessibility datasets [31,32].

We assessed enrichment of annotations within clusters by overlapping 100% credible set variants for signals in each cluster with cell type epigenomic annotations (chromatin states and cCREs). We calculated cell type probabilities for each cluster by summing the posterior probabilities of variants in cell type enhancers or promoters, divided by the number of signals in the cluster. We derived significance for cell type probabilities for each cluster using a permutation-based test. We permuted signals and cell type labels within each cluster and then recalculated cell type probabilities, as above. We then used cell type probabilities derived from 10,000 permutations as a background distribution and performed a one-tailed test to ascertain significance for each cell type.

We also assessed epigenomic enrichment in single cell pancreatic tissue using a second method. As previously described [8], we subset loci from the Beta-cell 1 and 2 clusters, annotated variants using cCREs from INS^{high} and INS^{low} beta cells, and applied fgwas [33] in the fine mapping mode. We considered annotations significantly enriched if the lower bound of the 95% confidence interval of the natural log enrichment was greater than 0.

Partitioned Polygenic Score (pPS) analysis in the Mass General Brigham Biobank

The Mass General Brigham (MGB) Biobank (formerly Partners Biobank) provides banked samples (plasma, serum, DNA and genomics data) collected from more than 120,000 consented patients seen at hospitals and clinics across the MGB system, including Brigham and Women's Hospital, Massachusetts General Hospital, Massachusetts Eye and Ear Infirmary, Faulkner Hospital, Newton-Wellesley Hospital, McLean Hospital, North Shore Medical Center and Spaulding Rehabilitation Hospital, all in the Boston area of Massachusetts [34,35]. Patients are recruited at clinical care appointments at more than 40 sites and clinics, and also electronically through the patient portal at MGB. Biobank subjects provide consent for the use of their samples and data in clinical research. Written consent was provided by all study

participants. Approval for analysis of Biobank data was obtained by the MGB IRB, study 2016P001018.

T2D status was defined based on algorithmically defined phenotypes developed by the Biobank Portal team using both structured and unstructured electronic medical record data and clinical, computational, and statistical methods [36]. Cases were selected by this curated phenotype to have T2D with PPV of 99% and required to be of at least age 35 to further minimize misclassification of T2D diagnosis. Additional phenotypic data (laboratory measures, vital signs, and anthropometric measures) were extracted, from which we generated median values over the most recent 5 years available within the years of 2015-2020.

Up to 36,000 samples were genotyped using three versions of the Biobank SNP array offered by Illumina that is designed to capture the diversity of genetic backgrounds across the globe. The first batch of data was generated on the Multi-Ethnic Genotyping Array (MEGA) array, and the second, third, and fourth batches were generated on the Expanded Multi-Ethnic Genotyping Array (MEGA Ex) array. All remaining data were generated on the Multi-Ethnic Global (MEG) BeadChip. The genotyping data were harmonized and quality controlled with a three-step protocol, including two stages of SNP removal and an intermediate stage of sample exclusion. The exclusion criteria for genetic variants were 1) missing call rate ≥ 0.05, 2) significant deviation from Hardy-Weinberg equilibrium ($P \leq 10^{-20}$ for the entire cohort), and 3) minor allele frequency (MAF) < 0.001. The exclusion criteria for samples were 1) gender discordance between the reported and genetically predicted sex, 2) subject relatedness (pairs with ≥ 0.125, from which we removed the individuals with the highest proportion of missingness), 3) missing call rates per sample ≥ 0.02, and 4) population structure showing more than four standard deviations within the distribution of the study population, according to the first four principal components (PCs). Phasing was performed with SHAPEIT [37] and then imputed with the Haplotype Consortium Reference Panel [38] using the Michigan Imputation Server [39].

We performed individual-level analyses on samples restricted to individuals from European ancestry based on self-reported ancestry and genetic PC's, totaling 25,419 individuals. SNPs were included in genetic risk scores as allele dosages. All SNPs were genotyped or imputed with high quality (r^2 values > 0.95). T2D partitioned polygenic scores (pPSs) for each cluster were generated by multiplying a variant's genotype dosage by its cluster weight. For each cluster pPS, only the top-weighted variants were included, as defined above. Logistic and linear regression were performed in R v3.6.2, adjusting for age, sex, and PC's.

Results

Ten T2D genetic clusters identified by high-throughput approach

We employed a novel high-throughput pipeline to enable extraction of loci from GWAS summary statistics files and generate a variant-trait association input matrix for clustering analysis (**Figure S1**). Our pipeline started with summary statistics from 13 T2D GWAS studies available in the AMP-CMDKP [9], from which we extracted 21,666 variants associated with T2D reaching genome-wide significance. After performing stringent LD-pruning and optimizing variants for inclusion in the variant-trait association matrix (see **Materials and methods**), 324 variants remained, representing independent T2D risk loci. For these 324 variants, we extracted summary statistics for traits linked to T2D from 75 GWAS studies of measures of glycemic, anthropometric, and laboratory traits as well as vital signs, all housed in either the AMP-CMDKP [9] or publicly available analysis of the UK Biobank [23]. We let the T2D genetics guide selection of 64 relevant traits to include in subsequent cluster analysis, such that each trait was significantly associated with at least one T2D variant. Soft clustering of the resulting 324 by 64 variant-trait association matrix was performed using bNMF.

The plurality of bNMF iteration results converged on ten clusters (36.3%), which captured the five clusters identified in our previous work [6] as well as five novel clusters (**Table S5, S6**). The remaining bNMF iterations converged on nested clusters, with 6 clusters in 0.3%,

7 clusters in 1.1%, 8 clusters in 8.3%, 9 clusters in 26.6%, 11 clusters in 22.6%, 12 clusters in 4.4% and 13 clusters in 0.4%. The same six clusters (Beta-cell 1, Beta-cell 2, Proinsulin, Obesity, Lipodystrophy, Liver/Lipid, as described below) were identified in 100% of iterations, based on inspection of constituent variants and traits. This high-throughput approach therefore robustly replicated clusters previously identified via a manual approach [6].

To further interpret these clusters, we examined their most highly weighted loci and traits, as well as the aggregate associations of cluster loci with the traits via GWAS pPS (**Table 1, Table S7, Figure S3**). The clusters were named after their most defining traits. In four of the clusters (Beta-cell 1, Beta-cell 2, Proinsulin, and Lipoprotein A), the T2D risk-increasing alleles were associated with reduced fasting insulin and reduced homeostatic assessment of beta cell function (HOMA-B; GWAS pPS *P*-values<0.05), consistent with pathways related to insulin deficiency (**Figure 1**). Another five clusters (Obesity, Lipodystrophy, Liver/Lipid, ALP negative, Hyper Insulin Secretion) reflected mechanisms of impaired insulin action, with the T2D risk alleles in these clusters associated with increased fasting insulin and homeostatic assessment of insulin resistance (HOMA-IR; GWAS pPS *P*-values<0.05). The remaining cluster (SHBG) was driven by one T2D allele which was not significantly associated with fasting insulin, but trended toward increased levels (**Table 1**, **Figure 1, Figure S3, Table S7**).

Of the four clusters related to insulin deficiency (Beta-cell 1, Beta-cell 2, Proinsulin, Lipoprotein A), Beta-cell 1 and Beta-cell 2 appeared to be a combination of the single Beta-cell cluster in our previous work [6], both including several well-known loci and traits related to pancreatic beta cell function, indicating mechanisms of beta cell dysfunction leading to T2D. In Beta-cell 1, the top-weighted traits were decreased corrected insulin response (CIR) and decreased disposition index (DI) (**Table S4**); the most strongly weighted loci included known beta cell loci *MTNR1B*, *CDKAL1, HHEX, C2CD4A, ANK1, ST6GAL1, SLC35C1,* and *SLC30A8* [40] (**Table 1, Table S5, Table S6**). In Beta-cell 2 cluster, the top-weighted traits included increased fasting proinsulin adjusted for fasting insulin, and reduced HOMA-B and fasting

insulin; *TCF7L2, SLC30A8, ADCY5, GCK, DGKB, C2CD4A* and *MTNR1B* were among the top-weighted loci [41] (**Table S4, S5**).

The Beta-cell 1 and Beta-cell 2 clusters differed from each other with regard to the magnitude of glycemic trait effects, with Beta-cell 1 (N loci=63) having more marked association with reduced DI compared to Beta-cell 2 (beta=-0.05, P=3.69×10^{-61} vs beta=-0.03, P=9.02×10^{-9}), while Beta-cell 2 (N loci=28) had a more marked association with increased fasting proinsulin adjusted for fasting insulin (beta=0.02, P=9.81×10^{-43} vs beta=0.006, P=9.81×10^{-7}) and fasting glucose (beta=0.02, P=1.87×10^{-88} vs beta=0.008, P=8.78×10^{-45}), compared to Beta-cell 1 cluster (**Figure 1**, **Table S7**). Proinsulin is a prohormone precursor to insulin, and elevated fasting proinsulin levels relative to fasting insulin levels indicates defective proinsulin processing, particularly related to beta cell stress [42,43]. The stronger association with increased proinsulin levels for Beta-cell 2 vs Beta-cell 1 could therefore indicate that Beta-cell 2 relates more specifically to beta cell stress.

The Proinsulin cluster, also identified in our previous work, had top-weighted traits of decreased fasting proinsulin adjusted for fasting insulin and reduced HOMA-B (**Table S4, S5**). The top-weighted loci included distinct signals in the *ARAP1/STARD10* locus; beta cell-selective deletion of *StarD10* in mice has previously been shown to cause impaired insulin secretion [44]. In contrast to the other insulin deficiency clusters, the Proinsulin cluster (N loci=18) was significantly associated with decreased fasting proinsulin adjusted for fasting insulin (GWAS pPS P=3.51×10^{-36}) (**Figure 1**, **Table S7**), indicating a mechanism of lack of proinsulin substrate for insulin synthesis.

The Lipoprotein A cluster was novel to the present analysis and had a single highly weighted trait, increased serum lipoprotein A (Lp(a)), and a single highly weighted locus, *SLC22A3/LPA* (**Table S4, S5**). *SLC22A3/LPA* has been previously associated with serum Lp(a) levels [45] and contains the gene *LPA* encoding Lp(a). The T2D-risk-increasing allele of Lipoprotein A cluster variant (rs487152) was strongly associated with increased Lp(a) levels in

the UK Biobank (GWAS pPS P=4.06×10^{-1586}) (**Table S7**), but the underlying mechanism relating to insulin deficiency is unknown.

Of the five clusters related to mechanisms of insulin response (Obesity, Lipodystrophy, Liver/Lipid, Hyper Insulin Secretion, ALP negative), three (Obesity, Lipodystrophy, and Liver/Lipid) were also identified in our previous work, but gained additional loci (and traits) in this expanded analysis.

The Obesity cluster had most-strongly weighted traits of increased body mass index (BMI), waist circumference, percent body fat, and C-reactive protein (CRP), and key genetic signals included the well-known obesity loci *FTO* and *MC4R* [46] (**Table S4, S5**). GWAS pPS for the Obesity cluster (N loci=35) identified significant associations with increased fasting insulin (P=7.92×10^{-22}), HOMA-IR (P=7.58×10^{-19}), BMI (P=1.87×10^{-1398}), percent body fat (P=6.94×10^{-83}) and CRP (P=6.47×10^{-260}), supporting a mechanism of obesity-mediated insulin resistance.

The Lipodystrophy cluster had top-weighted traits and loci suggestive of “lipodystrophy-like” or fat distribution-mediated insulin resistance as in our prior work [6]; these included decreased adiponectin, HDL cholesterol, and modified Stumvoll insulin sensitivity index (adjusted for age, sex and BMI), and increased triglycerides and waist-hip ratio, as well as top-weighted loci *IRS1, KLF14,* and *PPARG* [47,48] (**Table S4, S5**). The Lipodystrophy cluster (N loci=54) was associated with increased fasting insulin (P=3.16×10^{-43}), HOMA-IR (P=7.47×10^{-29}) and triglycerides (P=1.18×10^{-612}), decreased insulin sensitivity index (P=1.84×10^{-38}) and HDL (P=5.19×10^{-535}).

The Liver/Lipid cluster, also identified in our previous work [6], was defined by decreased triglycerides and gamma-glutamyl transferase levels, and multiple loci previously connected to hepatic lipid or glycogen metabolism, including *GCKR, HNF1A, PPP1R3B, TOMM40/APOE,* and *PNPLA3* (**Table S4, S5**) [49,50,51,52,53,54]. GWAS pPS for this cluster (N loci=11) were

associated with reduced triglycerides (P=3.64×10^{-181}) and interestingly also reduced CRP (P=7.75×10^{-106}) and white blood cell count (P=1.42×10^{-49}).

The two remaining insulin response clusters were novel (labeled ALP negative and Hyper Insulin Secretion). The ALP negative cluster had decreased alkaline phosphatase (ALP) level as its top-weighed trait, and the *ABO* locus as the top-weighted locus (**Table S4, S5**). GWAS pPS in this cluster (N loci=4) was associated with decreased serum ALP (P=1.97×10^{-1431}) and triglycerides (P=4.49×10^{-247}). The *ABO* locus has previously been connected to T2D risk, but the mechanism is unknown: *ABO* gene knock-out in a murine pancreatic beta-cell line has been shown to alter insulin secretion [55], and there may be differential risk of T2D by ABO blood group [56,57]. The Hyper Insulin Secretion cluster included top-weighted traits of increased DI and CIR, and loci *PPP1R3B, CNTN2, DTNB*, *SREBF1*, and *TNF* (**Table S4, S5**). The Hyper Insulin Secretion GWAS pPS (N loci=32) was associated with increased CIR (P=1.16×10^{-14}), DI (P=2.89×10^{-14}), BMI (P=1.01×10^{-26}), and reduced HDL (P=1.09×10^{-110}) and SHBG (P=1.07×10^{-100}).

The final cluster, labeled SHBG, was novel to the current work and not significantly associated with fasting insulin (GWAS pPS P=0.36). The cluster was driven by a single trait and locus: decreased SHBG levels and the *SHBG* locus (**Table S4, S5**). GWAS pPS in the SHBG cluster (N loci=1) was significantly associated with reduced SHBG (P=1.2×10^{-1784}) and reduced IGF-1 (P=4.12x10^{-13}).

T2D Clusters differ in tissue enrichment including single cell islets

To acquire further evidence for the suspected mechanistic pathways represented by clusters and assess the biological difference between the clusters, we analyzed the top weighted loci in each cluster for enrichment of epigenomic annotations across 28 tissues. The T2D clusters displayed clearly different patterns of tissue enhancer/promoter enrichment (**Figure 3A, Table S8a**). In line with expected mechanisms, the Beta-cell 1, Beta-cell 2, and

Proinsulin clusters were significantly enriched for pancreatic islet tissue (FDR<0.05). The Liver/Lipid and ALP negative clusters were significantly enriched in liver tissue (FDR<0.01). The Lipodystrophy cluster was strongly enriched for adipose tissue (FDR<0.01). Additionally, both Beta-cell 1 and 2 clusters had enrichment in adipose and the brain hippocampus (FDR<0.01). The Obesity cluster was most transcriptionally enriched in human epidermal keratinocytes (NHEK) and hASC-t3 pre-adipose cells, both at nominal significance (*P*<0.05, FDR=0.11); of note, a stigmata of insulin resistance commonly seen in obese patients with T2D is acanthosis nigricans, which is hyperpigmentation of skin driven by proliferation of epidermal keratinocytes [58].

We also interrogated newly available chromatin profiles from 14.3k pancreatic islet cells, which Chiou *et al.* subsetted based on their chromatin profiles [8]. There were two epigenomic subsets of islet beta cells, with the insulin gene INS among the genes with the most variable promoter accessibility, such that subsets were labeled Beta INS^{high} and Beta INS^{low}. The Beta INS^{high} islet cells were noted to be enriched for promoter accessibility for genes involved in insulin secretion, whereas the Beta INS^{low} was enriched for genes involved in stress-induced signaling response [8]. When assessing enrichment of our genetic clusters, we found that our Beta-cell 1 genetic cluster was enriched only in Beta INS^{high} cells (P=0.0001, FDR=0.0014), whereas our Beta-cell 2 genetic cluster was nominally enriched in both Beta INS^{high} and Beta INS^{low} cells (P=0.025, P=0.013, respectively, FDR=0.18 for both), (**Figure 3B, Table S8b**). Further supporting the delineation of the Beta cell loci into two separate sub-pathways, the same trend was observed in our fgwas enrichment analysis: Beta-cell 1 was significantly enriched only in INS^{high} (ln(enrichment) (95% CI) INS^{high} 2.32 (1.31-3.12); INS^{low} -0.36 (-1.79-0.55)) whereas Beta-cell 2 was significantly enriched in both single cell subsets (ln(enrichment) (95% CI), INS^{high} 1.61 (0.22 - 2.96); INS^{low} 2.11 (0.73 - 3.46)) (**Figure 3C**). Together these results supported that Beta-cell 1 and Beta-cell 2 clusters relate to distinct physiological mechanisms, with Beta-cell 2 again connected to a stress-induced pancreatic state.

Also of interest, within the pancreas single cell data, the Liver/Lipid cluster was most enriched for alpha cells, (P=0.007, FDR=0.099); alpha cells secrete glucagon, which acts to release glucose from the glycogen stores in the liver, providing further connection between these T2D genetic loci with liver function.

T2D clusters are differentially associated with clinical traits and outcomes

We next tested whether the pPS derived from the T2D genetic clusters were associated with clinical traits and outcomes in independent datasets. To do this, we generated cluster pPSs in the hospital-based MGB Biobank (N=25,419), on samples restricted to individuals from European ancestry based on self-reported ancestry and genetic PC's.

We first confirmed that cluster pPSs were associated with expected traits, as available, in this study population (**Figure S4b**). For example, increased Obesity pPS was associated with increased BMI (beta ln(BMI) per SD pPS=0.017, $P=4.65\times10^{-42}$) and the Lipoprotein A cluster was associated with increased Lipoprotein A levels (beta per SD pPS=18.76, $P=5.95\times10^{-7}$) (**Table S9**).

Next we assessed whether the cluster pPS were associated cardiometabolic clinical outcomes related to T2D: CAD, CKD, eGFR, hypertension, ischemic stroke, diabetic retinopathy, and diabetic neuropathy (**Table S4, Table S10, Figure 4a, Figure S4a**). Starting with GWAS (which were not included in the clustering), we identified GWAS pPS associations with all outcomes except diabetic retinopathy risk. Increased GWAS pPSs from the Beta-cell1, Proinsulin, Obesity, Lipodystrophy, and Lipoprotein A clusters were most significantly associated with increased risk of CAD ($P<5\times10^{-5}$). Increased GWAS pPSs in the Obesity and Liver/Lipid clusters were associated with increased risk of CKD ($P<5\times10^{-5}$), and GWAS pPSs in the Liver/Lipid, ALP negative, and SHBG clusters were associated with reduced eGFR ($P<5\times10^{-4}$). Increased GWAS pPSs of the Beta-cell1, Beta-cell2, Proinsulin, Obesity, Lipodystrophy clusters were significantly associated with increased risk of ischemic stroke

($P<5\times10^{-4}$). Increased Beta-cell1 and Hyper Insulin Secretion cluster GWAS pPS were associated with increased risk of neuropathy in T2D ($P<10^{-3}$).

Several of these cardiometabolic outcomes were available to study in MGB Biobank: eGFR, CKD, CAD and hypertension. We were able to replicate several of the GWAS pPS outcome associations, notably all the most significant associations (GWAS pPS $P<10^{-15}$). Individuals with increased Obesity cluster pPS had significantly increased risk of hypertension ($P=1.60\times10^{-7}$), those with increased Lipodystrophy cluster pPS had significantly increased risk of CAD ($P=4.75\times10^{-7}$), CKD ($P=4.01\times10^{-7}$), and hypertension ($P=1.82\times10^{-7}$), and those with increased Liver/Lipid cluster pPS had significantly reduced eGFR ($P=7.30\times10^{-4}$) (**Figure 4b, Table S11**). Of note, in both analyses, the Liver/Lipid pPS was nominally associated with reduced risk of CAD (**Figure 4, Table S9, Table S11**). Our results thus indicated that genetic pathways leading to T2D have distinct effects on other cardiometabolic conditions, supporting analyses performed with the original five T2D genetic clusters [59].

Loci from CAD and CKD share mechanistic pathways with T2D

To identify shared pathways among cardiometabolic outcomes and demonstrate portability of our high-throughput pipeline, we applied the same clustering approach to 219 CAD loci and 70 CKD loci. Five clusters were identified in clustering analysis of 219 CAD loci (ALP negative, Lipoprotein A, HDL negative, Cholesterol and Blood markers increased) and four clusters were identified for the 70 CKD loci (Blood markers increased, Blood markers decreased, Urea increased, Urea decreased, and Lipoprotein A) (**Tables S12-15**, **Figure S5**). Based on inspection of constituent variants and traits in the clusters of T2D, CAD, and CKD, one cluster, Lipoprotein A, was shared by all three diseases. Additionally, the ALP negative cluster was shared between T2D and CAD, and the Blood markers increased cluster between CAD and CKD.

Discussion

Novel approaches are needed to connect the currently identified hundreds of T2D genetic loci to disease pathways and also accommodate the rapid pace of new loci discovery. Here, we describe expanded clustering of T2D variants, using a high-throughput pipeline for extracting and preprocessing variants from multiple GWAS datasets and generating a variant-trait association matrix. The resulting matrix consisted of 324 T2D genetic variants and 64 diabetes-related metabolic traits from publicly available GWAS datasets. By applying bNMF soft clustering to this matrix, we identified ten robust clusters of T2D variants, representing biologically meaningful mechanistic pathways.

Among the ten clusters, we replicated the five identified in our previous work of 94 T2D variants (Beta-cell, Proinsulin, Obesity, Lipodystrophy, Liver/Lipid) [6], with the Beta-cell cluster now subdivided into two distinct clusters, and also identified four additional novel clusters related to pronounced insulin secretion, levels of alkaline phosphatase, lipoprotein-A, and sex hormone-binding globulin. In contrast to our prior work, which involved manual curation of loci published in GWAS manuscripts to generate the input list of variants, the current approach allowed for use of uncurated GWAS summary statistics files and included additional newly available datasets, more than tripling the number of input genetic loci. Additionally, the current work incorporated 17 more relevant GWAS traits (included only if associated with at least one T2D variant). Thus, replication of the previously identified five clusters provides strong validation of this high-throughput approach, with the newly identified clusters driven by traits or loci not available in the prior analysis. We provide relevant code for use by the broader complex genetic disease community.

Three of the ten T2D clusters identified in this work (Beta-cell 1, Beta-cell 2, and Proinsulin) clearly related to pancreatic beta-cell function, with the two Beta-cell clusters differing from the Proinsulin cluster with regard to the direction of association with fasting proinsulin adjusted for fasting insulin. The association of the Proinsulin cluster with lower fasting

proinsulin levels indicated a proximal defect in the insulin synthesis pathway. All three clusters were enriched in pancreatic islet tissue enhancers and promoters in the epigenomics analysis. Additionally, loci in the Beta-cell 1 cluster were significantly enriched for a subset of single beta cells predicted independent of our work based on RNA transcript levels to represent a normal state of pancreatic beta cell function, whereas loci in Beta-cell 2 cluster had a unique signal of enrichment for single beta cells predicted to be in a stressed state [8]; these functional distinctions between Beta-cell 1 and 2 supported our independent approach of phenotypically informed clustering T2D loci.

Three other T2D genetic clusters (Obesity, Lipodystrophy, Liver/Lipid) replicated findings from our prior work related to pathways of insulin resistance, gaining additional loci and traits compared to the prior analysis. Loci in these three clusters were most enriched for enhancers in tissues for the suspected mechanisms: pre-adipocytes, adipocytes, and liver tissue, respectively. The distinction between fat accumulation in the Obesity cluster and abnormal fat compartmentalization in the Lipodystrophy cluster may be supported by the differential enhancer enrichment shown for different developmental stages of the same adipocyte lineage.

In addition to a second Beta-cell cluster, there were four newly identified T2D genetic clusters from this work: ALP negative (containing the *ABO* locus), Lipoprotein A, SHBG, and Hyper Insulin Secretion.

The ALP negative cluster was driven by reduced serum alkaline phosphatase levels and four genetic loci, with the *ABO* locus most strongly weighted. Alkaline phosphatase is a circulating enzyme with isoforms that have been shown to vary in level by blood group [60]. This same cluster was also identified in the CAD clustering, with the sentinel SNP rs649129 in weak LD with the T2D SNP rs545971 (r^2=0.29 in 1000 Genomes European populations). The *ABO* gene has previously been connected to T2D, CAD, and thromboembolic risk [55,56,57,61], although the underlying mechanisms are not fully understood. In Trégouët *et al*. [62], rs657152 in *ABO* locus, which is in LD with our *ABO* variant (rs545971, r^2=0.98 in CEU) was associated

with increased risk of venous thromboembolism at genome-wide significance (P=2.22×10^{-13}). While we found that the ALP negative cluster was significantly associated with increased fasting insulin, suggesting a mechanistic pathway of insulin resistance, and enriched for transcriptional activity in liver tissue, further research will be needed to better understand how this cluster relates to T2D risk.

The Lipoprotein A cluster included a single locus (*SLC22A3*/*LPA*) and biomarker Lp(a), pointing to a genetic pathway whereby the genetic alteration at this locus is associated with increased Lp(a) levels and also increased risk of T2D. The relationship between Lp(a) and cardiometabolic disease is complex, and genetic interrogation of *LPA* has been complicated by the fact that plasma concentration of Lp(a) is influenced by kringle IV type 2 (KIV-2) repeats in addition to other genetic variation. While rs487152, marking the *SLC22A3*/*LPA* locus in our analysis was associated with both increased Lp(a) levels and T2D risk, several epidemiological studies have shown an inverse association between Lp(a) concentrations and risk of T2D [63,64]. In fact, it has been shown that almost half of the physiologic variation in Lp(a) concentrations can be explained by the number of kringle IV type 2 (KIV-2) repeats on *LPA* gene, with the number of KIV-2 repeats inversely associated with Lp(a) levels. Rare loss of function genetic variants associated with increased number of KIV-2 repeats (and thus lower levels of Lp(a)) have been shown by Mendelian Randomization to be causally associated with increased risk of T2D [65]. Another Mendelian Randomization study that failed to show an association with T2D, included a variant in very weak LD with the variant used in our analysis, rs487152 (rs10455872, r^2=0.091) [66]. Thus the differential relationship between Lp(a) and T2D observed for the variant in our work (rs487152) vs seen in prior epidemiological and genetic studies (which included variants not in LD with our variant) highlights that there are likely multiple genetic pathways impacting Lp(a) level that may have differential effects on T2D risk.

The SHBG cluster was also driven by a single locus and biomarker. Our results point to a genetic pathway whereby alteration of the *SHBG* locus leads to reduced SHBG levels and

increased T2D risk. Consistent with our findings, previous studies have suggested that low circulating levels of SHBG are causally related to increased risk of T2D in women and men [67,68].

We generated pPSs of top-weighted cluster loci and identified significant associations in individuals with expected traits (**Table S7, S9**) as well as with clinical outcomes for several of the clusters (Beta-cell1, Beta-cell2, Proinsulin, Obesity, Lipodystrophy, Liver/Lipid) (**Table S10, S11**). Such findings point to marked heterogeneity in T2D loci associations that could be missed without delineation of loci into clusters and suggest important mechanistic differences between clusters. At the same time, the effect sizes of the pPSs on clinical outcomes were likely too small to be of clinical utility at the individual-level. Nevertheless, as we have seen with traditional polygenic scores, expansion in the number of loci with time, can lead to improved disease risk prediction [3]; thus perhaps with further development of the pPS in the future, there will be clearer clinical application.

The strengths of this study include the high-throughput approach for preprocessing variants and traits from multiple GWAS datasets in a semi-automated way. The high-throughput pipeline allows for efficient updating as additional GWAS results become available. This method can also be readily applied to other diseases beyond T2D to identify key pathways and code has been made freely available. We included here application of the pipeline to CAD and CKD, demonstrating transferability of the approach and shared pathways among the three cardiometabolic outcomes.

Limitations of this study include clustering of only available phenotypes from GWAS. It is possible that additional pathways exist that are not captured using the set of traits included in the analysis. Additionally, due to methodological challenges and data availability we have focused on GWAS datasets from populations of European ancestry. Our clustering approach currently requires relative homogeneity across GWAS studies with regard to ancestral population; inclusion of GWAS from more than one ancestral population may lead to either

genetic variants not being present due allele frequency differences across populations (introducing missing data, which the clustering method cannot tolerate) or lead to the results from clustering method artifactually driven by patterns reflective of ancestry rather than disease biology. We are actively pursuing application of this method in non-European populations through additional efforts. It is worth noting that bNMF generates weights for all included elements in the matrix, and it is not known how best determine a cut-off threshold for cluster membership; we have applied a reasonable strategy to maximize signal to noise. Finally, whether the associations of specific genetically derived clusters with metabolic traits remain constant throughout the disease course has not been examined in this cross-sectional analysis.

In summary, we have identified ten robust genetic clusters pointing to mechanistic pathways of T2D using a high-throughput clustering pipeline of GWAS summary statistics. These clusters were readily interpretable, even if not yet fully understood, and displayed tissue-specific enrichment patterns even within single cell pancreatic tissue, supporting a biological basis of the genetic cluster assignment. We further utilized cluster pPSs to stratify patients genetically that resulted in associations with distinct clinical features and cardiometabolic outcomes. We demonstrate that our approach can be readily applied to other complex diseases, with identification of shared genetic pathways between T2D, CAD, and CKD. Thus, we contribute to further delineation of cardiometabolic disease genetic pathways using a data-driven approach informed by physiology and have made code for our pipeline freely available.

Description of Supplemental Data

Supplemental Data include five figures and fifteen tables.

Declaration of Interests

The authors declare no competing interests.

Data and Code Availability

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

Funding

This work was supported by FNIH RFP-13 and the MGH Transformative Scholars Award.

Contribution Statement

JC, MG, JCF and MSU conceived the research question. MSU, JK, MG, JC, KG and JCF conceived the methodology. HK, JC, MG, TM, JMM and MSU curated the data. HK, KS and JC conducted the analysis and visualized the results. HK, JC and MSU wrote the initial draft of the paper and incorporated co-author comments. KEW, KS, JBC, TM, MG, JMM, SK, JCF, KG and AKM provided feedback on the analysis, and critically reviewed the manuscript. All co-authors approved the final version of the paper. MSU and JCF are the guarantors of this work and, as such, had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Web Resources

Interactive results are viewable on the Common Metabolic Disease Knowledge Portal (https://hugeamp.org/).

References

1. Saeedi P, Petersohn I, Salpea P, et al (2019) Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9 edition. Diabetes Res Clin Pract 157:107843

2. Redondo MJ, Hagopian WA, Oram R, et al (2020) The clinical consequences of heterogeneity within and between different diabetes types. Diabetologia 63(10):2040–2048

3. Udler MS, McCarthy MI, Florez JC, Mahajan A (2019) Genetic Risk Scores for Diabetes Diagnosis and Precision Medicine. Endocr Rev 40(6):1500–1520

4. 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

5. 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

6. Udler MS, Kim J, von Grotthuss M, et al (2018) Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis. PLoS Med 15(9):e1002654

7. Mahajan A, Wessel J, Willems SM, et al (2018) Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. Nat Genet 50(4):559–571

8. Chiou J, Zeng C, Cheng Z, et al (2021) Single-cell chromatin accessibility identifies pancreatic islet cell type- and state-specific regulatory programs of diabetes risk. Nat Genet 53(4):455–466

9. Human Genetics Knowledge Portal - Home. https://hugeamp.org/. Accessed 19 Mar 2021

10. HaploReg v4.1. https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php. Accessed 19 Mar 2021

11. 1000 Genomes. https://www.internationalgenome.org/home. Accessed 19 Mar 2021

12. Home - SNP - NCBI. https://www.ncbi.nlm.nih.gov/snp/. Accessed 19 Mar 2021

13. Myers TA, Chanock SJ, Machiela MJ (2020) : An R Package for Rapidly Calculating Linkage Disequilibrium Statistics in Diverse Populations. Front Genet 11:157

14. NCI, CBIIT, DCEG, Machiela LDlink. https://ldlink.nci.nih.gov/?tab=home. Accessed 18 Mar 2021

15. van der Harst P, Verweij N (2018) Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. Circ Res 122(3):433–443

16. Nikpay M, Goel A, Won H-H, et al (2015) A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nat Genet

47(10):1121–1130

17. (2016) Coding Variation in ANGPTL4, LPL, and SVEP1 and the Risk of Coronary Disease. N Engl J Med 374(19):1898

18. Wuttke M, Li Y, Li M, et al (2019) A catalog of genetic loci associated with kidney function from analyses of a million individuals. Nat Genet 51(6):957–972

19. Salem RM, Todd JN, Sandholm N, et al (2019) Genome-Wide Association Study of Diabetic Kidney Disease Highlights Biology Involved in Glomerular Basement Membrane Collagen. J Am Soc Nephrol 30(10):2000–2016

20. van Zuydam NR, Ahlqvist E, Sandholm N, et al (2018) A Genome-Wide Association Study of Diabetic Kidney Disease in Subjects With Type 2 Diabetes. Diabetes 67(7):1414–1427

21. Locke AE, Steinberg KM, Chiang CWK, et al (2019) Exome sequencing of Finnish isolates enhances rare-variant association power. Nature 572(7769):323–328

22. Kurki MI, Karjalainen J, Palta P, et al (2022) FinnGen: Unique genetic insights from combining isolated population and national health register data. bioRxiv

23. UK Biobank — Neale lab. http://www.nealelab.is/uk-biobank. Accessed 18 Mar 2021

24. Tan VYF, Févotte C (2013) Automatic relevance determination in nonnegative matrix factorization with the β-divergence. IEEE Trans Pattern Anal Mach Intell 35(7):1592–1605

25. Kim J, Mouw KW, Polak P, et al (2016) Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. Nat Genet 48(6):600–606

26. Kasar S, Kim J, Improgo R, et al (2015) Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. Nat Commun 6(1):1–12

27. Harrer, M., Cuijpers, P., Furukawa, T. & Ebert, D. D. (2019). dmetar: Companion R Package For The Guide “Doing Meta-Analysis in R”. R package version 0.0.9000. URL http://dmetar.protectlab.org/. Accessed 29 Apr 2021

28. Wakefield J (2007) A Bayesian measure of the probability of false discovery in genetic epidemiology studies. Am J Hum Genet 81(2):208–227

29. Ernst J, Kellis M (2012) ChromHMM: automating chromatin-state discovery and characterization. Nat Methods 9(3):215–216

30. Varshney A, Scott LJ, Welch RP, et al (2017) Genetic regulatory signatures underlying islet gene expression and type 2 diabetes. Proc Natl Acad Sci U S A 114(9):2301–2306

31. Zhang K, Hocker JD, Miller M, et al (2021) A cell atlas of chromatin accessibility across 25 adult human tissues. Cold Spring Harbor Laboratory 2021.02.17.431699

32. Chiou J, Geusz RJ, Okino M-L, et al (2021) Interpreting type 1 diabetes risk with genetics and single-cell epigenomics. Nature 1–5

33. Pickrell JK (2014) Joint Analysis of Functional Genomic Data and Genome-wide

Association Studies of 18 Human Traits. Am J Hum Genet 94(4):559

34. Smoller JW, Karlson EW, Green RC, et al (2016) An eMERGE Clinical Center at Partners Personalized Medicine. Journal of Personalized Medicine 6(1):5

35. 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). https://doi.org/10.3390/jpm6010002

36. Yu S, Liao KP, Shaw SY, et al (2015) Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. J Am Med Inform Assoc 22(5):993–1000

37. Delaneau O, Zagury J-F, Marchini J (2013) Improved whole-chromosome phasing for disease and population genetic studies. Nat Methods 10(1):5–6

38. McCarthy S, Das S, Kretzschmar W, et al (2016) A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet 48(10):1279–1283

39. Das S, Forer L, Schönherr S, et al (2016) Next-generation genotype imputation service and methods. Nat Genet 48(10):1284–1287

40. Rosengren AH, Braun M, Mahdi T, et al (2012) Reduced Insulin Exocytosis in Human Pancreatic β-Cells With Gene Variants Linked to Type 2 Diabetes. Diabetes 61(7):1726–1733

41. Zhou Y, Park S-Y, Su J, et al (2014) TCF7L2 is a master regulator of insulin production and processing. Hum Mol Genet 23(24):6419–6431

42. Mezza T, Ferraro PM, Sun VA, et al (2018) Increased β-Cell Workload Modulates Proinsulin-to-Insulin Ratio in Humans. Diabetes 67(11):2389–2396

43. Røder ME, Porte D Jr, Schwartz RS, Kahn SE (1998) Disproportionately elevated proinsulin levels reflect the degree of impaired B cell secretory capacity in patients with noninsulin-dependent diabetes mellitus. J Clin Endocrinol Metab 83(2):604–608

44. Carrat GR, Hu M, Nguyen-Tu M-S, et al (2017) Decreased STARD10 Expression Is Associated with Defective Insulin Secretion in Humans and Mice. Am J Hum Genet 100(2):238–256

45. Qi Q, Workalemahu T, Zhang C, Hu FB, Qi L (2011) Genetic variants, plasma lipoprotein(a) levels, and risk of cardiovascular morbidity and mortality among two prospective cohorts of type 2 diabetes. Eur Heart J 33(3):325–334

46. Choquet H, Meyre D (2011) Genetics of Obesity: What have we Learned? Curr Genomics 12(3):169–179

47. Yaghootkar H, Scott RA, White CC, et al (2014) Genetic Evidence for a Normal-Weight “Metabolically Obese” Phenotype Linking Insulin Resistance, Hypertension, Coronary Artery Disease, and Type 2 Diabetes. Diabetes 63(12):4369–4377

48. Yaghootkar H, Lotta LA, Tyrrell J, et al (2016) Genetic Evidence for a Link Between

Favorable Adiposity and Lower Risk of Type 2 Diabetes, Hypertension, and Heart Disease. Diabetes 65(8):2448–2460

49. Speliotes EK, Yerges-Armstrong LM, Wu J, et al (2011) Genome-Wide Association Analysis Identifies Variants Associated with Nonalcoholic Fatty Liver Disease That Have Distinct Effects on Metabolic Traits. PLoS Genet 7(3):e1001324

50. Mahdessian H, Taxiarchis A, Popov S, et al (2014) TM6SF2 is a regulator of liver fat metabolism influencing triglyceride secretion and hepatic lipid droplet content. Proc Natl Acad Sci U S A 111(24):8913–8918

51. Kozlitina J, Smagris E, Stender S, et al (2014) Exome-wide association study identifies a TM6SF2 variant that confers susceptibility to nonalcoholic fatty liver disease. Nat Genet 46(4):352–356

52. Smagris E, Gilyard S, BasuRay S, Cohen JC, Hobbs HH (2016) Inactivation of Tm6sf2, a Gene Defective in Fatty Liver Disease, Impairs Lipidation but Not Secretion of Very Low Density Lipoproteins. J Biol Chem 291(20):10659–10676

53. Raimondo A, Rees MG, Gloyn AL (2015) Glucokinase regulatory protein: complexity at the crossroads of triglyceride and glucose metabolism. Curr Opin Lipidol 26(2):88–95

54. Smagris E, BasuRay S, Li J, et al (2015) Pnpla3I148M knockin mice accumulate PNPLA3 on lipid droplets and develop hepatic steatosis. Hepatology 61(1):108–118

55. Li-Gao R, Carlotti F, de Mutsert R, et al (2019) Genome-Wide Association Study on the Early-Phase Insulin Response to a Liquid Mixed Meal: Results From the NEO Study. Diabetes 68(12):2327–2336

56. Fagherazzi G, Gusto G, Clavel-Chapelon F, Balkau B, Bonnet F (2014) ABO and Rhesus blood groups and risk of type 2 diabetes: evidence from the large E3N cohort study. Diabetologia 58(3):519–522

57. Qi L, Cornelis MC, Kraft P, et al (2010) Genetic variants in ABO blood group region, plasma soluble E-selectin levels and risk of type 2 diabetes. Hum Mol Genet 19(9):1856–1862

59. Higgins SP, Freemark M, Prose NS (2008) Acanthosis nigricans: a practical approach to evaluation and management. Dermatol Online J 14(9):2

59. DiCorpo D, LeClair J, Cole JB, et al (2022) Type 2 Diabetes Partitioned Polygenic Scores Associate With Disease Outcomes in 454,193 Individuals Across 13 Cohorts. Diabetes Care 45(3):674–683

60. Domar U, Hirano K, Stigbrand T (1991) Serum levels of human alkaline phosphatase isozymes in relation to blood groups. Clin Chim Acta 203(2-3):305–313

61. Groot HE, Villegas Sierra LE, Abdullah Said M, Lipsic E, Karper JC, van der Harst P (2020) Genetically Determined ABO Blood Group and its Associations With Health and Disease. Arteriosclerosis, Thrombosis, and Vascular Biology 40:830–838

62. Trégouët D-A, Heath S, Saut N, et al (2009) Common susceptibility alleles are unlikely to contribute as strongly as the FV and ABO loci to VTE risk: results from a GWAS approach. Blood 113(21):5298–5303

63. Mora S, Kamstrup PR, Rifai N, Nordestgaard BG, Buring JE, Ridker PM (2010) Lipoprotein(a) and Risk of Type 2 Diabetes. Clin Chem 56(8):1252

64. Habib SS, Aslam M, Shah SFA, Naveed AK (2009) Lipoprotein (a) is associated with basal insulin levels in patients with type 2 Diabetes Mellitus. Arq Bras Cardiol 93(1):28–33

65. Tolbus A, Mortensen MB, Nielsen SF, Kamstrup PR, Bojesen SE, Nordestgaard BG (2017) Kringle IV Type 2, Not Low Lipoprotein(a), as a Cause of Diabetes: A Novel Genetic Approach Using SNPs Associated Selectively with Lipoprotein(a) Concentrations or with Kringle IV Type 2 Repeats. Clin Chem 63(12):1866–1876

66. Ye Z, Haycock PC, Gurdasani D, et al (2014) The Association Between Circulating Lipoprotein(a) and Type 2 Diabetes: Is It Causal? Diabetes 63(1):332–342

67. Ding EL, Song Y, Manson JE, et al (2009) Sex hormone-binding globulin and risk of type 2 diabetes in women and men. N Engl J Med 361(12):1152–1163

68. Perry JRB, Weedon MN, Langenberg C, et al (2010) Genetic evidence that raised sex hormone binding globulin (SHBG) levels reduce the risk of type 2 diabetes. Hum Mol Genet 19(3):535–544

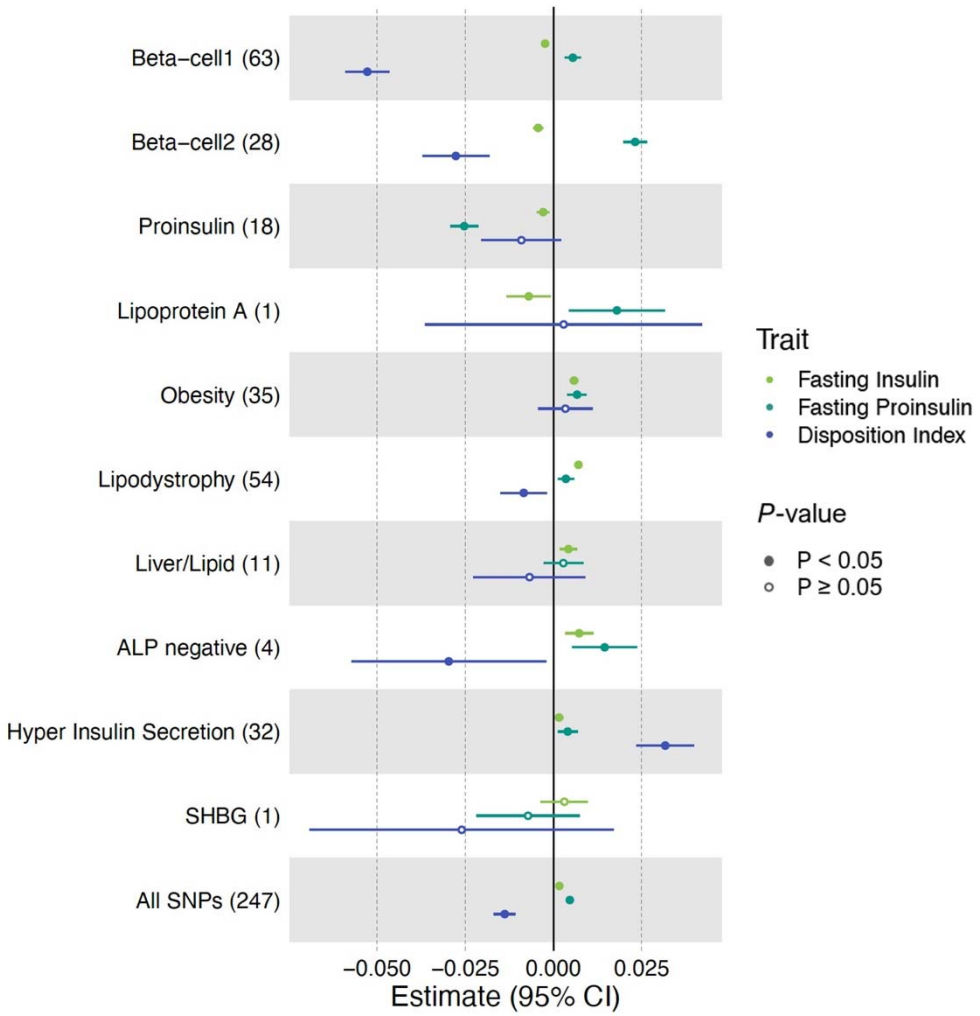
Table 1. Overview of T2D clusters

Physiological impact, key traits, key loci, suspected mechanism for each cluster. The numbers in parentheses next to cluster names indicate the numbers of top-weighted variants in each of the clusters.

Cluster	Physiological impact	Key traits	Key loci	Suspected mechanism	Note
Beta-cell 1 (63)	Insulin deficiency	CIR (-), DI (-)	*MTNR1B, CDKAL1, HHEX, C2CD4A, ANK1, ST6GAL1, SLC35C1, SLC30A8*	Beta cell function, glucose homeostasis	Beta Cell cluster from Udler et al. 2018 divided into 2 clusters
Beta-cell 2 (28)		fasting proinsulin adj FI (+), HOMA-B (-), fasting insulin (-)	*TCF7L2, SLC30A8, ADCY5, GCK, DGKB, MTNR1B, C2CD4A*	Beta cell function, insulin processing	Beta Cell cluster from Udler et al. 2018 divided into 2 clusters
Proinsulin (18)		fasting proinsulin adj FI (-), HOMA-B (-)	*ARAP1, STARD10*	Insulin synthesis	replicated Proinsulin cluster from Udler et al. 2018
Lipoprotein A (1)		lipoprotein A (+)	*SLC22A3/LPA*	Lipoprotein A metabolism	new cluster in this study
Obesity (35)	Insulin resistance	BMI (+), waistC (+), % body fat (+), CRP (+)	*FTO, MC4R*	Obesity-mediated insulin resistance	replicated Obesity cluster from Udler et al. 2018
Lipodystrophy (54)		adiponectin (-), ISI (-), HDL (-)	*IRS, PPARG, KLF14*	Fat distribution-mediated insulin resistance	replicated Lipodystrophy cluster from Udler et al. 2018
Liver/Lipid (11)		CRP (-), TG (-), GGT (-)	*GCKR, HNF1A, PPP1R3B, TOMM40, PNPLA3*	Liver/Lipid metabolism	replicated Liver/Lipid cluster from Udler et al. 2018
ALP negative (4)		ALP (-)	*ABO*	Alkaline phosphatase activity levels	new cluster in this study
Hyper Insulin Secretion (32)		DI (+), CIR (+)	*PPP1R3B, CNTN2, DTNB, TNF, SREBF1*	Insulin secretion, inflammation	new cluster in this study
SHBG (1)		SHBG (-)	*SHBG*	SHBG metabolism	new cluster in this study

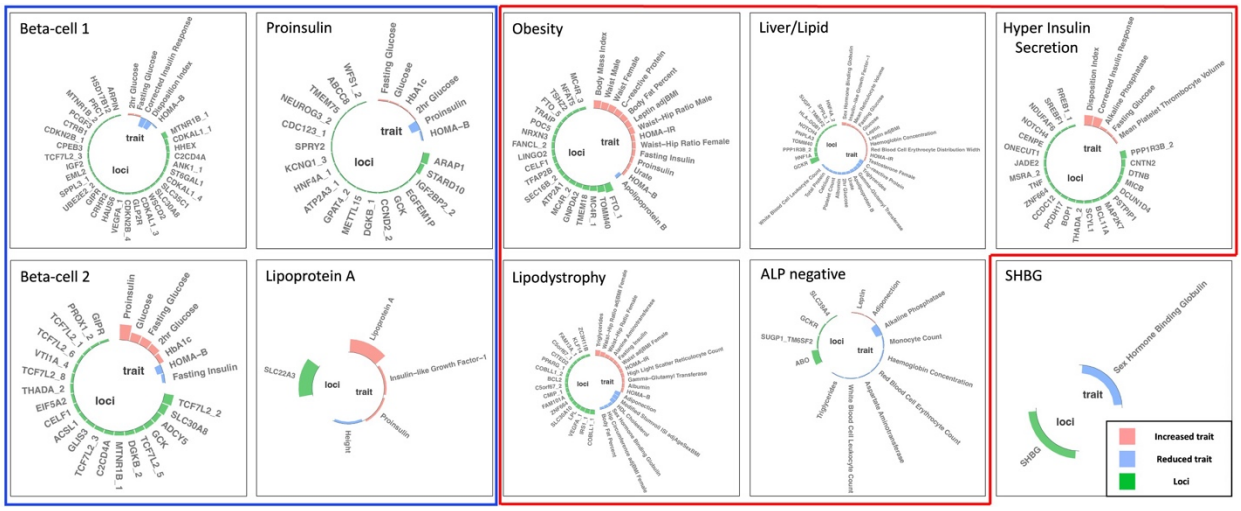
Figures

Figure 1. Cluster associations with metabolic traits using GWAS



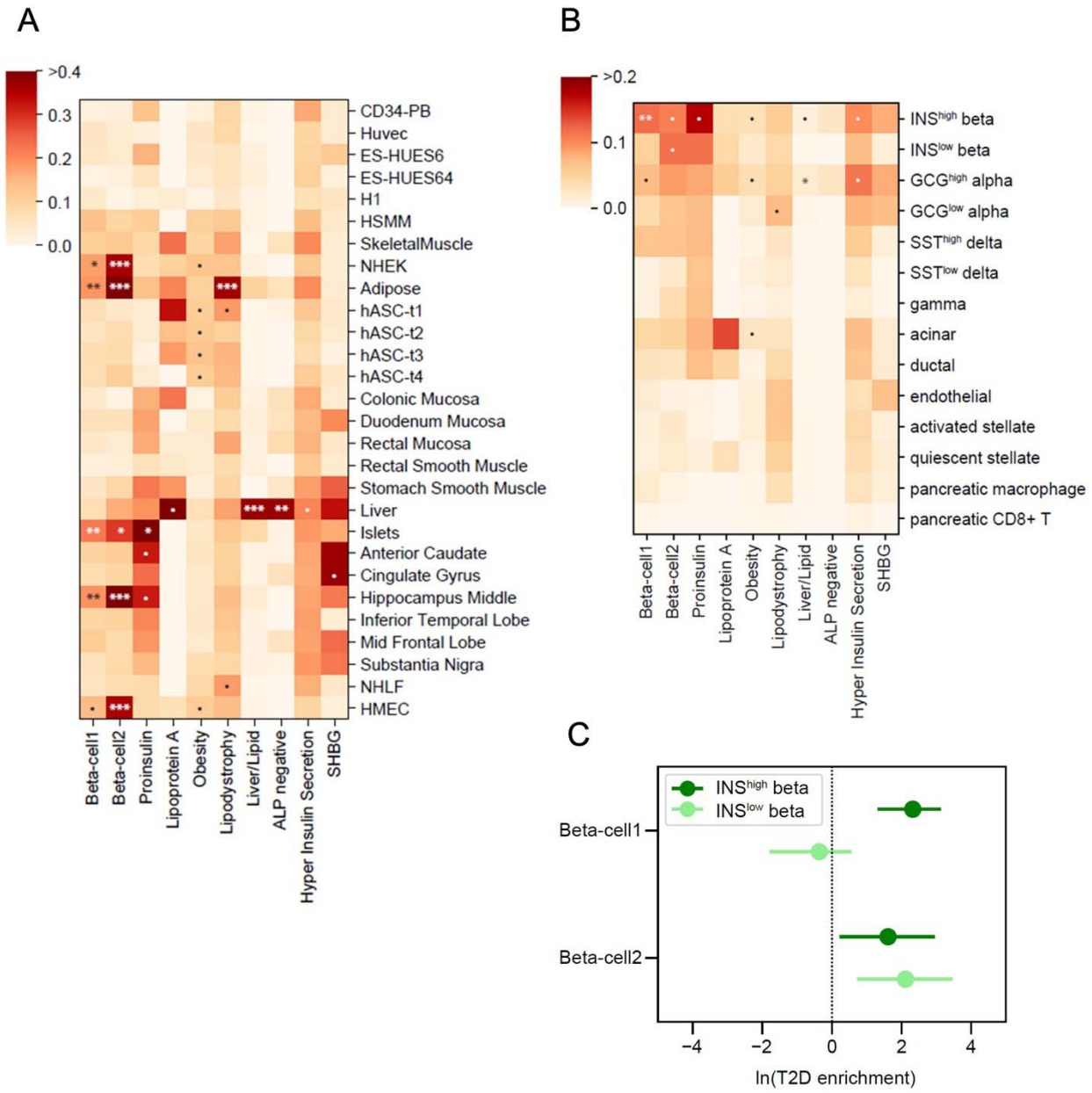
Standardized effect sizes with 95% confidence intervals of cluster pPS-trait associations derived from GWAS summary statistics shown in forest plot. Three metabolic traits (Fasting Insulin, Fasting proinsulin adjusted for fasting insulin, Disposition Index) that help discriminate clusters are displayed. The numbers in the parenthesis next to cluster names indicate the number of variants included in the analysis in each cluster. “All SNPs” include all the variants that are top-weighted in at least one cluster. Filled points indicate *P*-values less than 0.05.

Figure 2. Clusters of Type 2 Diabetes Loci



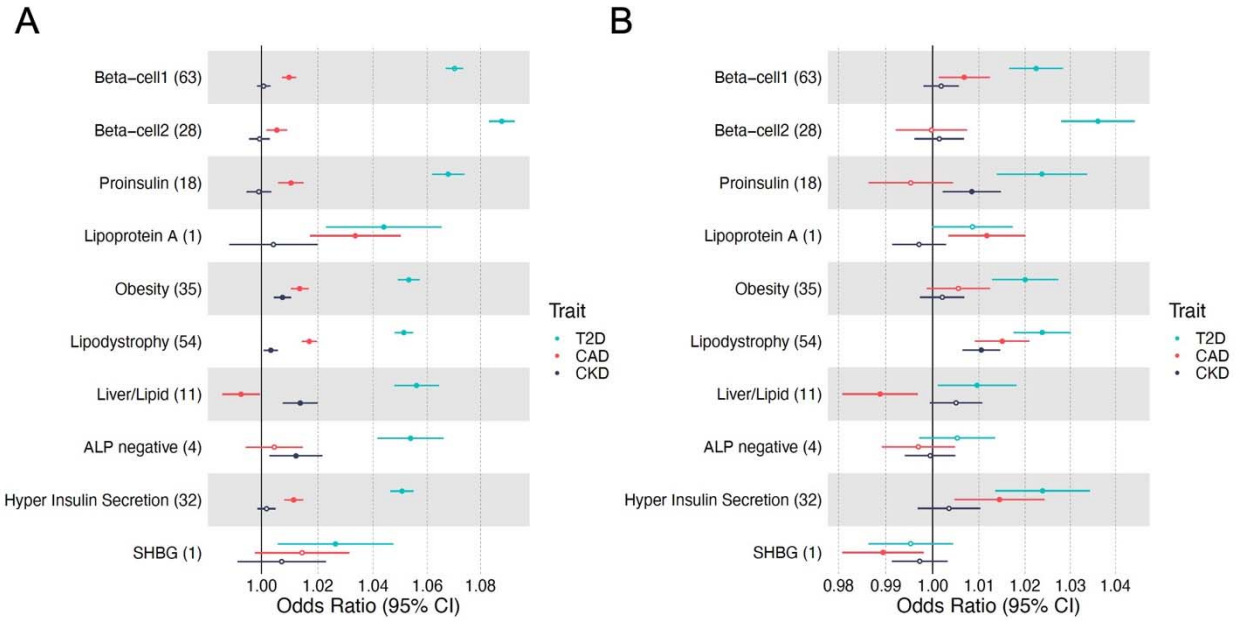
Top-weighted loci and traits in each of the ten clusters are represented in circular plots. Green bars represent top-weighted loci, red bars represent increased traits, and blue bars represent reduced traits in each cluster. A maximum of 35 elements (loci and traits) based on highest weights are displayed in each cluster. Blue outline indicates clusters associated with decreased fasting insulin levels, and red outline indicates clusters associated with increased fasting insulin levels.

Figure 3. Enrichment for tissue-specific enhancers in T2D clusters



(A) Heatmap of tissue enhancer/promoter enrichment analysis result. (B) Heatmap of pancreatic islet cell enrichment analysis result. Significance was indicated as follows: *** FDR < 0.001, ** FDR < 0.01, * FDR < 0.1, • P < 0.05. (C) Comparison of Beta-cell 1 and Beta-cell 2 clusters in fgwas enrichment analysis in functional and stressed beta-cell states shown in a forest plot.

Figure 4. Forest plot of cluster associations with outcomes using (A) GWAS and (B) individual-level data from MGB Biobank



(A) Standardized effect sizes with 95% confidence intervals of cluster pPS-outcome associations derived from GWAS summary statistics shown in forest plot. Three metabolic outcomes (T2D, CAD and CKD, all T2D unadjusted) are displayed. The numbers in the parenthesis next to cluster names indicate the number of variants included in the analysis in each cluster. Filled points indicate *P*-values less than 0.05. (B) Associations of pPSs in individuals in the MGB Biobank with clinical outcomes are shown in forest plot. Three outcomes including T2D are displayed.