THE UNIVERSITY OF CHICAGO

LEVERAGING EXPRESSION QUANTITATIVE TRAIT LOCI TO ELUCIDATE THE GENETIC ARCHITECTURE OF TYPE 2 DIABETES

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

COMMITTEE ON MOLECULAR METABOLISM AND NUTRITION

BY

JASON MATTHEW TORRES

CHICAGO, ILLINOIS

JUNE 2016

For my wife and family, whose encouragement made this possible.

Table of Contents

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

Type 2 diabetes (T2D) is a pernicious disease of impaired glucose homeostasis that affects more than 400 million people worldwide and is responsible for at least 6.8% of all deaths [163, 129]. It presents a major health burden due to myriad complications ranging from retinopathy to kidney failure and is worsened by a lack of adequate therapies [3]. Moreover, the rapid growth rate of T2D constitutes a global health crisis as the prevalence of T2D will exceed the rate of adult population growth by the year 2030 [162].

The epidemiological problem of T2D is exacerbated by the complexity of the disease itself. Familial aggregation of disease strongly supports a genetic basis for T2D [4, 81]. Yet, researchers striving to resolve the genetic architecture of T2D have had to confront the challenges of phenotypic heterogeneity, variation in age of onset, milieu of environmental risk factors such, and non-Mendelian pattern of inheritance [3, 162]. It is for these reasons that geneticist J. V. Neel designated T2D as the "geneticist's nightmare" [106].

1.1 T2D is characterized by dysfunction across multiple tissues that maintain glucose homeostasis

The pathobiology of T2D involves a complicated interplay between peripheral tissues that respond to insulin - the principal hormone responsible for lowering blood glucose levels - and the pancreatic β-cells that secrete insulin [25]. The development of insulin resistance (IR) in the liver promotes increased glucose production in the resting state [26]. Moreover, IR impairs suppression of hepatic glucose production (HGP) in response to insulin release following a meal (i.e. postprandial insulin response) and reduces glucose uptake in the liver, further diminishing glucose tolerance [56, 25]. IR in skeletal muscle also attenuates glucose uptake during the postprandial insulin response and results in postprandial hyperglycemia [43]. The pancreatic β-cells will compensate for IR in peripheral tissues by increasing insulin

secretion to maintain glucose tolerance [30]. However, this burden will cause β-cells to fail over time, chronically elevating fasting glucose concentration and ultimately promote the development of T2D [23]. Moreover, the rate of β-cell failure determines the rate of disease progression [25, 30].

Adipose tissue also plays an important role in maintaining glucose homeostasis as altered fat cell (i.e. adipocyte) metabolism contributes to glucose intolerance [25]. IR in adipocytes promotes resistance to the antilipolytic effect of insulin and results in day-long elevation of plasma free fatty acid (FFA) concentration [56]. The lipotoxicity that results from this process will in turn stimulates hepatic gluconeogenesis, further promotes IR in the liver and skeletal muscle, and worsens chronic hyperglycemia [80]. Moreover, dysfunctional adipocytes produce excessive amounts of insulin-sensitizing adipocytokines such as adiponectin and lose the capacity to store fat as they become enlarged [13, 131]. The overflowing lipid resulting from these processes will perturb metabolism in muscle, liver, and β-cells - further impairing IR and insulin secretion. In addition, hyperglucogonemia in pancreatic α-cells, increased glucose reabsorption in the kidneys, incretin deficiency and resistance in the gastrointestinal tract, and IR in the hypothalamus all contribute to glucose intolerance [34, 98, 112, 32, 25]. This systemic involvement of dysfunction across multiple tissues further illustrates the complexity of T2D.

1.2 The genetic basis of type 2 diabetes

The fact that the prevalence of T2D has markedly increased over the past few decades emphasizes the considerable impact of environmental factors (i.e. a calorie-rich diet and sedentary lifestyle) [3, 162]. However, there is also strong evidence supporting a pronounced genetic contribution to T2D. Genetic susceptibility is supported by twin studies where monozygotic twins show higher phenotypic concordance than dizygotic twins [108, 101]. Moreover, first degree relatives of affected individuals are at an increased risk for developing T2D [81, 85, 39]. For example, Hemminki et al. (2010) estimated relative risk for family members of diabetic

probands in a large-scale Swedish cohort of over 157,000 hospitalized patients and found a 2-3 fold increase in sibling relative risk [65]. The genetic contribution to T2D is further corroborated by familial estimates of disease heritability - the proportion of phenotypic variance explained by genetic variance - that range from 30-70% [4, 81, 172].

Although these studies support a genetic basis for T2D, the heterogenous nature of the phenotype, variable age of onset, and strong environmental influence greatly complicate genetic studies aimed at elucidating the causal variants and genes underlying genetic susceptibility [106].

1.3 Early gene-mapping studies

The complexity of T2D evinced by its non-Mendelian pattern of inheritance and strong environmental influence, however, did not dissuade researchers from conducting a series of gene-mapping studies that gleaned important insights into the molecular basis of T2D. These efforts included both family-based linkage studies and population-based candidate gene association studies.

Researchers performed genome-wide scans of T2D loci by applying both parametric (involving assumptions of penetrance and mode of inheritance) and non-parametric (based on allele sharing at putative disease loci) linkage studies involving 300-500 highly polymorphic microsatellite markers spaced at roughly 5-10 Mb intervals along chromosomes [90, 58, 114]. These studies provided evidence for several T2D loci indicated by co-segregation of disease with marker alleles within families. A study of Pima Native Americans uncovered evidence for linkage at the 1q21-q25 locus on chromosome 1 that was further corroborated in studies of families of European ancestry [62, 153, 39].

The first published genome-wide scan of T2D was based on a study of 330 affected sibling pairs from Mexican American families in South Texas and revealed a major susceptibility signal at the 2q37 locus [61]. In a follow-up study, Horikawa et al. (2000) applied positional cloning at the locus and determined that the linkage signal was attributable to polymor-

phisms within the *CAPN10* gene that encodes calpain-10 (including a common G → A missense variant) [70]. The calpain family of proteins are calcium-activated neutral proteases that cause activation or inactivation of proteins involved in intracellular signaling, proliferation, and differentiation - including adipocyte differentiation and possibly insulin action and insulin secretion [70]. In a different linkage study of Pima Native Americans, subjects homozygous for the G allele showed decreased rates of post-absorptive and insulin-stimulated glucose turnover - supporting an increased T2D risk through the development of IR [7]. Moreover, Baier et al. (2000) observed that decreased rates of glucose turnover coincided with lower mRNA levels of *CAPN10* in skeletal muscle [7]. Furthermore, Orho-Melander et al. (2002) observed a significant association of *CAPN10* polymorphisms with fasting insulin and homeostasis model assessment insulin resistance index (HOMA-IR) among a cohort of 298 non-diabetic control subjects [113]. They also identified significant associations with elevated FFA levels in both the control group and a cohort of 395 type 2 diabetics, providing further evidence that *CAPN10* risk alleles promote T2D through IR in peripheral tissues [113]. Importantly, the *CAPN10* locus represents the first successful positional cloning of a gene for a polygenic disease like T2D.

As a caveat, genome-wide association studies (GWAS; see next section) have not identified significant variant associations at the *CAPN10* locus. However, this may be due to the fact that genotyping microarrays used for GWAS of T2D poorly interrogate variants at this region [142].

Reynisdottir et. al. (2003) performed a non-parametric multipoint linkage analysis in an Icelandic cohort of 227 families and found evidence of linkage at the 5q34-q35.2 locus on chromosome 5 as well as a signal on chromosome 10 [126]. Grant et al. (2006) performed a follow-up association study at the chromosome 10q locus and found a significant association of the microsatellite marker, DG10S478, within intron 3 of the transcription factor 7-like 2 (*TCF7L2*) gene with T2D - a signal that was replicated in an independent Danish cohort [55]. The *CAPN10* and *TCF7L2* loci provide examples of genetic support for T2D gleaned

from genome-wide linkage studies, however, most loci discovered from such linkage-based investigations failed to replicate in independent studies [58]. Furthermore, the resolution of disease-linked loci is very broad (especially for non-parametric studies) and ranges from 15-30 cM (1 cM is approximately 1 million base pairs), making it difficult to determine causal variants and relevant genes [90]. Family-based linkage studies are well powered for identifying disease-linked loci for diseases that exhibit Mendelian patterns of inheritance involving highly penetrant causal alleles. This approach has therefore been successful for mapping disease loci underlying rare monogenic and syndromic forms of diabetes such as maturity-onset diabetes of the young (MODY) and Wolfram Syndrome [76, 64]. However, linkage studies have largely been underpowered to resolve genetic loci conferring risk for T2D due to a non-Mendelian inheritance, considerable influence of environmental risk factors, and the lack of highly penetrant causal variants with large effects [58].

Researchers have also employed candidate gene association studies to elucidate genetic risk factors for T2D. As opposed to family-based linkage studies, candidate gene association studies employ a hypothesis-driven approach where variants within genes suspected of playing a role in disease-relevant processes are genotyped and tested for association in a genetically unrelated cohort of case and control subjects. Alleles that segregate at higher frequencies among diabetic individuals are considered risk alleles whereas alleles that are significantly more common among controls are considered to be protective [119]. Moreover, causal variants segregating within a population may either be directly tested when genotyped or indirectly interrogated through non-random association with a marker allele (i.e. linkage disequilibrium or LD).

Although this approach provides more power to detect variants with smaller effects than that afforded by linkage studies, it requires a biologically grounded hypothesis for each gene tested that may be untenable for diseases with highly polygenic architectures. However, the application of such studies have revealed significant associations at coding variants for genes encoding peroxisome proliferator (*PPARG*) and the β-cell ATP-sensitive potassium

(K_{ATP}) channel Kir6.2 (*KCNJ11*) - both of which are targets for therapeutic agents (thiazolidinediones and sulfonylureas, respectively) [5, 40]. Coding mutations in *KCNJ11* and *ABCC8* - which encodes the sulfonylurea receptor 1 (SUR1) subunit of the β-cell K_{ATP} channel - have previously been shown to cause a rare form of diabetes that affects newborns known as neonatal diabetes (NND) [130, 54]. The pancreatic β-cell K_{ATP} channel encoded by these genes plays an essential role in insulin secretion as activating mutations in these genes prevent glucose-stimulated insulin secretion from the pancreas [130, 54]. Although the *KCNJ11* variants associated with T2D differ from those responsible for NND, it highlights a molecular process that plays a critical role in maintaining glucose homeostasis that promotes the development of diabetes when disrupted by genetic variation.

1.4 Genome-wide association studies reveal biological insights into T2D

The modest number of successes gleaned from family-based linkage and candidate gene studies of T2D can be attributed to the complexity of the disease itself; T2D is obfuscated by environmental factors, non-Mendelian inheritance, and a paucity of highly penetrant causal variants. However, a series of large-scale, collaborative efforts to elucidate the structure of human genetic variation along with advances in genotyping technology set the stage for genome-wide association studies. Beginning with the inception of the Human Genome Project in the mid-1980s and carrying forward to population studies of commonly segregating variants (i.e. The International HapMap Project and 1000 Genomes Project) - researchers delineated a framework that allowed them to surpass the power limitations of genome-wide linkage scans [145, 77].

Genome-wide association studies (GWAS) extend the study design of candidate gene association studies and involve testing thousands to millions of variants - predominantly single nucleotide polymorphisms (SNPs) - that span the entire human genome [154]. Population-level

haplotype maps garnered from the International HapMap consortium allowed researchers to exploit haplotype structure and delineate parsimonious sets of marker alleles that "tag" variants through LD (including putative causal variants) [77, 154]. During the first 2 years of GWAS, over 13 loci revealed significant associations with T2D - considerably more than previous decades of genetic studies - and have now culminated in over 90 loci that harbor significant associations with T2D [15, 104, 125]. Moreover, many T2D-associated variants have replicated in multiple studies and have revealed important insights into the genetic architecture of T2D.

For example, researchers have gleaned insight into consequences of risk alleles by comparing T2D-associated loci with results from GWAS of related traits. Obesity is a major risk factor for T2D and T2D-associated SNPs annotated to *FTO* and *MC4R* also significantly associated with obesity - indicating the these variants promote T2D through obesity-associated IR [46, 93]. Moreover, researchers have observed an overlap between T2D-associated loci and loci associated with physiological variation in glycemic traits [15]. Voight et al. (2010) observed that risk alleles at 10 of 32 mapped T2D loci also associated with reduced β-cell function (*CAMK1D*, *CDKAL1*, *CENTD2*, *HNF1B*, *IGF2BP2*, *KCNQ1*, *MTNR1B*, *SLC30A8*, *TCF7L2*, *THADA*) [157]. However, only 3 T2D loci showed association with reduced insulin sensitivity (*FTO*, *KLF14*, and *PPARG*) [157].

Despite the fact that over 50 loci have been associated with fasting and 2-hour glucose levels, only a subset overlap T2D-associated loci; 22 out of 36 loci associated with fasting glucose are directionally consistent (i.e. the disease promoting allele for T2D also associates with higher fasting glucose) and 4 out of 9 loci associated with plasma glucose following an oral glucose tolerance test were consistent [35, 134]. Moreover, directional divergence was observed for variants mapped to *MTNR1B*, *G6PC2*, and *GCK* [134, 104]. Notably, *GCK* encodes the pancreatic "glucose sensor" glucokinase that plays an integral role in the glucose-stimulated insulin response and has been previously mapped as a gene underlying a form of monogenic diabetes (i.e. MODY 2) [50]. The inconsistent direction of association at

these genes and partial overlap of T2D and glycemic trait loci indicates that although some variants promote elevated glucose levels and T2D risk, the genetic basis of T2D is influenced by variation that does not affect glucose homeostasis *per se*.

GWAS has also informed the relationship between dyslipidemia and T2D. Free fatty acids (FFA) have been shown to promote IR in liver and have lipotoxic effects in pancreatic β-cells [25, 151, 109]. Moreover, researchers have observed that elevated circulating triglyceride levels predict T2D independently of BMI and shown that triglyceride-lowering drugs may reduce T2D risk by improving insulin secretion [44, 25]. Although these observations suggest that risk variants may promote T2D by directly increasing triglyceride levels, De Silva et al. (2011) applied a Mendelian randomization-based study design and found no relationship between T2D rates and number of triglyceride-raising alleles [22]. These results therefore suggest that hypertriglyceridemia is a consequence rather than a cause of T2D [22].

Researchers have used results from GWAS to address the "Fetal origins" hypothesis that explains the epidemiological association between low birthweight and increased T2D risk later in life [10, 1]. Genetic variants in *ADCY5* and near *CCNL1* are not associated only with relative insulin deficiency but also with fetal growth and birthweight [10, 47]. Moreover, variants increasing T2D risk at additional T2D-associated loci near *CDKAL1*, *HHEX*, and *KCNQ1* are also associated with smaller birthweight [47, 71]. On the other hand, T2D risk alleles at the *GCK* and *MTNR1B* loci were associated with increased birthweight [47]. However, these associations were no longer significant when the researchers adjusted for the maternal genotype - indicating that the risk alleles promote hyperglycemia in the mother and subsequent hyperinsulinemia in the fetus [47]. In addition to environmental factors, these results collectively support a genetic contribution to intrauterine growth that increases T2D risk [102].

GWAS has also allowed researchers to confront questions pertaining to the population genetics of T2D risk alleles. For example, in 1962 J.V. Neel proposed a "thrifty genotype" hypothesis wherein the high prevalence of T2D and obesity alleles in contemporary populations

indicates that they may have conferred a selective advantage in human prehistory; variants promoting energy storage and weight gain may have been advantageous during periods of erratic food supply [107].

The "thrifty genotype" hypothesis would be corroborated by the presence of many derived alleles over ancestral alleles (i.e. selective sweeps) or marked differences in allele frequencies between ethnic groups [8]. However, there is little evidence of these signatures among T2D-associated variants as selection seemingly affected both protective and risk alleles in equal proportion at 65 loci significantly associated with T2D [6, 158]. Yet, there is a small but growing number of T2D loci where replication is not observed across ethnicities and could potentially be attributable to true population-specific risk alleles as opposed to differences in LD or being false positives [14, 20, 125].

1.5 The problem of "missing" heritability

There have also been a number challenges brought to light by GWAS - the most salient being that significant trait-associated variants present only modest effect sizes. For example, the most strongly associated SNP with T2D (at the *TCF7L2* locus) has an odds ratio (OR) of only ~ 1.4 in Europeans [55]. Moreover, the set of SNPs that reach genome-wide significance in GWAS explains only 5-10% of the phenotypic variance of T2D [15, 79]. This presents a problem of "missing" heritability where the trait heritability accounted for by GWAS-significant SNPs is considerably lower than that estimated from family studies [95]. Furthermore, this problem is not unique to T2D as GWAS of height - an extensively studied human phenotype associated with a high estimate of heritability ($\sim$ 80%) - has mapped variants at 180 loci yet account for only a combined 10% of trait heritability [86]. There have been several explanations put forward to address this discrepancy that range from a failure to account for interactions between SNPs (i.e. epistasis) to the possibility that family-based heritability estimates did not adequately adjust for shared environment and were therefore inflated [95, 15]. Additionally, GWAS has predominantly interrogated associations involving

SNPs due to the reliability and coverage of SNPs in genotyping platforms. Consequently, GWAS has interrogated structural variation to a lesser extent (although an appreciable proportion of common structural variants can be "tagged" by marker SNPs through LD) [103]. Therefore, heritability estimates from GWAS-significant SNPs may not fully reflect the impact of copy number variants (CNVs) (i.e. insertions and deletions) and copy neutral variants (i.e insertions and translocations) - variants that have been shown to significantly contribute to Mendelian disorders [95, 82].

Another widely considered explanation for missing heritability is that genotyping microarrays - although well-powered to detect associations for common variants segregating in populations with minor allele frequency (MAF) $> 1\%$ - poorly interrogate rarer variants (MAF $< 1\%$) with larger effect sizes [95, 52]. Purifying selection would work to purge highly deleterious alleles from populations, making it unlikely for there to be common variants with high effect sizes, yet it would not necessarily preclude the existence of rare variants with larger effects at loci harboring T2D associations [15]. Moreover, low frequency variants could have considerable effect sizes without showing a Mendelian pattern of inheritance and account for a sizeable proportion of phenotypic variance [100]. For example, 20 variants with allele frequency of 1% and odds ratio of 3 would theoretically account for most of the familial aggregation of T2D [100]. However, such variants were not detected in a large and sufficiently powered meta-analysis of over $6,000$ families [58]. Moreover, results from sequence-based association studies of T2D have not accounted for sizable proportions of "missing" heritability [91, 160].

Alternatively, there could be many common variants with relatively small effect sizes that may collectively account for much of the heritability not accounted for by GWAS-significant variants [95, 78]. The significance threshold ($p < 5 \text{ x } 10^{-8}$) employed in GWAS corrects for multiple testing [100, 154]. However, this threshold is overly stringent as it assumes independence between tests whereas the presence of LD between SNPs belies this assumption. Consequently, this would result in fewer SNPs reaching genome-wide significance and there-

fore contribute to the low proportion of explained variance observed for complex traits such as T2D. In 2010, Yang et al. developed a method for estimating heritability attribute to common genetic variation [78]. In contrast to family-based methods for estimating heritability that leverage expected or estimated genetic relatedness between family members [143], the Yang-Visscher method directly estimates genetic relatedness from GWAS data - where the ascertainment scheme involves unrelated case and control groups [78, 167]. This method allows for an unbiased estimate of narrow sense heritability (h^2 - the proportion of phenotypic variance explained by additive genetic effects). However, as the SNPs interrogated in GWAS typically segregate in populations at MAF $> 1\%$, this method provides an estimate for the phenotypic variance explained by the additive effect of common SNPs - a measure referred to as $h2_{\text{SNP}}$ or $h2_{\text{chip}}$ heritability. The application of this method to a dataset representing $3,925$ individuals of European ancestry at nearly $300,000$ SNP markers yielded an estimate of height heritability of 45% [78]. Further adjustment for imperfect LD between causal and marker variants due to differences in allele frequency distributions increased this estimate to $\sim 54\%$; an estimate considerably larger than those corresponding to "bottom-up" estimates of heritability attributable to GWAS-significant variants [78, 171].

Researchers have applied the Yang-Visscher method and similar approaches to show that common SNPs can account for significant proportions of phenotypic variance for numerous quantitative and disease traits such as body mass index (BMI), von Willebrand factor, QT interval, schizophrenia, and bipolar disorder [169, 123, 87, 21]. These results indicate that much of what might be considered "missing" heritability for a number of complex diseases may actually be "hidden" in the cumulative effect of many common SNPs with small effect sizes that fail to meet the stringent significance threshold applied in GWAS.

Indeed, researchers have estimated that 25-63% of the phenotypic variance of T2D in populations of European ancestry from Iceland and the UK can be explained by common variants represented in GWAS [104, 172, 140]. However, these results do not explain all the phenotypic variance of T2D and suggest that rare variants may constitute the remaining

heritability. Moreover, the fact that many common SNPs with small effect sizes explain a sizable fraction of heritability complicates the interpretability of results from GWAS; this observation makes it all the more difficult to resolve the causal variants and relevant genes that underlie disease risk.

1.6 The role of regulatory genetic variation on genetic risk of T2D

Another issue brought to light from GWAS on numerous complex traits and diseases is that the majority ($\sim$ 93%) of associated variants are located in non-coding genomic (i.e. intronic and intergenic) regions rather than regions encoding genes [67]. This suggests an important role for variants that regulate gene expression as opposed to variants that alter protein structure. Indeed, researchers have generated substantial evidence supporting a relationship between trait-associated variants and variants that influence chromatin accessibility and gene expression [99, 110, 111]. Many studies have mapped and investigated the phenotypic consequences of a class of genetic variation known as expression quantitative trait loci (eQTLs) - variants that show a strong statistical association with gene expression in a particular cell or tissue type [2]. Nicolae et al. (2010) showed that trait-associated variants from published GWAS were significantly enriched for eQTLs mapped in human lymphoblastoid cell lines (LCLs) [111]. Moreover, the observed enrichment were robust across eQTL association thresholds (with the enrichment being more pronounced for more significant eQTLs) and across multiple phenotypes [111]. Notably, sets of GWAS-significant SNPs for diseases with a recognized autoimmune etiology (i.e rheumatoid arthritis, Crohns disease, type 1 diabetes) exhibited greater enrichment for LCL eQTLs than did diseases where LCLs were less likely to be a relevant cell line (i.e. T2D) [111].

Maurano et al. (2012) similarly showed that trait-associated variants were enriched within deoxyribonuclease 1 (DNase1) hypersensitive sites (DHS) - chromatin accessible regions likely

to be actively involved in gene regulation [99]. They generated genome-wide DHS maps in 349 cell and tissue samples from the ENCODE and Roadmap Epigenomics Program and in 233 fetal tissue samples and identified approximately 3.9 million DHS across the genome in one or more tissues [99]. These DHS were highly enriched for significant GWAS SNPs corresponding to 207 and 447 quantitative and disease traits, respectively [99]. Furthemore, they observed that 76.6% of all non-coding GWAS SNPs were either located within a DHS or were in complete LD with SNPs in a nearby DHS [99]. Degner et al. (2012) provided additional evidence linking DHS regions to eQTLs by mapping loci that associated with DNase 1 sequencing (DNase1-seq) reads in 70 Yoruba LCLs corresponded to individuals of West African ancestry [27]. In addition to observing that such dsQTLs were enriched within transcription factor (TF) binding sites, they found that 16% of dsQTLs were also eQTLs and that, conversely, 55% of eQTLs were dsQTLs [27]. Moreover, although 23% of variants mapped as being both a dsQTL and eQTL were within 1 Kb of the transcription start site (TSS) of target genes, they also observed significant evidence for long-range regulation as far as 100 Kb away from the target gene [27].

Although T2D-associated variants were not highly enriched for eQTLs mapped in LCLs [111, 14], Below et al. (2010) found that eQTLs mapped in insulin-responsive peripheral tissues (skeletal muscle and adipose tissue) were enriched among the set of variants that associated with T2D in a GWAS of Mexican American individuals from South Texas [14]. This corroborates a tissue-specific context for regulatory variation wherein eQTL enrichment among sets of disease-associated variants reflects tissues relevant to pathogenesis. An advantage of leveraging eQTL in disease studies is the fact that eQTLs "spotlight" target genes and may help resolve the regulatory mechanism between causal SNPs and target genes implicated by GWAS. For example, T2D risk alleles upstream of *KLF14* (Morris 2012, Teslovich 2010) were also significantly mapped as eQTLs in adipose tissues, supporting a connection with their shared association in measure of insulin resistance [104, 144, 83]. Moreover, these same risk alleles also influenced the expression of other genes across the genome, constituting

distal-acting eQTLs that regulate gene expression through a trans (i.e. non-allele-specific) mechanism [83, 2]. This presumably results from the role of *KLF14* as a transcription factor that may control a coordinated network of *KLF14*-mediated transcriptional activation and suppression in adipose tissue [137]. Furthermore, genome-wide maps of regulatory elements in pancreatic islets (e.g. TF binding sites, open chromatin, putative enhancer elements) reveal that such regulatory regions are highly enriched for T2D-associated variants from GWAS [116, 118].

The substantial evidence linking regulatory variants (i.e. eQTLs) to disease-associated loci illuminates the genetic architecture of complex human diseases - including T2D. Importantly, insights gleaned from functional studies of disease-associated variants challenge the validity of implicating putative disease genes by proximity to associated loci. This problem is exemplified by molecular studies elucidating mechanistic consequences on gene expression. A seminal study of the *FTO* locus harboring the strongest GWAS association with obesity [46], showed that the associated variants within the first intron of *FTO* exhibited long-range interactions with the promoter of a downstream transcription factor - *IRX3*. Moreover, transgenic *in vivo* studies in mice revealed that perturbing the expression of *Irx3* directly affecting body mass composition and response to a high-fat diet [138]. Therefore, the reported disease gene at this locus (*FTO*) may not be the primary target gene impacted by the risk alleles for obesity and serves as a cautionary tale for interpreting GWAS results. Moreover, although GWAS provide reliable maps of T2D-associated loci, results from this and related studies warrant a careful consideration of the putative T2D genes reported from GWAS [138, 105].

1.7 Leveraging eQTL information to elucidate the genetic architecture of T2D

The enrichment of eQTLs mapped in insulin-responsive peripheral tissues among sets of T2D-associated variants suggests an important role for gene regulation in the genetic basis of T2D. In this body of work, I address outstanding questions about the overall contribution of regulatory variation to genetic susceptibility for T2D and the resolution of genes mapped as putative T2D genes. I employ heritability partitioning to determine the contribution of eQTLs mapped in human adipose tissue and skeletal muscle tissue to the proportion of T2D heritability attributable to common genetic variation. I then address the issue of gene mapping by predicting the genetic component of gene expression from eQTL data and sequentially test for association with T2D using a method that explicitly addresses the mechanism of transcription.

CHAPTER 2

PARTITIONING THE HERITABILITY OF TYPE 2 DIABETES WITH EQTLS MAPPED IN INSULIN-RESPONSIVE PERIPHERAL TISSUES

The following chapter was published in the American Journal of Human Genetics, 2014, Volume 95, Issue 5, Pages 521-534. The authors were as follows: Jason M. Torres, Eric R. Gamazon, Esteban J. Parra, Jennifer E. Below, Adan Valladares-Salgado, Niels Wacher, Miguel Cruz, Craig L. Hanis, and Nancy J. Cox

2.1 Abstract

Top signals from genome-wide association studies (GWAS) of type 2 diabetes (T2D) are enriched with expression quantitative trait loci (eQTLs) identified in skeletal muscle and adipose tissue. We therefore hypothesized that such eQTLs might account for a disproportionate share of the heritability estimated from all SNPs interrogated through GWAS. To test this hypothesis, we applied linear mixed models to the Wellcome Trust Case Control Consortium (WTCCC) T2D data set and to data sets representing Mexican Americans from Starr County, TX, and Mexicans from Mexico City. We estimated the proportion of phenotypic variance attributable to the additive effect of all variants interrogated in these GWAS, as well as a much smaller set of variants identified as eQTLs in human adipose tissue, skeletal muscle, and lymphoblastoid cell lines. The narrow-sense heritability explained by all interrogated SNPs in each of these data sets was substantially greater than the heritability accounted for by genome-wide-significant SNPs ($\sim$ 10%); GWAS SNPs explained over 50% of phenotypic variance in the WTCCC, Starr County, and Mexico City data sets. The estimate of heritability attributable to cross-tissue eQTLs was greater in the WTCCC data set and among lean Hispanics, whereas adipose eQTLs significantly explained heritability among Hispanics with a body mass index $\leq$ 30. These results support an important role for

regulatory variants in the genetic component of T2D susceptibility, particularly for eQTLs that elicit effects across insulin-responsive peripheral tissues.

2.2 Introduction

Numerous family-based studies of disease heritability have indicated that type 2 diabetes (T2D [MIM 125853]) is strongly heritable [121, 81, 108, 94, 4]. These results have motivated many large-scale linkage, candidate-gene, and genome-wide association studies (GWAS), which together have identified over 70 variants that significantly associate with T2D or glucose traits and replicate across studies [126, 70, 53, 45, 5, 159, 136, 157, 173, 133, 79]. However, this set of variants collectively explains only a fraction of the narrow-sense heritability previously estimated from family studies (31%-69%) [4] and thereby constitutes a problem of "missing" heritability [79, 15, 95, 75]. Several hypotheses, including the possibility that genetic susceptibility to T2D is driven by rare variants that are poorly interrogated through GWAS [37, 31, 127, 51, 57], have since been put forward to explain the missing heritability of T2D. Consequently, a number of large-scale exome and whole-genome sequencing efforts to identify such rare variants underlying T2D risk within specific populations are ongoing [91, 73].

Alternatively, the majority of "missing" heritability might instead be "hidden" and explained by the cumulative effect of many variants with small effect sizes [123, 155, 51, 115]. Recently, investigators have tested this hypothesis by applying linear mixed models (LMMs), which treat individual SNP effects as random effects and allow for the estimation of total additive genetic variance [169, 155, 87, 88]. The application of these models has provided evidence that GWAS-interrogated variants explain sizable portions of the missing heritability for several complex traits, such as height, fasting glucose, and Tourette syndrome (MIM 137580) [169, 87, 152, 21]. Indeed, recent studies have suggested that much of the missing heritability of T2D can be accounted for by the cumulative effect of common variants effectively interrogated through GWAS and therefore support a highly polygenic architecture for

T2D susceptibility [140, 104]. This presents a major challenge in identifying disease-related genes given that the vast majority of interrogated SNPs are in non-coding regions and weakly influence disease risk [15, 127]. Although T2D heritability might be attributable to the combined action of many variants, we hypothesize that the variants driving susceptibility are concentrated in pathways that influence molecular processes within metabolically important tissues.

Previous studies have shown that expression quantitative trait loci (eQTLs) are highly enriched within sets of disease-associated variants [111, 66, 99]. Moreover, the extent of enrichment is greater for eQTLs mapped in tissues relevant to the disease [111]. Importantly, eQTLs mapped in human adipose and skeletal-muscle tissue are enriched among sets of T2D-associated SNPs, whereas eQTLs mapped in human lymphoblastoid cell lines (LCLs) exhibit no such enrichment [111, 14, 38]. It is thus reasonable to hypothesize that at least some of the overall heritability of T2D might be attributable to variants that regulate gene expression in insulin-responsive peripheral tissues (IRPTs).

In this study, we applied LMMs to the Wellcome Trust Case Control Consortium (WTCCC) GWAS data set for T2D and to data sets including Mexican Americans from Starr County, TX, and Mexicans from Mexico City to investigate the T2D variance explained by the combined effect of common variants with individually small effects across data sets including populations with distinct ancestries. Furthermore, we leveraged information on eQTLs mapped in human adipose tissue, skeletal muscle, and LCLs to elucidate the heritability contribution from regulatory variants mapped in IRPTs.

Here, we report that for each GWAS data set, the additive effect of the total set of interrogated SNPs accounts for more heritability than estimates previously reported for GWAS-significant variants. SNP subsets composed of eQTLs mapped in skeletal muscle and subcutaneous adipose tissues and of eQTL subsets identified in multiple tissues explain higher phenotypic variance than expected given the proportion of SNPs included in the analysis. These results suggest that T2D has a highly polygenic architecture that is disproportionately

driven by regulatory variants, including those active in IRPTs.

2.3 Materials and Methods

2.3.1 Ethics Statement

All study participants who donated adipose and muscle biopsies provided written informed consent under protocols originally approved by the institutional review board at the University of Arkansas for Medical Sciences. Informed consent was obtained from each participant from Mexico City, and research was approved by the ethical research boards of the Medical Center "Siglo XXI" and the University of Toronto. Informed consent was obtained from each participant from the Starr County Health Study, and research was approved by the institutional review boards of the University of Texas Health Science Center at Houston.

2.3.2 GWAS and eQTL Data Sets

We analyzed genotype data from the WTCCC GWAS of T2D in $1,924$ case and $2,938$ control subjects [159]. The data set included 469,557 SNPs genotyped on the Affymetrix GeneChip 500K Mapping Array Set and was called with the CHIAMO algorithm. SNPs showing significant departures from Hardy-Weinberg equilibrium (HWE) could inflate estimates of narrow-sense heritability [88]. We therefore excluded $33,618$ markers with p values $< 5\%$ from a HWE exact test. An additional set of $63,236$ SNPs with a minor allele frequency (MAF) $< 1\%$ were excluded, resulting in a pruned set of $370,139$ SNPs that were carried forward in our estimation of phenotypic variances explained by GWAS-interrogated SNPs. Moreover, we ensured that SNPs with genotype missingness $> 1\%$ were not included in our analyses, and therefore only SNPs with < 20 missing genotypes were included.

We also analyzed two GWAS data sets of Hispanic populations of Mexican descent in this study. The first data set was previously used in a GWAS [14] of T2D in a Mexican American population from Starr County, TX (SCT data set), and consisted of 837 case and

781 control subjects (Supplementary Table 2.1). As previously described [14], individuals were genotyped on the Affymetrix Genome-Wide SNP Array 6.0, and the set of typed SNPs was imputed to a HapMap2 reference panel, resulting in a set of ~ 1.8 million SNPs after quality-control measures. A total set of $1,733,064$ SNPs remained after we removed SNPs with HWE-departure p values $< 5\%$ and SNPs with MAF $< 1\%$.

The second Hispanic GWAS data set came from a Mexican population from Mexico City (MCM data set) and consisted of 967 case and 343 control subjects (Supplementary Table 2.1). Individuals were genotyped on the Affymetrix Genome-Wide SNP Array 5.0, and the set of typed SNPs was imputed to both a phased HapMap2 and an unphased HapMap3 reference panel as previously described [117]. We applied the same quality-control measures we used for the WTCCC and SCT data sets to the MCM data set, and the remaining set of $2,431,591$ SNPs was carried forward for heritability analyses (Supplementary Table 2.2). In order to increase the sample size for estimation of heritability among populations of Mexican descent, we merged the imputed SCT and MCM data sets together by chromosome and applied the same MAF and HWE-departure p value thresholds as in the individual GWAS data sets. This resulted in a data set including $1,652,821$ SNP genotypes (Supplementary Table 2.2) for a total of $2,928$ subjects ($1,804$ case and $1,124$ control individuals). We applied a stringent genotype missingness threshold of 1% in the merged Hispanic data set (and thus SNPs with > 20 missing genotypes were not included in our analyses).

A set of eQTLs mapped in 90 HapMap CEU (Utah residents with ancestry from northern and western Europe from the CEPH collection) lymphoblastoid cell lines (LCLs) were acquired from the SNP and CNV Annotation (SCAN) database, an online repository that provides physical and functional annotations for SNPs, copy number variants, and genes [49]. As previously described [49], we mapped eQTLs in LCL samples by applying a family-based quantitative transmission-disequilibrium test (QTDT) to a set of more than two million SNPs with MAF $> 5\%$ in order to evaluate associations with the reliable expression of $> 13,000$ gene transcripts measured with the Affymetrix GeneChip Human Exon 1.0 ST Array. For

this study, we included all SNPs with QTDT p values $< 1.0 \times 10^{-4}$ in the SCAN database.

Extraction and eQTL mapping of primary human adipose and skeletal muscle tissue are described in Elbein et al [38]. In brief, samples of subcutaneous adipose tissue (extracted via needle biopsy or incision) and skeletal muscle from the *vastus lateralis* were extracted by Bergstrom needle biopsy from 62 individuals with ages ranging from 20 to 55 years and body mass index (BMI) ranging from 19 to 42 kg/m2. Of these 62 individuals, 40 were of European American ancestry and 22 were of African American ancestry. Genotyping was performed with the Illumina 1M platform, and RNA-expression profiling was performed with the Agilent Human Whole Genome 4 x 44 array [38]. The set of typed SNPs was imputed to a HapMap2 reference panel. For each tissue, we mapped eQTLs by testing each SNP for the additive effect of allele dosage on covariate-adjusted normalized probe-level expression intensity under a linear model that included the combined set of samples from both European and African American cohorts:

$$P = \mu + R + \epsilon,$$

where P is the normalized probe-level expression value, μ is the expected value of the probe expression intensity, R is the covariate for genetic ancestry (i.e., first principal component), and ϵ is the residual error term. For the eQTL-based analyses, we included SNPs with gene-expression association p values $< 1.0 \times 10^{-4}$.

2.3.3 Estimation of Phenotypic Variance Explained by GWAS SNPs

For each GWAS data set, a genetic relatedness matrix (GRM) was calculated from the full set of autosomal SNPs (i.e., not pruned for LD) with the program Genome-wide Complex Trait Analysis (GCTA), developed by Yang et al [167]. We included only SNPs with MAF $> 1\%$ in our analyses because we were interested in evaluating the heritability attributable to common SNPs.We applied a relatedness threshold ($\hat{\pi}$) to prune pairs of individuals sharing

many genes identically by descent. This ensured that effects from non-genetic factors, such as a common environment shared between close relatives, did not upwardly bias estimates of chip heritability (i.e., the proportion of phenotypic variance explained by the additive effect of GWAS-interrogated SNPs) [88]. The inclusion of close relatives (i.e., monozygotic twins and first-degree relatives) would dominate heritability estimates and yield results similar to those of an additive genetic model even in the absence of a shared environment [156]. When a pair of individuals share more than the specified genetic relatedness, GCTA employs a parsimonious algorithm that selectively removes an individual so as to maximize the number of individuals retained in the GRM [167]. No pair of subjects in the WTCCC data set shared genetic relatedness in excess of 5%, whereas a considerable number of subjects in the Hispanic data sets surpassed this threshold. The primary analyses presented here utilized all the samples in the Hispanic data set. However, the results from our heritability analyses of this data set are robust across relatedness thresholds (see below).

PC analyses (PCAs) [122] were performed for each data set with the PCA method implemented in GCTA. Mexican American and Mexican individuals showing poor clustering in PC cluster plots were identified, and a GRM was recalculated without these individuals for each of the individual Hispanic data sets and the merged Hispanic data set (Supplementary Figure 2.6).

The proportion of narrow-sense heritability explained by the additive effect of GWAS SNPs (i.e., chip heritability) was estimated with the restricted maximum-likelihood estimation (REML) method implemented in GCTA. The actual binary phenotype of T2D status was transformed to an underlying liability model in which the disease threshold was dependent upon the prevalence of T2D in each population [167]. A prevalence of 8% was specified for the REML estimation of additive genetic variance in the WTCCC data set. Population disease-prevalence values of 20% and 10% were specified for the analyses of the SCT and MCM data sets, respectively (Supplementary Table 2.2). A population prevalence of 16% was specified for heritability estimation in the merged His- panic data set and was calculated

as a weighted average of the T2D prevalence from the two samples.

2.3.4 Covariate Selection

LMM models including fixed effects from potential covariates for T2D (i.e., age, BMI, sex, and PCs) were evaluated for each data set with available information (Supplementary Table 2.3). Sex and the first five PCs were included as covariates in the analysis of the WTCCC data set. Sex, age (at diagnosis for case subjects and at enrollment for control subjects), and BMI were included as covariates in the analyses of the Hispanic data sets. Moreover, differences in global ancestry can potentially confound genetic association studies in admixed populations [148]. To control for these effects, we included the first ten PCs as covariates in our estimation of heritability in the SCT, MCM, and merged Hispanic data sets.

2.3.5 Estimation of Phenotypic Variance Explained by SNPs that Colocalize with Known T2D-Associated Variants

A set of 141 unique SNPs significantly associated with T2D was obtained from the National Human Genome Research Institute (NHGRI) online catalog of published GWAS associations. Reference sets of SNPs composed of the 141 NHGRI SNPs and all HapMap2 SNPs within 1 Kb, 10 Kb, 100 Kb, 500 Kb, and 1 Mb of each of the NHGRI SNPs were constructed. For each GWAS data set, multiple joint analyses of narrow-sense heritability involving each NHGRI SNP set were performed. The total set of GWAS SNPs was partitioned into two unique SNP subsets whereby the first subset was composed of SNPs represented in a NHGRI subset (e.g., 1 Kb) and the second subset was composed of the complement set of all SNPs not present in the first subset. A GRM was calculated for each subset with the –make-grm function implemented in GCTA, and the two GRMs were then incorporated into a joint analysis of the heritability attributable to both the NHGRI SNPs and the complement set of non-NHGRI SNPs. The additive genetic variance attributed to each subset was jointly

estimated with the multiple GRM (–mgrm) and REML (–reml) functions implemented in GCTA.

2.3.6 Estimation of Phenotypic Variance Explained by eQTL SNP Subsets

We estimated the variance explained by eQTL sets by using the following mixed-effects model:

$$Y = Xb + \sum_{T} g_T + C + e$$

$$\text{var(Y)} = \sum_{T} A_T \sigma_T^2 + A_C \sigma_C^2 + I\sigma_e^2,$$

where Y is a vector of phenotypes and b is a vector of fixed effects (e.g., PCs, BMI, etc.). Here, A_T is the GRM estimated from the eQTL set T, and g_T denotes the genetic effect attributable to the eQTL set T with variance of g_T equal to $A_T\sigma_T^2$. C is the genetic effect of the complement set of SNPs. The phenotype is modeled as the sum of these genetic effects and the relevant covariates (fixed effects) and a residual. Variances were estimated with REML. The heritability attributable to the eQTL set T was then calculated as the fraction of the phenotypic variance σ_Y^2:

$$h_T^2 = \frac{\sigma_T^2}{\sigma_Y^2}$$

In our analysis, we estimated the heritability explained by various eQTL sets (derived from several tissues) by partitioning, in multiple analyses, the genome into disjoint sets (composed of classes of eQTLs and the complement SNP set). We did not prune SNP subsets for LD because this could result in spurious "enrichment" of heritability (see Results).

2.3.7 Description of eQTL Subsets Evaluated in Heritability Analyses

For convenience, we provide a brief description of eQTL subsets we evaluated in our partition analyses of T2D heritability (Figure 2.1): eQTLs mapped in LCL samples from HapMap2 CEU individuals, but not included in either the adipose or the skeletal-muscle eQTL sets (L); eQTLs included in the adipose eQTL reference set, but not in any other eQTL reference set (A); eQTLs in the skeletal-muscle eQTL reference set, but not in any other eQTL reference set (M); eQTLs included in both the LCL and the adipose eQTL reference sets, but not in the skeletal-muscle eQTL reference set (AL); eQTLs present in both the LCL and the skeletal-muscle eQTL reference sets, but not in the adipose eQTL reference set (ML); eQTLs represented in both the adipose and the skeletal- muscle eQTL reference sets, but not in the LCL eQTL reference set (AM); the intersection of eQTLs represented in the LCL, adipose, and skeletal muscle reference sets (AML); and the set of eQTLs mapped in IRPTs, defined as the union of eQTL sets A $\cup$ M $\cup$ AM.

Baseline-Subset Analysis. For this analysis, the total set of SNPs was partitioned into eight SNP subsets representing all possible disjoint eQTL sets (Figures 2.1A and 2.1B): L, A, M, AL, ML, AM, AML, and the complement set of SNPs not included in the eQTL subsets.

Analysis of IRPT versus LCL eQTLs. We sought to compare the heritability for regulatory variants identified in IRPTs as a group (A $\cup$ M $\cup$ AM) to that from eQTLs mapped in LCLs, a representative non-insulin responsive cell type. Moreover, we were interested in comparing these estimates to the heritability from eQTLs mapped across tissues (AML) in this analysis. The total set of SNPs in each GWAS data set was partitioned into non-overlapping subsets: L, IRPT, AML, and the complement set of SNPs not included in these eQTL subsets.

Expanded Analysis of IRPT versus LCL eQTLs. This analysis was similar to the analysis of IRPT versus LCL eQTLs (IRPT-LCL analysis) with the distinction that the

IRPT set was ungrouped into the A, M, and AM subsets. Alternatively, this analysis was similar to the baseline-subset analysis with the modification that the AL and ML subsets were included in the complement SNP subset. This analysis, unlike the IRPT-LCL analysis, allowed for a comparison of tissue-specific eQTLs while involving fewer subsets than the baseline-subset analysis.

Index-Subset Analysis. In order to further investigate relationships between eQTL subsets that significantly explain heritability, we performed a series of three-way partition analyses whereby we jointly estimated heritability for two "index" subsets along with the complement set of SNPs not included in either of the index subsets. For example, in the WTCCC analyses (see below), we observed that the muscle-specific and cross-tissue eQTL subsets explained disproportionate phenotypic variances relative to SNP-set proportions (Figure 3). We subsequently performed an index-subset analysis by grouping all cross-tissue eQTLs subsets (AL $\cup$ ML $\cup$ AM $\cup$ AML) and comparing this subset with the muscle-specific (M) and complement SNP subsets.

2.3.8 Estimation of Phenotypic Variance Explained by cis- and trans-eQTLs

In order to further delineate the eQTL subsets that explained the most phenotypic variance relative to SNP-set proportion in the WTCCC data set (i.e., cross-tissue and muscle-specific eQTLs; see Results), we partitioned according to *cis* and *trans* classification. eQTLs in each reference data set were designated as either *cis* or *trans* according to proximity to either the gene start site or the gene end site of the target transcripts. eQTLs within 4 Mb of a start or end site for a target gene were classified as *cis*-eQTLs, whereas eQTLs associ- ated with genes beyond this threshold were designated as trans- eQTLs. SNPs associated with gene expression in multiple genes were classified as *cis* if at least one significant association was a *cis* association. Otherwise, a SNP was designated as a *trans*-eQTL.

Cross-tissue eQTLs were classified as *cis*-eQTLs if at least one association was a *cis* association. The total set of SNPs in the WTCCC data set was partitioned into three unique SNP subsets: cross-tissue *cis*-eQTLs, cross-tissue *trans*-eQTLs, and the remaining set of all SNPs not represented in the other two subsets. As in the eQTL-subset analyses described above, we calculated a GRM for each subset from the corresponding SNP genotypes and used the three GRMs to jointly estimate the heritability attributable to each subset by using REML. We performed a similar partition analysis in the WTCCC data set for skeletal muscle-specific eQTLs.

2.3.9 BMI-Stratified Analysis of the Merged Hispanic Data Set

In order to evaluate T2D-heritability profiles among subjects of Mexican descent in the context of obesity,we stratified the merged Hispanic data set according to BMI whereby subjects with BMI < 30 were included in a "lean" group (997 case and 640 control subjects) and subjects with BMI ≥ 30 were included in a "non-lean" group (779 case and 475 control subjects). We performed two separate heritability analyses for each BMI-stratified group: (1) index-subset analysis (see above) with muscle-specific eQTLs (M) and cross-tissue eQTLs (AL $\cup$ ML $\cup$ AM $\cup$ AML) as the index subsets; and (2) index-subset analysis with adipose-specific eQTLs (A) and cross-tissue eQTLs as the index subsets. The first ten PCs were included as covariates, and the analyses were performed both with and without the inclusion of BMI as a covariate.

2.3.10 Permutation Analysis

In order to identify significant differences in the proportions of phenotypic variance explained by eQTL subsets, we performed a series of permutation-based analyses. For each joint heritability analysis, we investigated statistical differences in heritability estimates between subsets by defining a test statistic:

$$T = \frac{\hat{h}_x^2 - \hat{h}_y^2}{\sqrt{\mathrm{SE}_x^2 + \mathrm{SE}_y^2}}$$

where $\hat{h}_x^2$ and $\hat{h}_y^2$ are the estimates of heritability attributable to SNP subsets x and y, respectively, and SE_x and SE_y are the standard errors for SNP subsets x and y, respectively. We then permuted jointly the phenotype assignment and corresponding covariate values of each individual in the GWAS data set while maintaining the genotype structure. We used the permuted phenotypes to recalculate REML estimates of heritability with SE and obtained a test statistic. We repeated this process multiples times in order to generate a distribution of test statistics and obtained an empirical p value from the number of instances a permutation test statistic equaled or exceeded the observed test statistic from the non-permuted analysis. For each joint heritability analysis in the WTCCC and Hispanic data sets, we performed a minimum of 200 phenotype permutations with REML.

In order to evaluate significant differences in heritability estimates between corresponding eQTL subsets in the lean and non-lean Hispanic cohorts, we defined the following test statistic:

$$T = \frac{\hat{h}_{x,\text{non-lean}}^2 - \hat{h}_{x,\text{lean}}^2}{\sqrt{\mathrm{SE}_{x,\text{non-lean}}^2 + \mathrm{SE}_{x,\text{lean}}^2}}$$

where $\hat{h}_{x,\text{non-lean}}^2$ and $\hat{h}_{x,\text{lean}}^2$ are the estimates of heritability attributable to SNP subset x in the non-lean and lean cohorts, respectively, and $\mathrm{SE}_{x,\text{non-lean}}^2$ and $\mathrm{SE}_{x,\text{lean}}^2$ are the standard errors for SNP subset x in the non-lean and lean cohorts, respectively. An empirical p value was similarly obtained from a distribution of permuted test statistics. For each inter-cohort comparison of eQTL heritability made between the non-lean and lean Hispanic groups, we performed 1,000 permutations.

2.3.11 Adjustment for Imperfect LD at Causal Variants

In order to evaluate the effect of imperfect LD between GWAS-interrogated SNPs and underlying causal variants on the observed relationships between eQTL subsets in our joint partition analyses, we applied a MAF-stratified correction to GRMs as described in Yang et al [169]. The genomic relationship between individuals j and k at causal variants (G_{jk}) is approximated by the relatedness calculated from GWAS-interrogated SNPs (A_{jk}). Yang et al. determined empirically that when the MAF at causal variants is $\leq \theta$, the regression coefficient from the regression of G_{jk} on A_{jk} is:

$$\beta = 1 - \frac{(c + 1/N)}{\text{var}(A_{jk})},$$

where c depends on the MAF distribution of the causal SNPs and $1/N$ represents the sampling error from estimating the genetic relationship from N SNPs. We adjusted the GRM for each eQTL subset in the IRPT-LCL partition analysis with the –grm-adj function in GCTA by varying the c parameter (c = 6.2 x 10^{-6}, 3.4 x 10^{-6}, 1.8 x 10^{-6}, 7.8 x 10^{-7}, and 9.2 x 10^{-9} for θ = 0.1, 0.2, 0.3, 0.4, and 0.5, respectively) and estimated variances with REML.

We also considered the effect of an uneven distribution of local LD across the set of GWAS-interrogated SNPs; regions with high LD are likely to disproportionately contribute to estimates of heritability. Speed et al. proposed estimating an adjusted GRM, $\mathbf{A^*}$, which is comparable to the standard GRM calculated in GCTA ($\mathbf{A} = \mathbf{XX'}/m'$), where column X_j of the standardized genotype matrix ($\mathbf{X}$) for SNP j in SNP set m is replaced by weighted vector $\sqrt{w_j X_j}$ and m' is replaced by $\sum_j w_j$ [139]. We estimated weightings for the entire set of GWAS SNPs with the –cut-weights, –calc-weights, and –join-weights options implemented in the program LDAK (Linkage Disequilibrium Adjusted Kinships). We then calculated adjusted GRMs for each SNP subset in the IRPT-LCL partition analysis with the –cut-kins and –calc-kins functions and applied the –partition-prefix and –partition-number (set

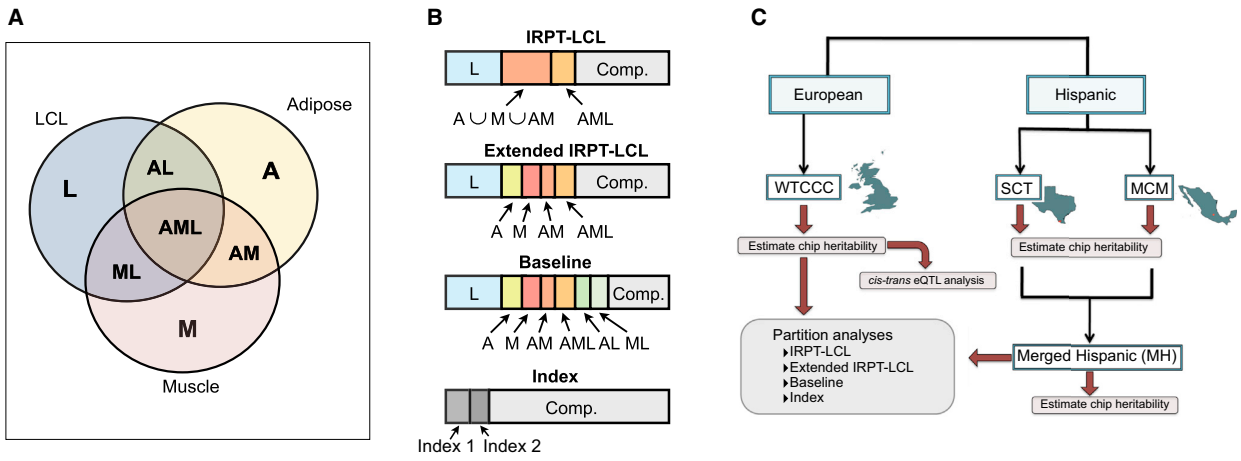

Figure 2.1: **Study overview.** (A) Delineation of the eQTL subsets evaluated in the decomposition of T2D heritability.(B) Diagram showing the constitution of each of the major partitions evaluated in this study. (C) Schematic of the heritability analyses performed on the WTCCC, SCT, and MCM data sets.

to four) arguments to enable genomic partitioning. Estimates of heritability for each subset were calculated through REML implemented in GCTA as described above.

Figures were created in R [124] with base graphics and the ggplot2 package [164].

2.4 Results

2.4.1 Estimation of T2D Heritability Explained by GWAS SNPs

We applied an LMM method to estimate the phenotypic T2D variance explained by the additive effect of GWAS-interrogated SNPs (i.e., chip heritability) in three independent data sets: the WTCCC, SCT, MCM data sets (Figure 2.1A). For each of these populations, the narrow-sense heritability explained by the complete set of GWAS-interrogated SNPs was high and exceeded the range for previously reported GWAS-significant SNPs (Figure 2.2A; Table S2). The SE values of the heritability estimates were inversely related to sample size and thus were larger for the Hispanic data sets (Supplementary Figure 2.8).

To further investigate the heritability contribution from loci strongly implicated from GWAS, we referred to the NHGRI online catalog of published GWAS variants and jointly

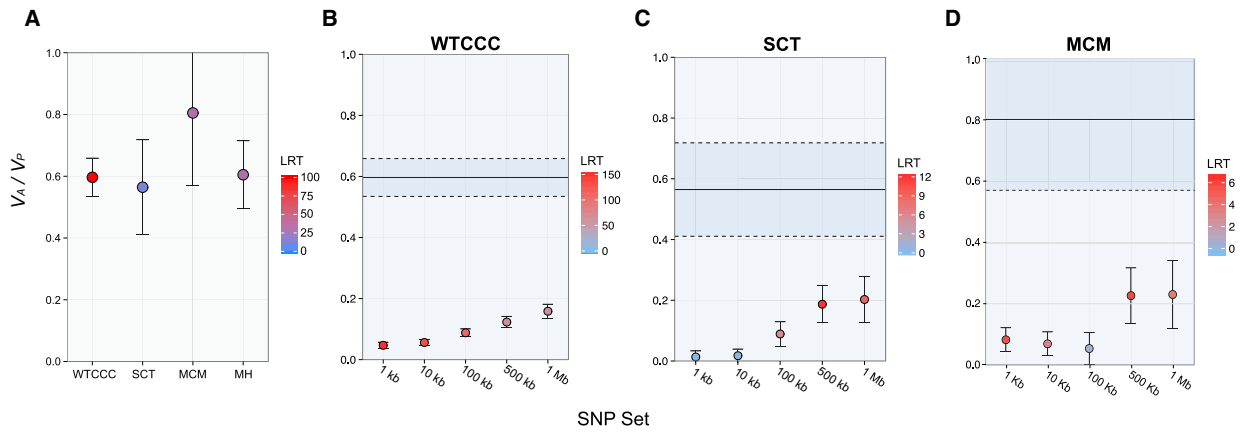

Figure 2.2: **Estimates of narrow-sense T2D heritability explained by GWAS-interrogated SNPs.** (A) The REML estimates of phenotypic variance explained by the additive effect of SNPs interrogated in GWAS (V_A/V_P) on T2D are shown for the WTCCC, SCT, and MCM data sets. (BD) Heritability estimates for SNP subsets composed of T2D-associated variants from the NHGRI GWAS catalog and HapMap2 SNPs within 1 Kb, 10 Kb, 100 Kb, 500 Kb, and 1 Mb are shown for the WTCCC (B), SCT (C), and MCM (D) data sets. Total chip heritability and SE for each GWAS are given by the solid and dashed black lines, respectively. The color corresponds to the significance of each heritability estimate determined by the test statistic from the likelihood-ratio test (LRT).

estimated the variance explained by subsets enriched with T2D-associated SNPs. For each GWAS data set, the significance of heritability estimates corresponding to the NHGRI subsets was high, and the explained phenotypic variance generally increased with the number of SNPs included in each subset (Figures 2.2B-D; Supplementary Table 2.4). However, in each data set, the heritability estimates attributable to subsets composed of T2D-associated SNPs were much lower than the total estimate of chip heritability (Figures 2.2A-D; Supplementary Tables 2.2 and 2.4).

2.4.2 Estimating T2D Heritability Explained by Metabolic-Tissue eQTLs in the WTCCC Data Set

Our group has previously shown that eQTLs mapped in human adipose and skeletal muscle tissue are enriched among the set of variants that show genome-wide significant association with T2D in populations of European and Mexican descent [14, 38]. We therefore hypoth-

esized that eQTLs mapped in metabolically important tissues might also account for much of the heritability of T2D. In order to evaluate this hypothesis, we partitioned the set of GWAS-interrogated SNPs into subsets composed of eQTLs and performed joint analyses of variance components (Figures 2.1A-C). We included eQTLs mapped in two IRPTs, adipose tissue, and skeletal muscle, as well as eQTLs mapped in LCLs (Supplementary Table 2.5).

We performed an IRPT-LCL analysis (see Material and Methods) in order to compare the eQTL-derived heritability in IRPTs (i.e., adipose tissue and skeletal muscle) with that in LCLs, a representative non-insulin responsive cell type. Of the four subsets evaluated in the IRPT-LCL partition, the IRPT, AML, and complement subsets significantly explained phenotypic variance (Figure 2.3A; Supplementary Table 2.6). Moreover, only the IRPT and AML subsets accounted for proportions of explained heritability that exceeded their SNP-set proportions; the AML set accounted for three times more heritability than would be expected from the number of represented SNPs (Figure 2.3B; Supplementary Table 2.6). Interestingly, the LCL subset yielded a low and non-significant estimate of heritability despite the fact that it composed 27% of all GWAS-interrogated SNPs in the WTCCC data set (Figures 2.3A and 2.3B; Supplementary Table 2.6). We performed a permutation-based analysis and observed that the variance explained by the IRPT eQTLs was significantly greater than the variance explained by the L subset ($p < 0.01$).

In order to resolve the tissue-specific contributions to the GWAS heritability estimate, we performed an expanded version of the IRPT-LCL analysis described above whereby we ungrouped the IRPT subset into the constituent A, M, and AM subsets. The subsets composed of muscle-specific eQTLs (M), cross-tissue eQTLs (AML), and complement SNPs significantly explained phenotypic variance, whereas the LCL (L) and adipose (A) subsets yielded low and non-significant estimates of heritability (Figure 2.3C; Supplementary Table 2.7). Of these subsets, only the muscle-specific and cross-tissue eQTL subsets accounted for proportions of heritability that exceeded their SNP-set proportions (Figure 2.3C; Supplementary Table 2.7). Moreover,we found that the difference in heritability estimates between

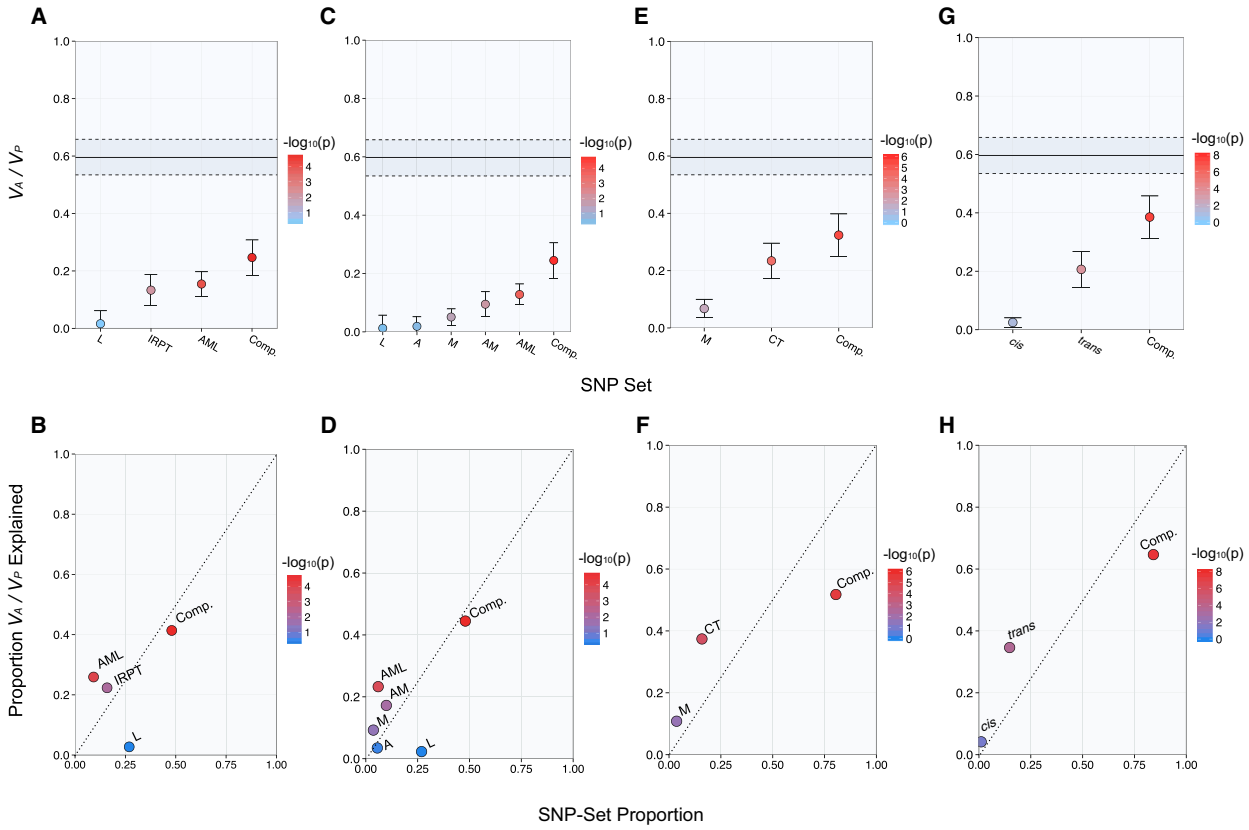

Figure 2.3: **Heritability of T2D Explained by Metabolic-Tissue eQTLs in the WTCCC GWAS Data Set.** The narrow-sense heritability estimates (V_A/V_P) attributable to non-overlapping SNP subsets (top panels). The proportion of chip heritability explained by each subset is plotted with SNP-set proportion (bottom panels). Color is designated by the $\log_{10}$ of the LRT p value, and estimates are shown with SE. (A and B) IRPT-LCL analysis. (C and D) Expanded IRPT-LCL analysis. (E and F) Index-subset analysis with muscle-specific (M) and cross-tissue (CT) eQTLs as index sets. (G and H) *cis-trans* analysis of cross-tissue eQTLs.

the muscle-specific eQTL (M) and adipose-specific eQTL (A) subsets was significant in a permutation-based analysis ($p = 0.025$). However, in a separate analysis, we did not observe a significant difference between the variances explained by the M $\cup$ ML and A $\cup$ AL eQTL subsets ($p = 0.72$; Supplementary Table 2.8).

We performed a baseline-subset analysis to evaluate the phenotypic variance explained by each possible eQTL subset delineated in Figure 2.1B. We observed that each subset composed of cross-tissue eQTLs (i.e., AL, ML, AM, and AML) significantly explained proportions of chip heritability that exceeded their SNP-set proportions (Supplementary Figure 2.8; Supplementary Table 2.9).

We further investigated relationships between eQTL subsets that disproportionately explained phenotypic variance by performing an index-subset analysis (see Material and Methods) whereby we compared two index subsets along with the complement set of SNPs. For this analysis, we evaluated the muscle-specific eQTL subset (M) and a cross-tissue eQTL subset composed of eQTLs mapped in two or more tissues (AL $\cup$ ML $\cup$ AM $\cup$ AML). Each of the evaluated subsets significantly explained phenotypic variance, but only the muscle-specific (M) and cross-tissue eQTL subsets explained proportions of heritability that exceeded their SNP-set proportions (Figures 2.3E and 2.3F; Supplementary Table 2.9). Moreover, the magnitude of this difference for each of these subsets was greater than a factor of 2 (Figure 2.3F; Supplementary Table 2.10). Interestingly, the complement set of SNPs composed 80% of the GWAS-interrogated SNPs yet accounted for only 52% of the explained heritability (Figure 2.3F; Supplementary Table 2.10).

Although the LCL-specific eQTL subset (L) did not significantly explain phenotypic variance in any of the above analyses, we performed an additional joint analysis to resolve the heritability contribution from LCL eQTLs. We estimated the variance explained by the A $\cup$ M $\cup$ AM and AL $\cup$ ML $\cup$ AML eQTL subsets and observed that both subsets significantly explained phenotypic variance and were enriched with heritability (Supplementary Table 2.11). However, a permutation-based analysis showed that the enrichment for the A $\cup$

M ∪ AM subset was not significantly greater than that for the AL ∪ ML ∪ AML subset ($p = 0.18$). This supports a role for cross-tissue eQTLs mapped in LCLs and is consistent with the heritability enrichment observed for the AML subset.

Finally, we considered that the higher heritability attributable to cross-tissue eQTLs in this study might be due to the possibility that an eQTL mapped in two or more tissues can more robustly be classified as an eQTL. This could then explain why cross-tissue eQTL subsets consistently accounted for more heritability than the LCL eQTL sub-sets. However, we do not believe this to be the case given that a comparison of matched sets of highly significant LCL and cross-tissue eQTLs showed that cross-tissue eQTLs, but not LCL eQTLs, yielded significant estimates of heritability (Supplementary Table 2.12). Taken together, these results indicate that subsets composed of muscle-specific eQTLs and cross-tissue eQTLs explain disproportionate amounts of heritability for T2D in the WTCCC data set.

2.4.3 Evaluating the Heritability Captured by cis- and trans-eQTLs in the WTCCC Data Set

The cross-tissue eQTL subset explained the most heritability in the WTCCC data set (see above); we therefore evaluated the respective heritability contributions from cross-tissue *cis*- and *trans*-eQTLs (see Material and Methods). Although the subset of cross-tissue *cis*-eQTLs accounted for a proportion of heritability that exceeded its SNP-set proportion, the REML estimate of variance was low and not significant at the 5% significance level (Figures 2.3G and 2.3H; Supplementary Table 2.13). In contrast, the set of cross-tissue trans-eQTLs, although defined by a loose p value threshold for association with gene expression in each tissue, significantly explained most of the heritability accounted for by the total set of cross-tissue eQTLs (Figures 2.3G and 2.3H; Supplementary Table 2.13).We observed similar results when we performed an analysis of the respective contributions of *cis*- and *trans*-eQTLs represented in the set of eQTLs specific to skeletal muscle (Supplementary Figure 2.10; Supplementary Table 2.14).

2.4.4 Estimating T2D Heritability Explained by Metabolic-Tissue eQTLs in GWAS Data Sets Representing Populations of Mexican Descent

Because heritability-partitioning analyses were unstable in the two Hispanic data sets (Supplementary Tables 2.15-2.17), we combined individuals from both data sets and merged both sets of imputed SNP genotypes to increase the sample size to nearly $3,000$ subjects. The total set of GWAS-interrogated SNPs in the merged data set significantly explained 60% of the phenotypic variance in the merged Hispanic data set (Figure 2.2A; Supplementary Table 2.2).We then performed an IRPT-LCL heritability analysis (see Material and Methods) and found that only the IRPT subset (A $\cup$ M $\cup$ AM) significantly explained phenotypic variance with an enrichment relative to the SNP-set proportion (Figures 2.4A and 2.4B; Supplementary Table 2.18). We performed a permutation-based analysis to determine whether the variance explained by the IRPT subset was significantly greater than the variance explained by the LCL-specific (L) subset. Similar to our results from the corresponding WTCCC analysis, the IRPT eQTL subset significantly explained more variance than the LCL subset ($p < 0.01$).

Next, we performed an expanded IRPT-LCL analysis (see Material and Methods) to further resolve the tissue-specific contribution to heritability in the merged Hispanic data set. Adipose-specific eQTLs and cross-tissue eQTLs mapped in adipose and skeletal-muscle tissue significantly explained phenotypic variance and accounted for most of the estimated heritability (Figure 2.4C; Supplementary Table 2.19). However, muscle-specific eQTLs did not significantly explain phenotypic variance in the merged Hispanic data set (Figures 2.4C and 2.4D; Supplementary Table 2.19), and a permutation-based analysis showed that the enrichment of adipose-specific eQTLs was significantly greater than that of muscle-specific eQTLs ($p < 0.01$).

We performed a baseline-subset analysis (see Material and Methods) and observed that the set of eQTLs mapped in both adipose and skeletal-muscle tissue (AM), but not LCLs, disproportionately explained the most chip heritability relative to the SNP-set proportion

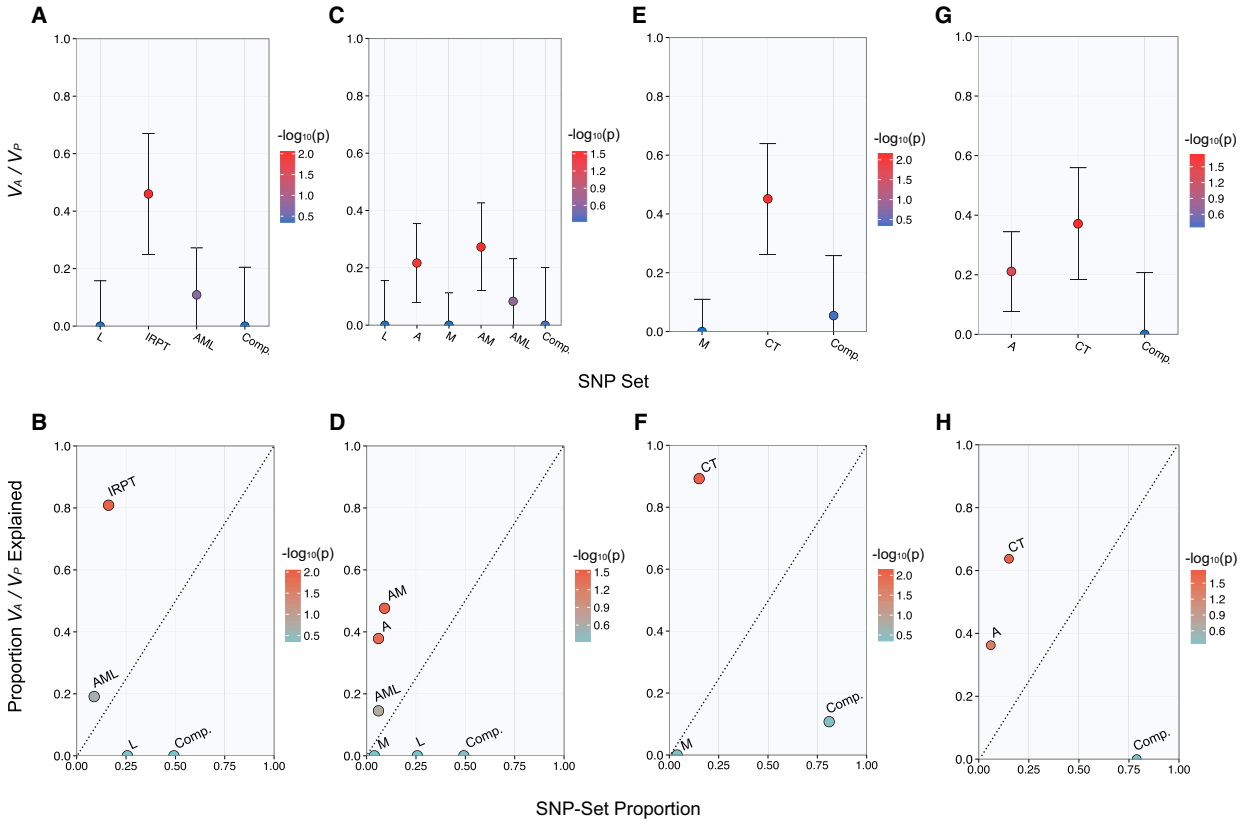

Figure 2.4: **Heritability of T2D explained by metabolic-tissue eQTLs in the merged Hispanic data set.** The narrow-sense heritability estimates (V_A/V_P) attributable to nonoverlapping SNP subsets (top panels). The proportion of chip her- itability explained by each subset is plotted with SNP-set proportion (bottom panels). Color is designated by the $\log_{10}$ of the LRT p value, and estimates are shown with SE. (A and B) IRPT-LCL analysis. (C and D) Expanded IRPT-LCL analysis. (E and F) Index-subset analysis with muscle-specific (M) and cross-tissue (CT) eQTLs as index sets. (G and H) Index-subset analysis with adipose-specific (A) and cross-tissue (CT) eQTLs as index sets.

(Supplementary Figure 2.4; Supplementary Table 2.20). Moreover, we corroborated that adipose-specific (A) and cross-tissue eQTLs, but not muscle-specific (M) eQTLs, significantly and disproportionately explained phenotypic variance in a set of index-subset analyses (Figures 2.4E-H; Supplementary Tables 2.21 and2.22).

We performed an additional index-subset analysis and observed that the A $\cup$ AL subset was significantly enriched with heritability in relation to the M $\cup$ ML subset ($p < 0.01$) (Supplementary Table 2.23). Furthermore, subsets of cross-tissue eQTLs mapped in LCLs did not contribute to heritability in the Hispanic data set, given that the AL $\cup$ ML $\cup$ AML subset did not significantly explain variance and yielded a heritability estimate that was significantly lower than that for the A $\cup$ M $\cup$ AM subset ($p < 0.01$) (Supplementary Table 2.24).

Although the merged Hispanic data set was better powered to resolve heritability estimates between eQTL subsets, the heritability estimate for the complement subset in the IRPT-LCL and expanded IRPT-LCL analyses was non-significant with high SE. However, in a two-way partition analysis of L and complement subsets, the complement subset (including eQTLs mapped in adipose tissue and skeletal muscle) yielded a high and significant estimate of heritability (Supplementary Table 2.25). Taken together, these results differ from the WTCCC results in that they support a greater heritability contribution from subsets composed of eQTLs mapped in human adipose tissue.

2.4.5 BMI-Stratified Analysis of Heritability Explained by eQTL Subsets in the Hispanic Data Set

We observed that 44% of the subjects in the merged Hispanic data set had a BMI $\geq$ 30. Therefore, the fact that more heritability was attributable to adipose-specific eQTLs in this data set than in the WTCCC data set might be explained by differences in the BMI distribution (Supplementary Figure 2.6). We explored this hypothesis by first stratifying the merged Hispanic data set into a lean group composed of subjects with a BMI $<$ 30 and

a non-lean group composed of subjects with a BMI ≥ 30. Although individuals with BMI values less than 30 but greater than 25 can be considered overweight, we included them in the lean cohort to enable a more even comparison both in terms of sample size and as a contrast to obese subjects.

We performed a set of index-subset analyses in each group to compare the heritability contributions from adipose- and muscle-specific eQTLs in relation to cross-tissue eQTLs. The set of cross-tissue eQTLs consistently explained the most heritability within the lean group, whereas the heritability estimates corresponding to the muscle-specific (M) and adipose-specific (A) eQTL sets were not significant and centered on 0 (Figure 2.5A-D; Supplementary Tables 2.26 and 2.27). In contrast, cross-tissue eQTLs did not significantly explain phenotypic variance in the non-lean group in an index-subset analysis that included muscle-specific (M) eQTLs as an index set (Figures 2.5E and 2.5F; Supplementary Table 2.23). Moreover, phenotypic variance was most significantly explained by the complement set of SNPs in this analysis (Figures 2.5E-F; Supplementary Table 2.28). However, both the magnitude and the significance of the heritability estimate attributable to the complement set were considerably reduced when we jointly estimated the heritability explained by adipose-specific eQTLs (Figures 2.5G-H; Supplementary Table 2.29). Furthermore, adipose-specific eQTLs consistently explained the most phenotypic variance within the non-lean group whether or not we included BMI as a covariate for T2D in the REML estimation of genetic variance (Figures 2.5G and 2.5H; Supplementary Tables 2.29 and 2.30). All results shown in Figure 2.5 correspond to analyses in which we included BMI as a covariate.

Next, we performed a permutation-based analysis to evaluate the difference in enrichment by the adipose-specific eQTL subset between the two cohorts and found that the heritability explained by the adipose-specific eQTL subset in the non-lean cohort was significantly greater than that in the lean cohort ($p < 0.001$). Moreover, we performed a set of analysis to assess the difference in the variances explained by the A $\cup$ AM subset between cohorts and similarly found a greater enrichment for this subset in the non-lean cohort than in the lean cohort

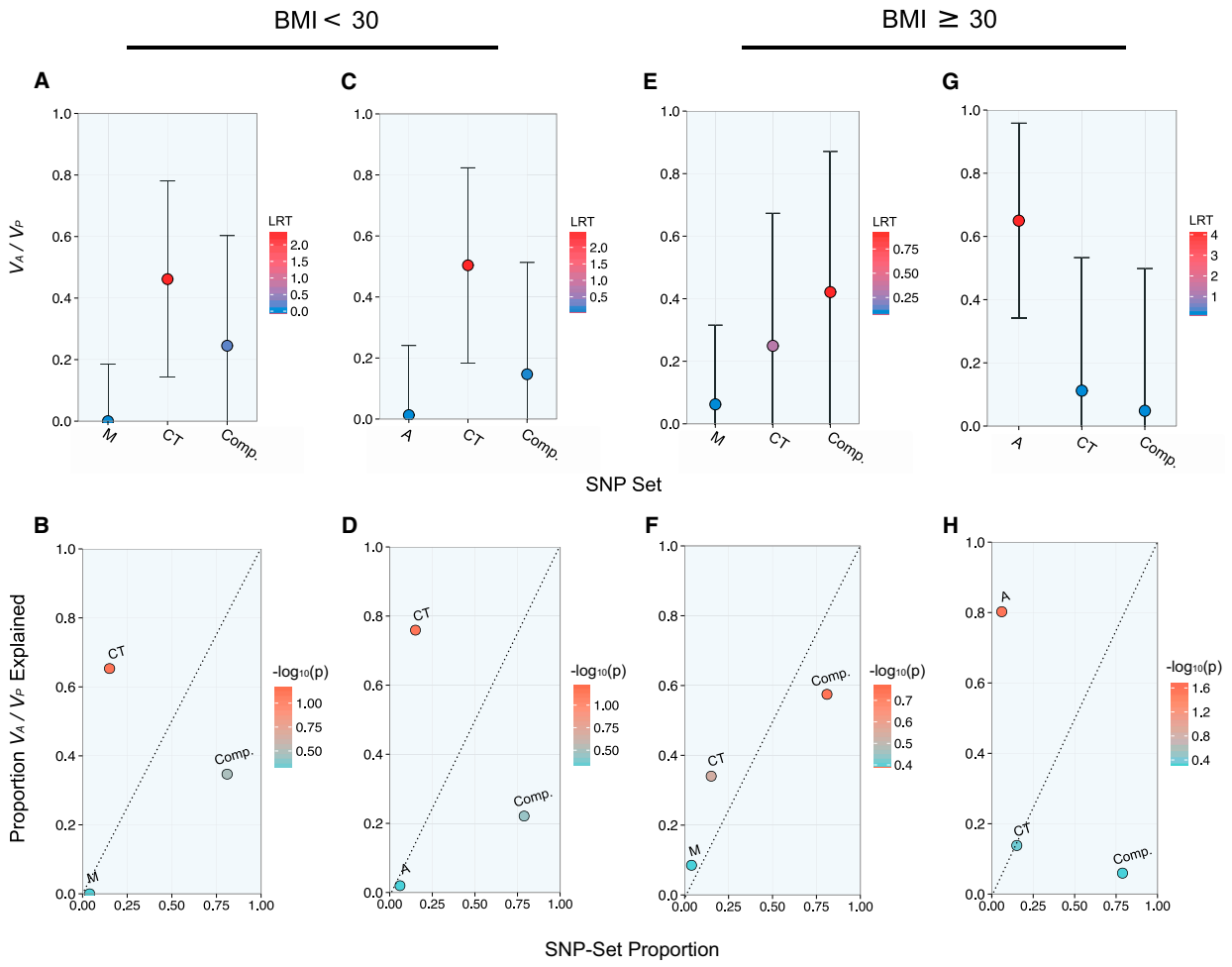

Figure 2.5: **Partition of metabolic-tissue heritability among Hispanics in low- and high-BMI subgroups.** The narrow-sense heritability estimates (V_A/V_P) attributable to non-overlapping SNP subsets are shown with SE and color corresponding to LRT (top panels). The proportion of chip heritability explained by each subset is plotted with SNP-set proportion and is color coded by the $\log_{10}$ of the LRT p value (bottom panels). Results from an index-subset analysis with muscle-specific (M) and cross-tissue (CT) eQTLs as the index sets are shown for Hispanic subjects with a BMI < 30 (A and B) and subjects with a BMI ≥ 30 (E and F). Results from an index-subset analysis with adipose-specific (A) and cross-tissue (CT) eQTLs as the index sets are shown for Hispanic subjects with a BMI < 30 (C and D) and subjects with a BMI ≥ 30 (F and G).

($p = 0.003$; Supplementary Table 2.31).

2.4.6 Joint Estimation of eQTL Heritability and Adjustment for Incomplete LD with Causal Variants

Imperfect LD between causal variants and tagging SNPs can affect the estimation of heritability. Thus, Yang et al. proposed a GRM adjustment that corrects for greater prediction error when the MAF distribution of causal variants deviates from that of the GWAS-interrogated SNPs [166, 168]. In order to determine how imperfect LD at causal loci might affect our joint analyses of heritability, we investigated our results from the IRPT-LCL analysis for both GWAS data sets by adjusting across different values of θ such that the assumed MAF at causal SNPs was MAF $\leq \theta$ (see Material and Methods). Although the significance of the heritability estimates decreased and the SE increased with decreasing values of θ, the relationships between the eQTL subsets reported above were maintained (Supplementary Figures 2.12 and 2.13).

Speed et al. proposed a weighted GRM approach that involves the calculation of local LD throughout the set of GWAS SNPs and that adjusts the GRM according to an appropriate weighting for each SNP [139]. In order to investigate the effect of local LD correction on joint estimation of eQTL heritability, we incorporated this weighted GRM approach implemented in the program LDAK into our IRPT-LCL analysis in both GWAS data sets (Material and Methods). The relationships between eQTL subsets reported from this analysis were maintained (i.e., eQTLs mapped in IRPT disproportionately explained T2D heritability, whereas eQTLs mapped in LCL did not) (Supplementary Figures 2.14 and 2.15; Supplementary Tables 2.32 and 2.33).

2.4.7 Evaluation of the Robustness of eQTL Relationships

As mentioned above, we included the full set of individuals in the merged Hispanic data set in our estimation of heritability. We evaluated relatedness thresholds ranging from 5% to 20% and found no meaningful change in the total estimate of heritability, although there was, as expected, an increase in the SE with each corresponding decrease in sample size (Supplementary Table 2.34). Moreover, we observed consistent relationships between heritability estimates in eQTL partition analyses across relatedness thresholds (Figure S11). We also observed that the relationships between eQTL subsets in the IRPT-LCL partition analysis were maintained across MAF thresholds in both the WTCCC and the Hispanic data sets (Supplementary Figures 2.17 and 2.18; Supplementary Tables 2.35 and 2.36).

2.5 Discussion

In this study, we found that the cumulative effect of GWAS-interrogated SNPs explained a large proportion of phenotypic variation for T2D ($> 50\%$) in the WTCCC data set and in two independent data sets representing Hispanic populations of Mexican descent (Figure 2.2A). Moreover, the heritability attributable to the total set of GWAS-interrogated SNPs for each GWAS data set was greater than the heritability attributable to SNP subsets enriched with previously reported T2D-associated variants (Figure 2.2A-D). These results provide evidence that a considerable proportion of heritability of T2D in these data sets is driven by many SNPs that individually do not attain genome-wide significance in GWAS and therefore support a highly polygenic architecture for T2D.

We have also shown that SNP subsets composed of eQTLs mapped in IRPTs explain disproportionate heritability in these data sets. Our results showing differences between the partitioning of heritability by adipose eQTLs between the WTCCC and Hispanic data sets, and between leaner and more obese subjects within the Hispanic data set, might reflect disparate genetic architectures for T2D in which inherited variants underlying disease sus-

ceptibility among lean individuals are distinct from those contributing to risk among obese subjects. Indeed, large-scale GWAS on T2D stratified by BMI have identified numerous loci that significantly associate with T2D among lean individuals, but not obese individuals [120].

Previous studies have shown that when total sets of GWAS-interrogated SNPs are partitioned by chromosome, larger SNP subsets corresponding with longer chromosomes explain more phenotypic variance than smaller subsets [21, 169]. However,we did not observe a similar relationship between SNP-set proportion and proportion of explained heritability for the eQTL subsets in our study. As a class of variants, eQTLs share a set of characteristics (e.g., allele-frequency distribution, proximity to genes). Therefore, eQTLs mapped in LCL samples enable a more meaningful comparison with adipose and skeletal muscle eQTLs than a random set of SNPs simply matched for MAF. Notably, the LCL eQTL subset consistently composed approximately 25% of the total SNPs across data sets yet accounted for much less heritability and yielded non-significant heritability estimates (Figures2.3A-D and 2.4A-D). We considered the possibility that observed differences in heritability estimates might be attributable to differences in the properties of SNP subsets independently of eQTL associations. Notably, the permutation-based analyses we conducted preserved both the MAF and the LD properties of the SNPs. Therefore, in these permutation studies, the relationships between heritability estimates attributable to eQTL subsets were robust to any differences in the SNP properties of eQTLs. These results suggest that heritability can be highly "concentrated" by small subsets of variants that influence gene expression in tissues relevant to T2D.

The fact that eQTLs mapped in Europeans and African Americans "concentrated" heritability in Hispanics suggests that at least a proportion of these variants are functionally relevant in Hispanics and can be considered cross-population eQTLs. Recently, Marten et al. reported that eQTLs previously mapped in LCLs exhibited similar effects across samples from seven different populations represented in the Human Genome Diversity Panel [97]. Al-

though some eQTLs might be highly population specific, we could not ascertain such eQTLs in the Hispanic data set because comprehensive eQTL mapping has not been reported for these samples or others of Mexican descent. Thus, heritability estimates reported in this study might be a lower bound for the heritability that might be attributable to eQTLs mapped in these tissues in individuals of Mexican ancestry.

It is important to note that a significant correlation between genome-wide ancestry and T2D has not been observed in the Starr County population [14]. In contrast, a strong correlation between PC 1 and T2D has been previously reported in the MCM data set. However, the control individuals in the MCM study had greater European ancestry that was strongly confounded by socioeconomic status (SES) [14, 117]. Moreover, a relationship between SES and T2D has not been observed in the Starr County study because there is much less variability in SES among subjects in the SCT data set than among subjects in the MCM data set. Therefore, even though we observed a significant effect from PC 1 in the merged Hispanic data set ($p < 2 \text{ x } 10^{-16}$) from a regression of T2D on PC 1), we cannot conclude that this was driven by differences in genetic ancestry between case and control individuals.

In this study, we evaluated T2D variance explained by eQTLs mapped in IRPTs. This approach differs from a genetic-correlation analysis of T2D and gene expression in that the former tests functional classes of inherited variants and the latter constitutes a gene-based study. Our study provides insight into the overall genetic architecture of T2D susceptibility and prioritizes tissues for expression-based analyses. Therefore, these study designs are complementary approaches that yield distinct information.

A straightforward interpretation of these results suggests that eQTLs regulating gene expression in skeletal muscle and adipose tissues are important drivers of T2D heritability. Indeed, skeletal muscle is an important IRPT that plays a predominant role in the postprandial insulin response [24]. Moreover, increased secretion of free fatty acids from adipocytes can promote T2D pathogenesis by promoting hepatic gluconeogenesis, insulin resistance in

peripheral tissues, and impaired insulin secretion from pancreatic β-cells [16, 128, 19]. However, it is possible that many of the eQTLs considered tissue specific in this study might also influence gene expression in other key metabolic tissues, such as the liver or pancreatic islet. Similarly, some of the heritability attributed to *trans*-eQTLs might be secondary to *cis*-regulatory effects in tissues that could not be included in this analysis. Therefore, the eQTL contribution to T2D heritability observed in this study could be attributable to biologic effects in tissues not evaluated here. In order to better resolve variation driving disease susceptibility within and across populations, it will be important to extend studies of T2D heritability and include regulatory variants mapped in a more comprehensive collection of tissues [92].

Partitioning heritability by eQTL classes also provides an avenue for identifying genes and pathways that are relevant to disease susceptibility and that might complement current practices for implicating disease-associated genes from associated loci. As previously reported, the majority of trait-associated variants for T2D and other complex diseases are located in non-coding regions, such as intergenic sites [95, 15]. This complicates interpretability, and investigators often report the gene most proximal to the associated variant as the putative disease-associated gene. However, it is possible that trait-associated variants occupy enhancer elements that modulate the expression of genes distal to the associated loci. For example, non-coding variants within introns of *FTO* (MIM 610966) strongly associate with obesity (MIM 601665) and have replicated across multiple studies [41]. However, recent molecular studies have provided evidence that the obesity-associated intronic sequences within *FTO* function as long-range enhancers of the homeobox gene *IRX3* (MIM 612985) and increase *IRX3*, but not *FTO*, expression in the human brain [138]. Moreover, the dramatic weight reduction observed in *Irx3*-deficient mice corroborates a direct relationship between *IRX3* expression and body-fat composition [138]. Incorporating eQTL information into studies of heritability not only can provide insight into the relative contribution of tissue-specific regulatory variation to disease susceptibility but also might help implicate

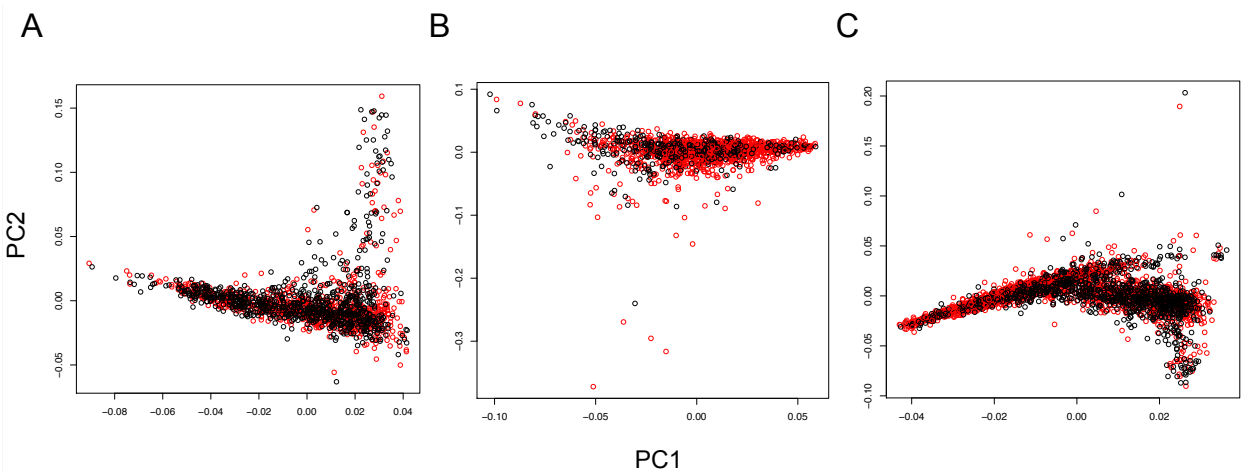

Figure 2.6: **Principal components (PC) analyses of the Hispanic GWAS datasets.** Plots of PC1 versus PC2 are shown for the SCT (A), MCM (B), and MH (C) datasets. Black circles indicate controls and red circles indicate cases.

Dataset	**Cases** (n)	**Controls** (n)	**Age** (years)	**Sex** (n, female)	**BMI** (kg/m^2)
WTCCC	1,924	2,938	NA	2,298	NA
Starr County	837	781	43.0 ± 11.0	1062	31.0 ± 6.4
Mexico City	967	343	45.5 ± 7.3	833	29.1 ± 4.6

Table 2.1: **Characteristics of study populations.** Data are mean $\pm$ SD, unless otherwise indicated. NA, not available.

genes while eschewing assumptions based on proximity.

2.5.1 Acknowledgments

We thank Lea K. Davis, Hae Kyung Im, and Anna Plushnikov for insightful comments and suggestions regarding heritability partitioning. We also thank Graeme I. Bell, David Carmody, and Honggang Ye for useful discussions regarding analyses of type 2 diabetes in the Mexican American and Mexican data sets.

2.6 Supplemental Figures and Tables

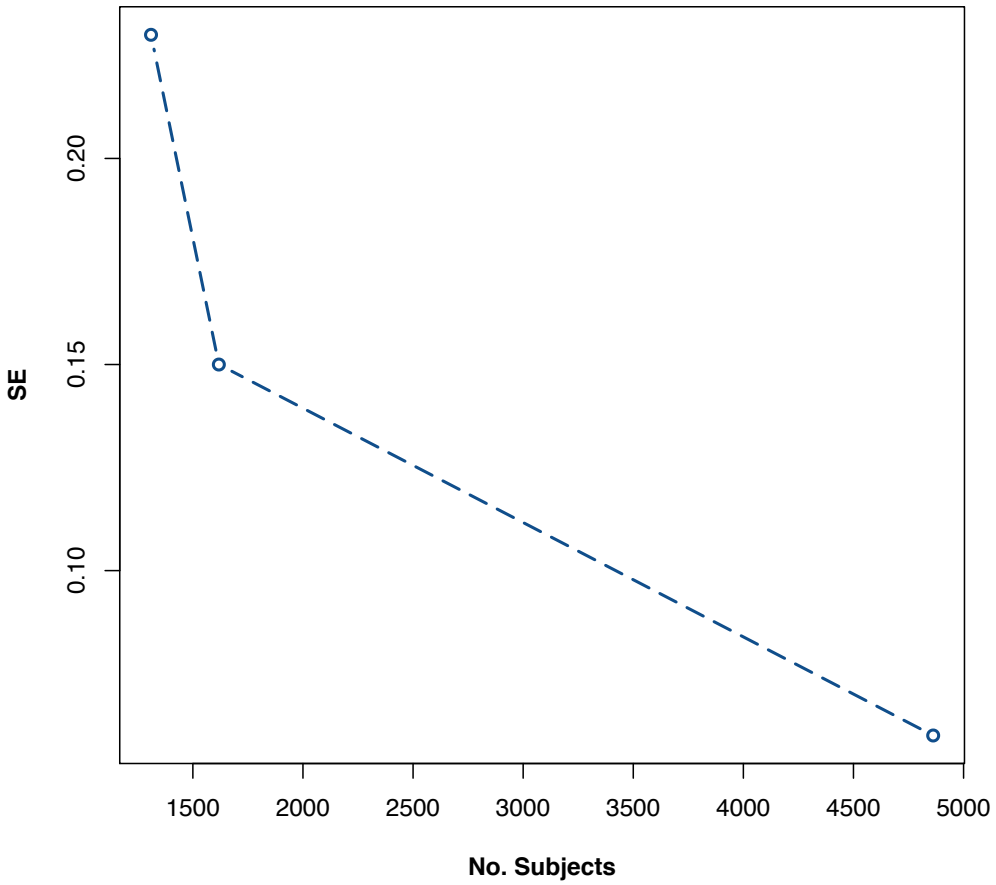

Figure 2.7: **Standard error versus number of subjects represented in each GWAS dataset.** From left to right; MCM, SCT, WTCCC.

Dataset	V_A/V_P	SE	LRT	P-value	No. SNPs	Prevalence
WTCCC	0.60	0.06	98.24	1.9×10^{-23}	370,139	8%
Starr County	0.56	0.15	13.50	1.2×10^{-04}	1,733,064	20%
Mexico City	0.80	0.23	32.18	7.0×10^{-09}	2,431,591	10%
Merged Hispanic	0.60	0.11	32.79	5.0×10^{-09}	1,652,799	16%

Table 2.2: **Estimates of narrow-sense heritability explained by GWAS-interrogated SNPs.** REML estimates of phenotypic variance explained by the additive effect of GWAS SNPs are given with standard error (SE), LRT statistic, p-value, number of SNPs, and prevalence of T2D for the WTCCC, Starr County, Mexico City, and Merged Hispanic dataset.

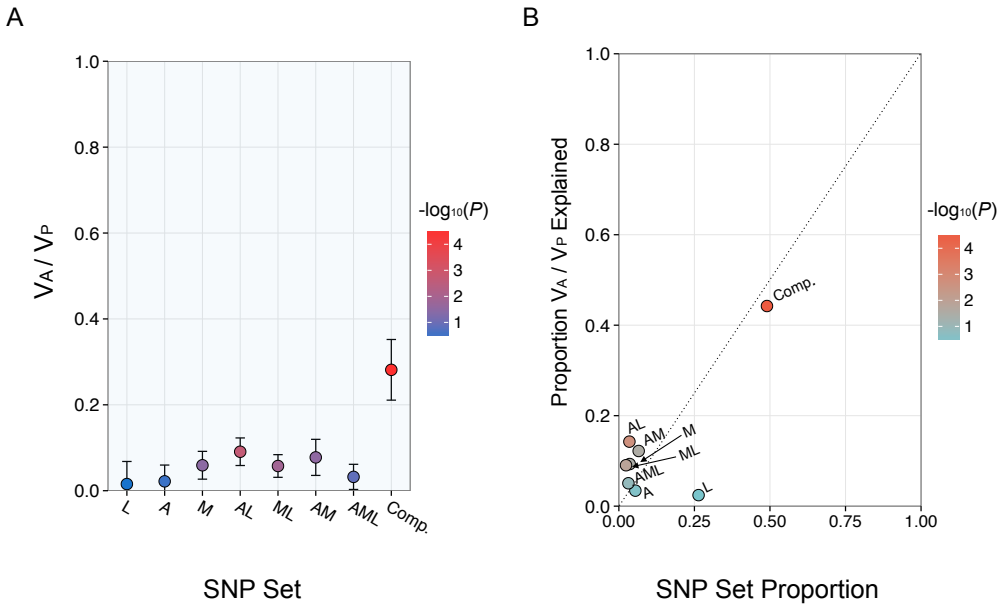

Figure 2.8: **Baseline subset analysis of the WTCCC dataset.** (A) The narrow-sense heritability estimates (V_A/V_P) attributable to disjoint SNP subsets and (B) proportion of explained variance by SNP set proportion for each subset are shown with color corresponding to $\log_{10}$(p-value).

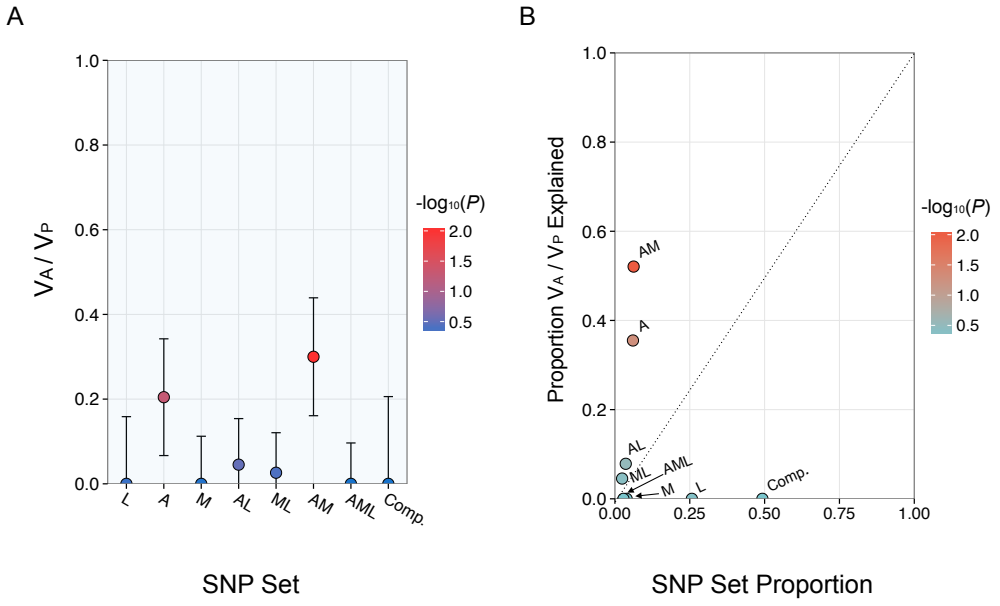

Figure 2.9: **Baseline subset analysis of the merged Hispanic dataset.** (A) The narrow-sense heritability estimates (V_A/V_P) attributable to disjoint SNP subsets and (B) proportion of explained variance by SNP set proportion for each subset are shown with color corresponding to $\log_{10}$(p-value).

WTCCC				
Covariates	***Va/Vp***	**SE**	**LRT**	**P-value**
—	0.61	0.06	107.85	1.4×10^{-25}
Sex	0.60	0.06	104.57	7.6×10^{-25}
PC1-5	0.60	0.06	98.24	1.9×10^{-23}
Sex+PC1-5	0.59	0.06	95.51	7.4×10^{-23}
Sex+PC1-10	0.56	0.06	79.445	2.5×10^{-19}
Starr County				
Covariates	***Va/Vp***	**SE**	**LRT**	**P-value**
—	0.64	0.15	27.25	8.9×10^{-08}
Sex	0.59	0.15	22.64	9.7×10^{-07}
Age	0.65	0.15	21.85	1.5×10^{-06}
BMI	0.64	0.15	26.66	1.2×10^{-07}
PC1-5	0.59	0.15	17.88	1.2×10^{-05}
Sex+Age	0.61	0.15	18.97	6.7×10^{-06}
Sex+BMI	0.58	0.15	21.14	2.1×10^{-06}
Age+BMI	0.64	0.15	21.65	1.6×10^{-06}
Sex+Age+BMI	0.59	0.15	18.098	1.0×10^{-05}
Sex+Age+BMI+PC1-5	0.56	0.15	13.50	0.0001
Sex+Age+BMI+PC1-10	0.55	0.15	12.411	0.0002
Mexico City				
Covariates	***Va/Vp***	**SE**	**LRT**	**P-value**
—	1.07	0.22	82.95	4.2×10^{-20}
Sex	1.07	0.22	79.63	2.3×10^{-19}
Age	0.91	0.23	46.59	4.4×10^{-12}
BMI	0.94	0.22	70.58	2.2×10^{-17}
PC1-5	1.08	0.23	71.30	1.5×10^{-17}
Sex+Age	0.95	0.23	46.52	4.5×10^{-12}
Sex+BMI	0.96	0.22	69.41	4.0×10^{-17}
Age+BMI	0.78	0.23	38.13	3.3×10^{-10}
Sex+Age+BMI	0.84	0.23	38.92	2.2×10^{-10}
Sex+Age+BMI+PC1-5	0.82	0.23	32.34	6.5×10^{-09}
Sex+Age+BMI+PC1-10	0.80	0.23	32.18	7.0×10^{-09}

Table 2.3: **Evaluation of covariates included in estimation of chip heritability.** REML estimates of phenotypic variance explained by the additive effect of GWAS SNPs are given for each GWAS dataset with standard error (SE), LRT statistic, p-value, number of SNPs, and prevalence of T2D.

WTCCC					
Subset	***Va/Vp***	**SE**	**LRT**	**P-value**	**No. SNPs**
1KB	0.05	0.01	150.09	8.3×10^{-35}	97
10KB	0.06	0.01	130.12	1.9×10^{-30}	429
100KB	0.09	0.01	106.47	2.9×10^{-25}	3043
500KB	0.12	0.02	64.31	5.3×10^{-16}	13931
1MB	0.16	0.02	62.65	1.2×10^{-15}	27116
Starr County					
Subset	***Va/Vp***	**SE**	**LRT**	**P-value**	**No. SNPs**
1KB	0.01	0.02	0.47	0.25	268
10KB	0.02	0.02	0.70	0.20	1546
100KB	0.09	0.04	5.57	9.1×10^{-03}	12063
500KB	0.19	0.06	11.64	3.2×10^{-04}	58852
1MB	0.20	0.08	8.47	1.8×10^{-03}	115591
Mexico City					
Subset	***Va/Vp***	**SE**	**LRT**	**P-value**	**No. SNPs**
1KB	0.08	0.04	6.53	5.3×10^{-03}	445
10KB	0.07	0.04	3.79	0.03	2684
100KB	0.05	0.05	1.08	0.15	20347
500KB	0.23	0.09	6.80	4.6×10^{-03}	91926
1MB	0.23	0.11	4.57	0.02	177167

Table 2.4: **Estimates of narrow-sense heritability explained by subsets comprised of T2D-associated SNPs.** Subsets are comprised of T2D-associated SNPs from the NHGRI catalogue and HapMap2 SNPs within 1 Kb, 10 Kb, 100 Kb, 500 Kb, and 1 Mb of the associated SNPs. REML estimates of phenotypic variance explained by the additive effect of SNPs from each subset are given with standard error (SE), LRT statistic, p-value, and number of SNPs. Results are listed for each GWAS dataset; WTCCC, SCT, MCM.

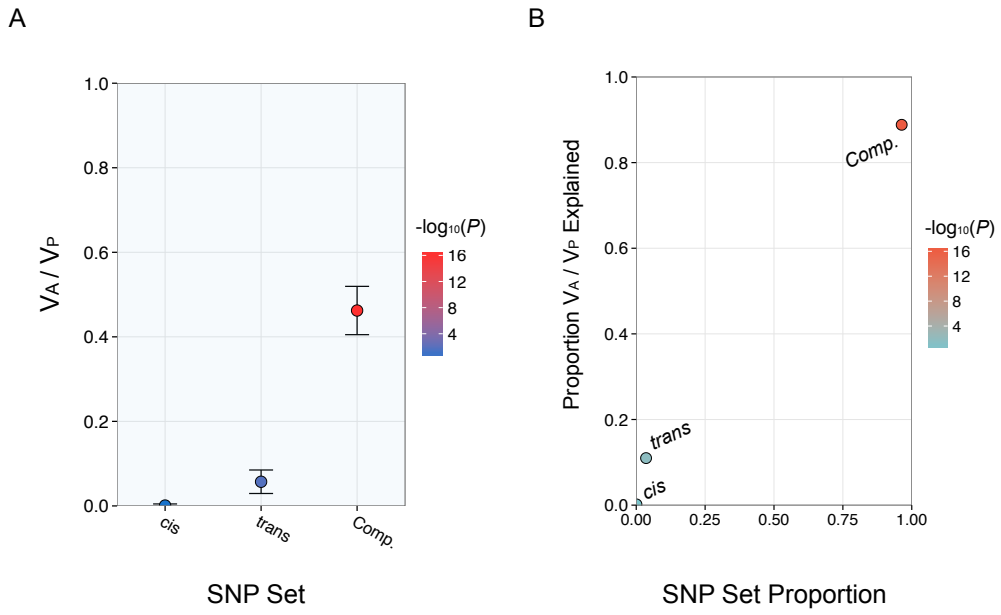

Figure 2.10: **Analysis of heritability from *cis* and *trans* skeletal muscle eQTLs in the WTCCC dataset.** (A) The narrow-sense heritability estimates (V_A/V_P) attributable to disjoint SNP subsets and (B) proportion of explained variance by SNP set proportion for each subset are shown with color corresponding to $\log_{10}$(p-value).

WTCCC			**Hispanic**		
Set	**No. SNPs**	**SNP Set Proportion**	**Set**	**No. SNPs**	**SNP Set Proportion**
LCL	134,600	0.354	LCL	569,600	0.345
Adipose	72,290	0.190	Adipose	309,355	0.187
Muscle	60,691	0.160	Muscle	254,357	0.154
GWAS	380,660	1.00	GWAS	1,652,799	1.00

Table 2.5: **Proportion of SNPs designated as eQTLs in each GWAS dataset.** The number of SNPs represented in the reference eQTL set and SNP set proportion are shown for each eQTL classification in the WTCCC and merged Hispanic datasets. There are a total of 838, 302 SNPs, 616, 455 SNPs, and 523, 682 SNPs in the LCL adipose, and skeletal muscle eQTL reference sets, respectively.

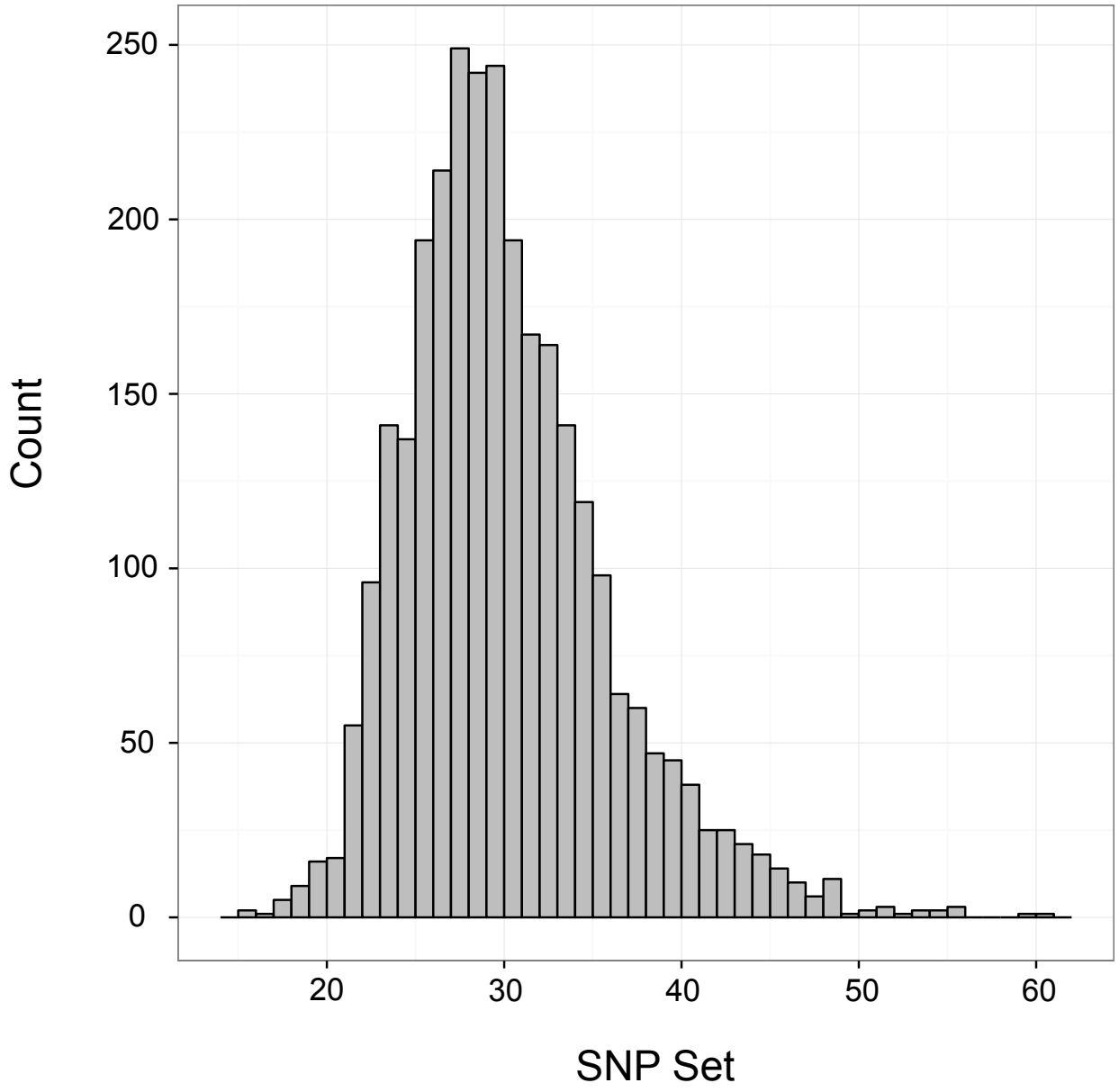

Figure 2.11: **BMI distribution of the merged Hispanic dataset.** Mean value with standard deviation, 30.12 ± 5.74 $kg/m2$.

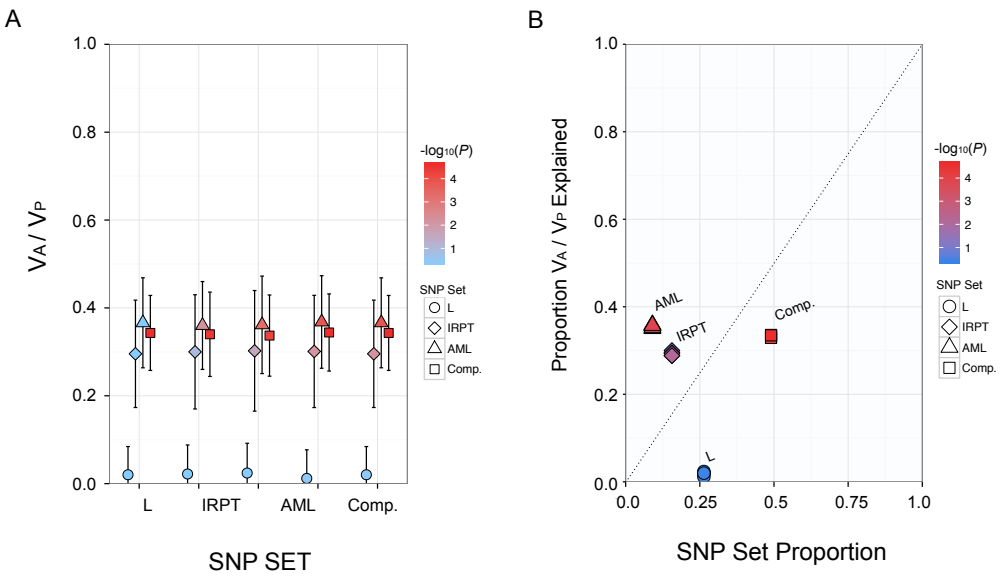

Figure 2.12: **GRM-adjusted partition analysis of the WTCCC dataset.** (A) The narrow-sense heritability estimates (V_A/V_P) attributable to disjoint SNP subsets in the IRPT-LCL partition are shown at each assumed MAF threshold (θ) for causal variants and (B) proportion of explained variance by SNP set proportion for each subset are shown with color corresponding to $\log_{10}$(p-value). The shape of each point designates the subset identity.

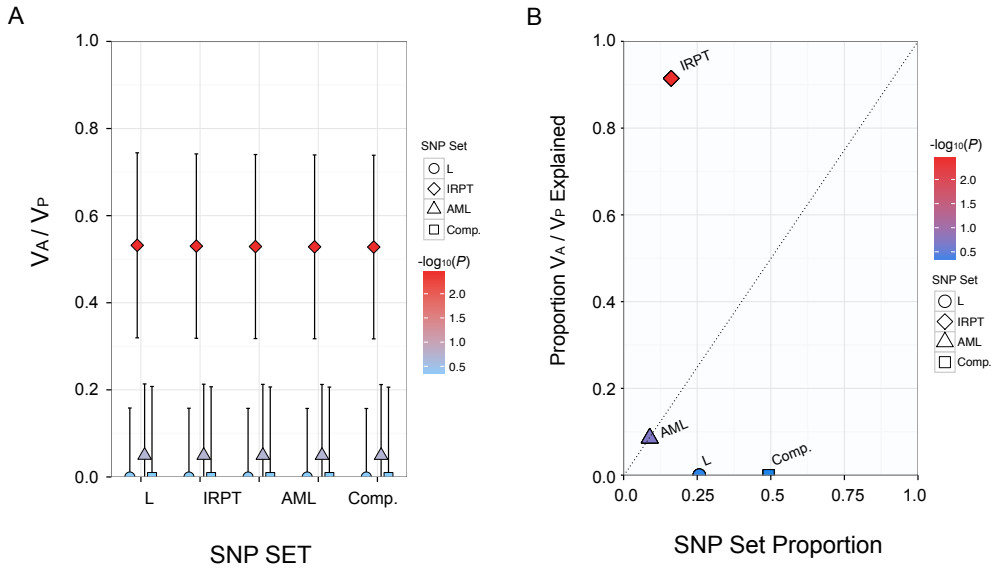

Figure 2.13: **GRM-adjusted partition analysis of the merged Hispanic dataset.** (A) The narrow-sense heritability estimates (V_A/V_P) attributable to disjoint SNP subsets in the IRPT-LCL partition are shown at each assumed MAF threshold (θ) for causal variants and (B) proportion of explained variance by SNP set proportion for each subset are shown with color corresponding to $\log_{10}$(p-value). The shape of each point designates the subset identity.

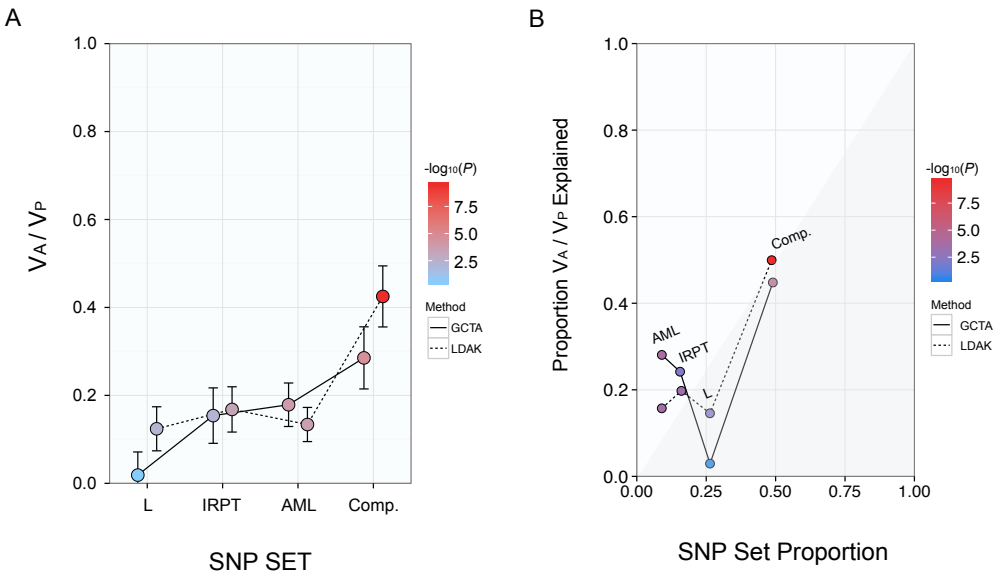

Figure 2.14: **GRM-adjusted partition analysis of the WTCCC dataset.** (A) The narrow-sense heritability estimates (V_A/V_P) attributable to disjoint SNP subsets in the IRPT-LCL partition and (B) proportion of explained variance by SNP set proportion for each subset are shown with color corresponding to $\log_{10}$(p-value). Estimates corresponding to the unadjusted GRMs (GCTA) and LD-adjusted GRMs (LDAK) are designated by solid and dashed lines, respectively.

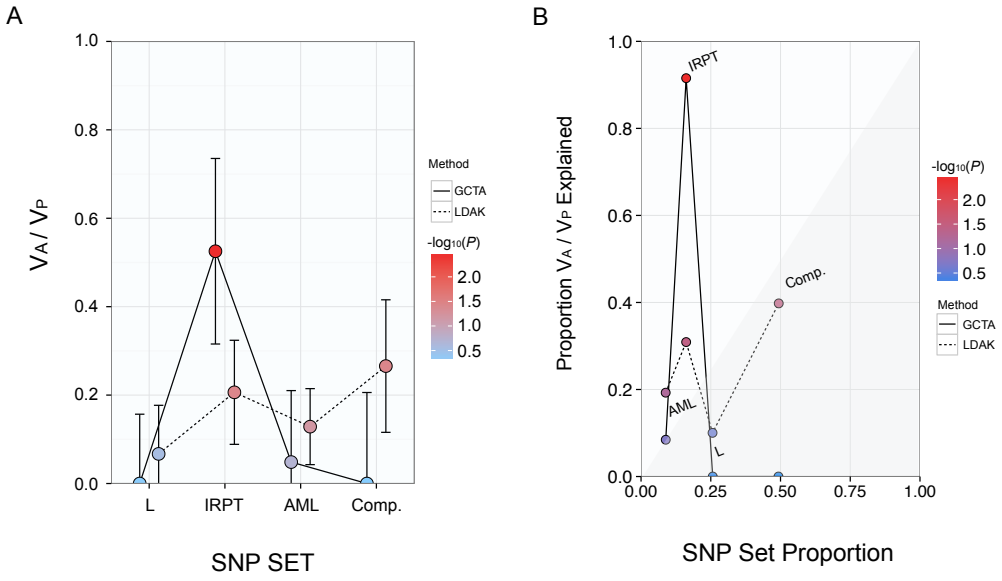

Figure 2.15: **GRM-adjusted partition analysis of the merged Hispanic dataset.** (A) The narrow-sense heritability estimates (V_A/V_P) attributable to disjoint SNP subsets in the IRPT-LCL partition and (B) proportion of explained variance by SNP set proportion for each subset are shown with color corresponding to $\log_{10}$(p-value). Estimates corresponding to the unadjusted GRMs (GCTA) and LD-adjusted GRMs (LDAK) are designated by solid and dashed lines, respectively.

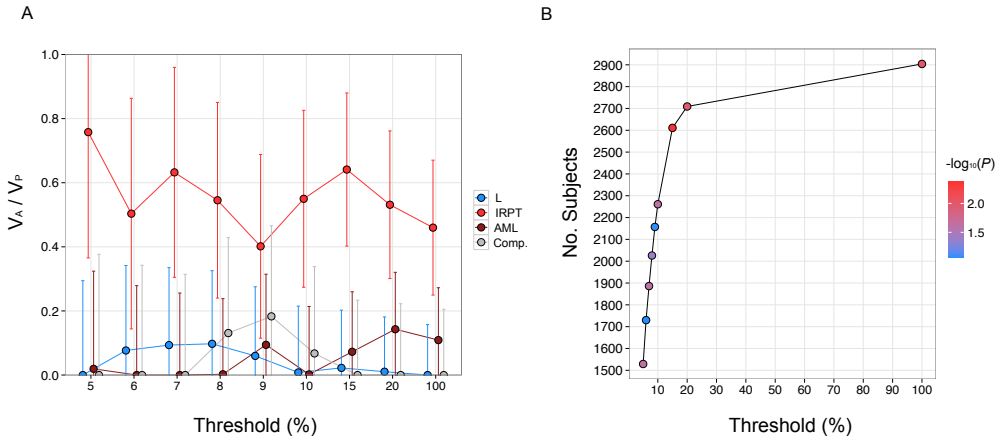

Figure 2.16: **The effect of relatedness thresholds on an eQTL partition analysis of T2D heritability.** (A) The IRPT-LCL eQTL partition analysis (explained in detail in the Methods section) was performed at various relatedness thresholds in the merged Hispanic dataset The estimate of heritability (V_A/V_P) is shown for each disjoint subset designated by color at each threshold. (B) The number of retained individuals at each threshold is shown with color corresponding to the significance ($\log_{10}$(p-value)) of the V_A/V_P estimate for the insulin-responsive peripheral tissue (IRPT) subset.

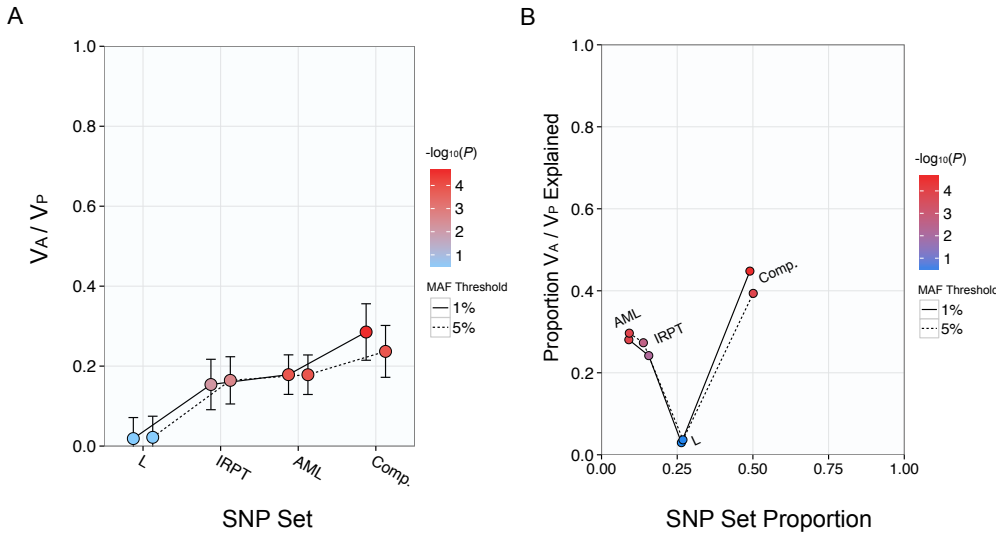

Figure 2.17: **Effect of MAF threshold on partition analysis of the WTCCC dataset.** (A) The narrow-sense heritability estimates (V_A/V_P) attributable to disjoint SNP subsets in the IRPT-LCL partition and (B) proportion of explained variance by SNP set proportion for each subset are shown with color corresponding to $\log_{10}$(p-value). Solid and dashed lines indicate results from a joint analysis performed with a MAF threshold of 1% and 5%, respectively.

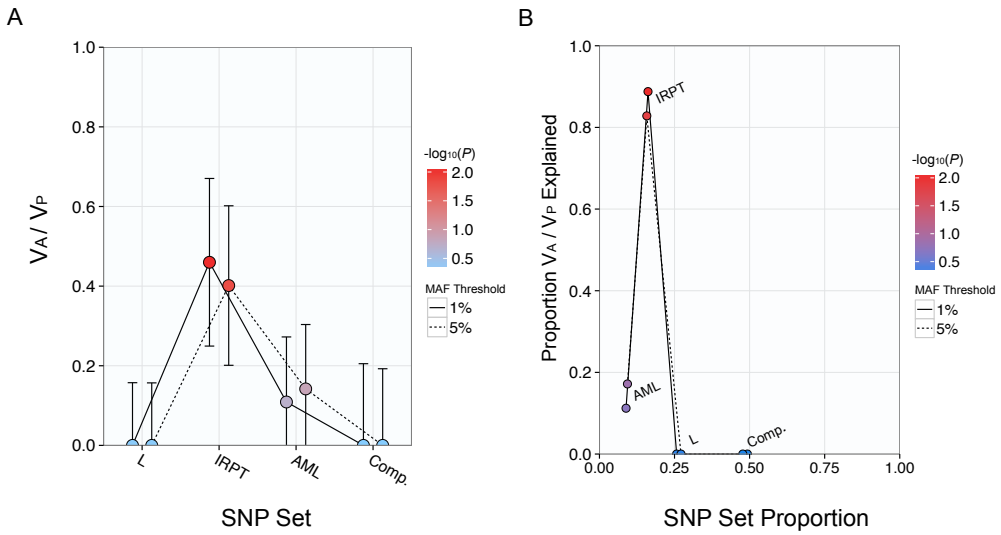

Figure 2.18: **Effect of MAF threshold on partition analysis of the merged Hispanic dataset.** (A) The narrow-sense heritability estimates (V_A/V_P) attributable to disjoint SNP subsets in the IRPT-LCL partition and (B) proportion of explained variance by SNP set proportion for each subset are shown with color corresponding to $\log_{10}$(p-value). Solid and dashed lines indicate results from a joint analysis performed with a MAF threshold of 1% and 5%, respectively.

Subset	*Va/Vp*	SE	LRT	P-value	SNP Set Proportion	Proportion of *Chip* Heritability	Factor
L	0.02	0.05	0.12	0.36	0.27	0.03	0.1
IRPT	0.13	0.05	6.16	6.5×10^{-03}	0.16	0.24	1.4
AML	0.15	0.04	13.36	1.3×10^{-04}	0.09	0.28	2.8
Complement	0.25	0.06	16.46	2.5×10^{-05}	0.48	0.45	0.9

Table 2.6: **Estimates of narrow-sense heritability explained by eQTL subsets in the WTCCC dataset.** REML estimates of phenotypic variance explained by the additive effect of SNPs from each subset are given with standard error (SE), LRT statistic, p-value, SNP set proportion, and proportion of chip heritability explained.

Subset	*Va/Vp*	SE	LRT	P-value	SNP Set Proportion	Proportion of *Chip* Heritability	Factor
L	0.01	0.04	0.08	0.39	0.27	0.02	0.1
A	0.02	0.03	0.35	0.28	0.06	0.03	0.6
M	0.05	0.03	3.36	0.03	0.04	0.09	2.3
AM	0.09	0.04	5.06	0.01	0.10	0.17	1.6
AML	0.13	0.04	12.89	1.6×10^{-04}	0.06	0.23	3.6
Complement	0.24	0.06	16.54	2.4×10^{-05}	0.48	0.44	0.9

Table 2.7: **Expanded IRPT-LCL analysis of the WTCCC dataset.** REML estimates of phenotypic variance explained by the additive effect of SNPs from each subset are given with standard error (SE), LRT statistic, pvalue, SNP set proportion, and proportion of chip heritability explained. Factor relates SNP set proportion to proportion of chip heritability explained.

Subset	*Va/Vp*	SE	LRT	P-value	SNP Set Proportion	Proportion of *Chip* Heritability	Factor
A $\cup$ AL	0.11	0.05	5.06	0.01	0.09	0.17	2.0
M $\cup$ ML	0.18	0.04	9.0	1.4×10^{-03}	0.06	0.19	3.2
Complement	0.39	0.08	27.31	8.7×10^{-08}	0.85	0.63	0.7

Table 2.8: **Index subset analysis of the WTCCC dataset.** REML estimates of phenotypic variance explained by the additive effect of SNPs from each subset are given with standard error (SE), LRT statistic, p-value, SNP set proportion, and proportion of chip heritability explained. Factor relates SNP set proportion to proportion of chip heritability explained.

Subset	*Va/Vp*	SE	LRT	P-value	SNP Set Proportion	Proportion of *Chip* Heritability	Factor
L	0.02	0.05	0.09	0.39	0.26	0.02	0.1
A	0.02	0.04	0.35	0.28	0.05	0.03	0.6
M	0.06	0.03	3.39	0.03	0.04	0.09	2.6
AL	0.09	0.03	8.21	0.00	0.03	0.14	4.1
ML	0.06	0.03	4.98	0.01	0.02	0.09	3.8
AM	0.08	0.04	3.55	0.03	0.07	0.12	1.9
AML	0.03	0.03	1.26	0.13	0.03	0.05	1.6
Complement	0.28	0.07	16.03	0.00	0.49	0.44	0.9

Table 2.9: **Baseline subset analysis of the WTCCC dataset.** REML estimates of phenotypic variance explained by the additive effect of SNPs from each subset are given with standard error (SE), LRT statistic, p-value, SNP set proportion, and proportion of chip heritability explained. Factor relates SNP set proportion to proportion of chip heritability explained.

Subset	*Va/Vp*	SE	LRT	P-value	SNP Set Proportion	Proportion of *Chip* Heritability	Factor
M	0.07	0.03	4.53	0.02	0.04	0.11	3.1
CT	0.23	0.06	14.94	5.5×10^{-05}	0.16	0.37	2.5
Complement	0.32	0.07	19.80	4.3×10^{-06}	0.80	0.52	0.7

Table 2.10: **Index subset analysis of the WTCCC dataset.** REML estimates of phenotypic variance explained by the additive effect of SNPs from each subset are given with standard error (SE), LRT statistic, p-value, SNP set proportion, and proportion of chip heritability explained. Factor relates SNP set proportion to proportion of chip heritability explained. M and CT denote muscle-specific and cross-tissue eQTL subsets, respectively.

Subset	*Va/Vp*	SE	LRT	P-value	SNP Set Proportion	Proportion of *Chip* Heritability	Factor
A $\cup$ M $\cup$ AM	0.18	0.06	8.83	1.5×10^{-03}	0.16	0.29	1.8
AL $\cup$ ML $\cup$ AML	0.15	0.05	10.43	6.2×10^{-03}	0.09	0.24	2.7
Complement	0.30	0.08	15.59	3.9×10^{-05}	0.75	0.48	0.6

Table 2.11: **Index subset analysis of the WTCCC dataset.** REML estimates of phenotypic variance explained by the additive effect of SNPs from each subset are given with standard error (SE), LRT statistic, p-value, SNP set proportion, and proportion of chip heritability explained. Factor relates SNP set proportion to proportion of chip heritability explained.

Subset	*Va/Vp*	SE	LRT	P-value	SNP Set Proportion	Proportion of *Chip* Heritability	Factor
L	0.00	0.04	0.00	0.50	0.16	0.00	0.0
CT	0.24	0.06	16.98	1.9×10^{-05}	0.16	0.38	2.4
Complement	0.40	0.07	33.51	3.5×10^{-09}	0.68	0.62	0.9

Table 2.12: **chip heritability explained by matched sets of LCL-specific and cross-tissue eQTLs in the WTCCC dataset.** In this analysis, the L subset contained the set of the most significant LCL-specific eQTLs so that the SNP set proportion was identical to that of the Cross-tissue set. REML estimates of phenotypic variance explained by the additive effect of GWAS SNPs are given for each GWAS dataset with standard error (SE), LRT statistic, p-value, number of SNPs, and prevalence of T2D.

Subset	*Va/Vp*	SE	LRT	P-value	SNP Set Proportion	Proportion of *Chip* Heritability	Factor
Cross Tissue (*Cis*)	0.02	0.02	2.47	0.06	0.01	0.04	3.9
Cross Tissue (*Trans*)	0.21	0.06	11.78	3.0×10^{-04}	0.15	0.33	2.3
Complement	0.39	0.07	29.45	2.9×10^{-08}	0.84	0.63	0.7

Table 2.13: ***cis-trans* analysis of cross-tissue eQTLs in the WTCCC dataset.** REML estimates of phenotypic variance explained by the additive effect of SNPs from each subset are given with standard error (SE), LRT statistic, p-value, SNP set proportion, and proportion of chip heritability explained. Factor relates SNP set proportion to proportion of chip heritability explained.

Subset	*Va/Vp*	SE	LRT	P-value	SNP Set Proportion	Proportion of *Chip* Heritability	Factor
Muscle (*Cis*)	0.0009	0.004	0.06	0.40	0.0008	0.002	2.2
Muscle (*Trans*)	0.06	0.03	4.29	0.02	0.04	0.11	3.1
Complement	0.46	0.06	68.53	6.3×10^{-17}	0.96	0.89	0.9

Table 2.14: ***cis-trans* analysis of muscle-specific eQTLs in the WTCCC dataset.** REML estimates of phenotypic variance explained by the additive effect of SNPs from each subset are given with standard error (SE), LRT statistic, p-value, SNP set proportion, and proportion of chip heritability explained. Factor relates SNP set proportion to proportion of chip heritability explained.

Subset	*Va/Vp*	SE	LRT	P-value	SNP Set Proportion	Proportion of *Chip* Heritability	Factor
L	0.00	0.27	0.00	0.49	0.26	0.00	0.0
A	0.08	0.25	0.05	0.41	0.06	0.12	2.2
M	0.27	0.20	2.17	0.07	0.04	0.41	11.4
AM	0.26	0.27	0.83	0.18	0.09	0.39	4.5
AML	0.06	0.26	0.69	0.20	0.06	0.08	1.5
Complement	0.00	0.35	0.00	0.50	0.50	0.00	0.0

Table 2.15: **Expanded IRPT-LCL analysis of the SCT dataset.** REML estimates of phenotypic variance explained by the additive effect of SNPs from each subset are given with standard error (SE), LRT statistic, p-value, SNP set proportion, and proportion of chip heritability explained. Factor relates SNP set proportion to proportion of chip heritability explained.

Subset	*Va/Vp*	SE	LRT	P-value	SNP Set Proportion	Proportion of *Chip* Heritability	Factor
L	0.10	0.41	0.00	0.50	0.23	0.10	0.4
A	0.00	0.40	0.00	0.50	0.06	0.00	0.0
M	0.25	0.31	3.19	0.04	0.04	0.25	7.1
AM	0.48	0.47	0.00	0.50	0.10	0.49	4.9
AML	0.00	0.43	0.00	0.50	0.05	0.00	0.0
Complement	0.15	0.58	0.00	0.50	0.52	0.16	0.3

Table 2.16: **Expanded IRPT-LCL analysis of the MCM dataset.** REML estimates of phenotypic variance explained by the additive effect of SNPs from each subset are given with standard error (SE), LRT statistic, p-value, SNP set proportion, and proportion of chip heritability explained. Factor relates SNP set proportion to proportion of chip heritability explained.

WTCCC				
Prevalence (%)	***Va/Vp***	**SE**	**LRT**	**P-value**
5	0.52	0.05	98.24	1.9×10^{-23}
10	0.64	0.07	98.24	1.9×10^{-23}
15	0.73	0.08	98.24	1.9×10^{-23}
20	0.79	0.08	98.24	1.9×10^{-23}
25	0.85	0.09	98.24	1.9×10^{-23}
Starr County				
Prevalence (%)	***Va/Vp***	**SE**	**LRT**	**P-value**
5	0.37	0.10	13.50	0.0001
10	0.45	0.12	13.50	0.0001
15	0.52	0.14	13.50	0.0001
20	0.56	0.15	13.50	0.0001
25	0.60	0.16	13.50	0.0001
Mexico City				
Prevalence (%)	***Va/Vp***	**SE**	**LRT**	**P-value**
5	0.65	0.19	32.18	7.0×10^{-09}
10	0.80	0.23	32.18	7.0×10^{-09}
15	0.92	0.27	32.18	7.0×10^{-09}
20	1.00	0.29	32.18	7.0×10^{-09}
25	1.07	0.31	32.18	7.0×10^{-09}

Table 2.17: **Evaluation of the prevalence parameter specified in the estimation of chip heritability.** REML estimates of phenotypic variance explained by the additive effect of GWAS SNPs are given for each GWAS dataset with standard error (SE), LRT statistic, p-value, number of SNPs, and prevalence of T2D.

Subset	*Va/Vp*	SE	LRT	P-value	SNP Set Proportion	Proportion of *Chip* Heritability	Factor
L	0.00	0.16	0.01	0.47	0.26	0.00	0.0
IRPT	0.46	0.21	5.60	0.01	0.16	0.81	5.0
AML	0.11	0.16	0.72	0.20	0.09	0.19	2.2
Complement	0.00	0.21	0.00	0.50	0.49	0.00	0.0

Table 2.18: **Estimates of narrow-sense heritability explained by eQTL subsets in the merged Hispanic dataset.** REML estimates of phenotypic variance explained by the additive effect of SNPs from each subset are given with standard error (SE), LRT statistic, p-value, SNP set proportion, and proportion of chip heritability explained. Factor relates SNP set proportion to proportion of chip heritability explained.

Subset	*Va/Vp*	SE	LRT	P-value	SNP Set Proportion	Proportion of *Chip* Heritability	Factor
L	0.00	0.16	0.00	0.50	0.26	0.00	0.0
A	0.22	0.14	3.26	0.04	0.06	0.38	6.3
M	0.00	0.11	0.00	0.50	0.04	0.00	0.0
AM	0.27	0.15	3.54	0.03	0.09	0.48	5.2
AML	0.08	0.15	0.60	0.22	0.06	0.15	2.4
Complement	0.00	0.20	0.01	0.45	0.49	0.00	0.0

Table 2.19: **Expanded IRPT-LCL analysis of the merged Hispanic dataset.** REML estimates of phenotypic variance explained by the additive effect of SNPs from each subset are given with standard error (SE), LRT statistic, p-value, SNP set proportion, and proportion of chip heritability explained. Factor relates SNP set proportion to proportion of chip heritability explained.

Subset	*Va/Vp*	SE	LRT	P-value	SNP Set Proportion	Proportion of *Chip* Heritability	Factor
L	0.00	0.16	0.03	0.43	0.26	0.00	0.0
A	0.20	0.14	2.65	0.05	0.06	0.36	6.0
M	0.00	0.11	0.00	0.48	0.04	0.00	0.0
AL	0.05	0.11	0.24	0.31	0.04	0.08	2.2
ML	0.03	0.09	0.09	0.38	0.02	0.05	1.9
AM	0.30	0.14	5.55	0.01	0.06	0.52	8.3
AML	0.00	0.10	0.00	0.50	0.03	0.00	0.0
Complement	0.00	0.21	0.00	0.50	0.49	0.00	0.0

Table 2.20: **Baseline subset analysis of the merged Hispanic dataset.** REML estimates of phenotypic variance explained by the additive effect of SNPs from each subset are given with standard error (SE), LRT statistic, p-value, SNP set proportion, and proportion of chip heritability explained. Factor relates SNP set proportion to proportion of chip heritability explained.

Subset	*Va/Vp*	SE	LRT	P-value	SNP Set Proportion	Proportion of *Chip* Heritability	Factor
M	0.00	0.11	0.00	0.50	0.04	0.00	0.0
CT	0.45	0.19	5.93	0.01	0.15	0.89	5.9
Complement	0.05	0.20	0.05	0.41	0.81	0.11	0.1

Table 2.21: **Index subset analysis of the merged Hispanic dataset.** REML estimates of phenotypic variance explained by the additive effect of SNPs from each subset are given with standard error (SE), LRT statistic, p-value, SNP set proportion, and proportion of chip heritability explained. Factor relates SNP set proportion to proportion of chip heritability explained. M and CT denote muscle-specific and cross-tissue eQTL subsets, respectively.

Subset	*Va/Vp*	SE	LRT	P-value	SNP Set Proportion	Proportion of *Chip* Heritability	Factor
A	0.21	0.13	3.10	0.04	0.06	0.36	6.0
CT	0.37	0.19	4.37	0.02	0.15	0.64	4.2
Complement	0.00	0.21	0.00	0.50	0.79	0.00	0.0

Table 2.22: **Index subset analysis of the merged Hispanic dataset.** REML estimates of phenotypic variance explained by the additive effect of SNPs from each subset are given with standard error (SE), LRT statistic, pvalue, SNP set proportion, and proportion of chip heritability explained. Factor relates SNP set proportion to proportion of chip heritability explained. A and CT denote adipose-specific and cross-tissue eQTL subsets, respectively.

Subset	*Va/Vp*	SE	LRT	P-value	SNP Set Proportion	Proportion of *Chip* Heritability	Factor
A ∪ AL	0.32	0.16	4.35	0.02	0.10	0.63	6.5
M ∪ ML	0.05	0.13	0.12	0.36	0.06	0.09	1.4
Complement	0.14	0.21	0.47	0.25	0.84	0.28	0.34

Table 2.23: **Index subset analysis of the merged Hispanic dataset.** REML estimates of phenotypic variance explained by the additive effect of SNPs from each subset are given with standard error (SE), LRT statistic, pvalue, SNP set proportion, and proportion of chip heritability explained. Factor relates SNP set proportion to proportion of chip heritability explained.

Subset	*Va/Vp*	SE	LRT	P-value	SNP Set Proportion	Proportion of *Chip* Heritability	Facto
A ∪ M ∪ AM	0.53	0.19	8.29	1.2×10^{-03}	0.16	0.91	5.6
AL ∪ ML ∪ AML	0.05	0.14	0.10	0.38	0.09	0.09	1.0
Complement	1.0×10^{-06}	0.22	0.00	0.50	0.75	1.7×10^{-06}	2.3×10

Table 2.24: **Index subset analysis of the merged Hispanic dataset.** REML estimates of phenotypic variance explained by the additive effect of SNPs from each subset are given with standard error (SE), LRT statistic, pvalue, SNP set proportion, and proportion of chip heritability explained. Factor relates SNP set proportion to proportion of chip heritability explained.

Subset	*Va/Vp*	SE	LRT	P-value	SNP Set Proportion	Proportion of *Chip* Heritability	Factor
L	0.01	0.13	0.01	0.47	0.26	0.02	0.1
Complement	0.48	0.16	8.896	1.4×10^{-03}	0.74	0.98	1.3

Table 2.25: **Index subset analysis of the merged Hispanic dataset.** REML estimates of phenotypic variance explained by the additive effect of SNPs from each subset are given with standard error (SE), LRT statistic, p-value, SNP set proportion, and proportion of chip heritability explained. Factor relates SNP set proportion to proportion of chip heritability explained. L denotes LCL-specific eQTL subset.

Subset	*Va/Vp*	SE	LRT	P-value	SNP Set Proportion	Proportion of *Chip* Heritability	Factor
M	0.00	0.19	0.00	0.50	0.04	0.00	0.0
CT	0.46	0.32	2.32	0.06	0.15	0.65	4.3
Complement	0.25	0.36	0.24	0.31	0.81	0.35	0.4

Table 2.26: **Index subset analysis of the merged Hispanic dataset with BMI** < 30. REML estimates of phenotypic variance explained by the additive effect of SNPs from each subset are given with standard error (SE), LRT statistic, p-value, SNP set proportion, and proportion of chip heritability explained. Factor relates SNP set proportion to proportion of chip heritability explained. M and CT denote muscle-specific and cross-tissue eQTL subsets, respectively.

Subset	*Va/Vp*	SE	LRT	P-value	SNP Set Proportion	Proportion of *Chip* Heritability	Factor
A	0.01	0.23	0.00	0.48	0.06	0.02	0.3
CT	0.50	0.32	2.45	0.06	0.15	0.76	5.0
Complement	0.15	0.37	0.15	0.35	0.79	0.22	0.3

Table 2.27: **Index subset analysis of the merged Hispanic dataset with BMI** < 30. REML estimates of phenotypic variance explained by the additive effect of SNPs from each subset are given with standard error (SE), LRT statistic, p-value, SNP set proportion, and proportion of chip heritability explained. Factor relates SNP set proportion to proportion of chip heritability explained. A and CT denote adipose-specific and cross-tissue eQTL subsets, respectively.

Subset	*Va/Vp*	SE	LRT	P-value	SNP Set Proportion	Proportion of *Chip* Heritability	Factor
M	0.06	0.25	0.06	0.41	0.04	0.08	2.2
CT	0.25	0.42	0.34	0.28	0.15	0.34	2.3
Complement	0.42	0.45	0.91	0.17	0.81	0.57	0.7

Table 2.28: **Index subset analysis of the merged Hispanic dataset with BMI** ≥ 30. REML estimates of phenotypic variance explained by the additive effect of SNPs from each subset are given with standard error (SE), LRT statistic, p-value, SNP set proportion, and proportion of chip heritability explained. Factor relates SNP set proportion to proportion of chip heritability explained. M and CT denote muscle-specific and cross-tissue eQTL subsets, respectively.

Subset	*Va/Vp*	SE	LRT	P-value	SNP Set Proportion	Proportion of *Chip* Heritability	Factor
A	0.65	0.31	4.15	0.02	0.06	0.80	13.3
CT	0.11	0.42	0.07	0.40	0.15	0.14	0.9
Complement	0.05	0.45	0.01	0.46	0.79	0.06	0.1

Table 2.29: **Index subset analysis of the merged Hispanic dataset with BMI** ≥ 30**.** REML estimates of phenotypic variance explained by the additive effect of SNPs from each subset are given with standard error (SE), LRT statistic, p-value, SNP set proportion, and proportion of chip heritability explained. Factor relates SNP set proportion to proportion of chip heritability explained. A and CT denote adipose-specific and cross-tissue eQTL subsets, respectively.

Subset	*Va/Vp*	SE	LRT	P-value	SNP Set Proportion	Proportion of *Chip* Heritability	Factor
A	0.62	0.31	3.87	0.02	0.06	0.69	11.4
CT	0.00	0.42	0.00	0.50	0.15	0.00	0.0
Complement	0.28	0.45	0.45	0.25	0.79	0.31	0.4

Table 2.30: **Index subset analysis of the merged Hispanic dataset with BMI ≥ 30.** BMI was not included as a covariate in this analysis. REML estimates of phenotypic variance explained by the additive effect of SNPs from each subset are given with standard error (SE), LRT statistic, p-value, SNP set proportion, and proportion of chip heritability explained. Factor relates SNP set proportion to proportion of chip heritability explained. A and CT denote adipose-specific and cross-tissue eQTL subsets, respectively.

Subset	*Va/Vp*	SE	LRT	P-value	SNP Set Proportion	Proportion of *Chip* Heritability	Factor
"Lean" cohort							
A $\cup$ AM	0.15	0.30	0.25	0.31	0.12	0.22	1.8
Complement	0.54	0.31	2.8	0.05	0.88	0.78	0.9
"Non-lean" cohort							
A $\cup$ AM	0.69	0.38	3.23	0.04	0.12	0.77	6.3
Complement	0.20	0.38	0.29	0.30	0.88	0.23	0.3

Table 2.31: **Index subset analysis of the merged Hispanic.** SNPs were partitioned into the set comprised of the union of adipose-specific eQTLs and eQTLs mapped in both adipose and muscle tissue (A $\cup$ AM) and the complement set of SNPs. For the lean (BMI < 30) and non-lean (BMI ≥ 30) cohort, REML estimates of phenotypic variance explained by the additive effect of SNPs from each subset are given with standard error (SE), LRT statistic, p-value, SNP set proportion, and proportion of chip heritability explained. Factor relates SNP set proportion to proportion of chip heritability explained.

Subset	*Va/Vp*	SE	LRT	P-value	SNP Set Proportion	Proportion of *Chip* Heritability	Factor
				GCTA			
L	0.02	0.0	0.12	0.36	0.26	0.03	0.1
IRPT	0.15	0.06	6.16	6.5×10^{-03}	0.16	0.24	1.5
AML	0.18	0.05	13.36	1.2×10^{-04}	0.09	0.28	3.1
Comp.	0.29	0.07	16.46	2.5×10^{-05}	0.49	0.45	0.9
				LDAK			
L	0.12	0.05	6.34	5.9×10^{-03}	0.26	0.15	0.6
IRPT	0.17	0.05	11.00	4.5×10^{-04}	0.16	0.20	1.3
AML	0.13	0.04	12.26	2.3×10^{-04}	0.09	0.16	1.8
Comp.	0.43	0.07	38.25	$3.11 \times 10z^{-10}$	0.49	0.50	1.0

Table 2.32: **GRM-adjusted estimates of narrow-sense heritability explained by eQTL subsets in the WTCCC dataset.** Heritability estimates corresponding to the unadjusted GRM (GCTA) and LD-adjusted GRM (LDAK) for each subset are given with standard error (SE), LRT statistic, p-value, SNP set proportion, and proportion of chip heritability explained. Factor relates SNP set proportion to proportion of chip heritability explained.

Subset	*Va/Vp*	SE	LRT	P-value	SNP Set Proportion	Proportion of *Chip* Heritability	Factor
				GCTA			
L	0.00	0.16	0.01	0.47	0.26	0.00	0.0
IRPT	0.53	0.21	7.2	3.6×10^{-03}	0.16	0.92	5.8
AML	0.05	0.16	0.76	0.19	0.09	0.08	0.9
Comp.	0.00	0.21	0.00	0.5	0.49	0.00	0.0
				LDAK			
L	0.07	0.11	0.38	0.27	0.26	0.10	0.4
IRPT	0.21	0.12	3.20	0.037	0.16	0.31	1.9
AML	0.13	0.09	2.28	0.07	0.09	0.19	2.1
Comp.	0.27	0.15	3.20	0.04	0.50	0.40	0.8

Table 2.33: **GRM-adjusted estimates of narrow-sense heritability explained by eQTL subsets in the merged Hispanic dataset.** Heritability estimates corresponding to the unadjusted GRM (GCTA) and LD-adjusted GRM (LDAK) for each subset are given with standard error (SE), LRT statistic, p-value, SNP set proportion, and proportion of chip heritability explained. Factor relates SNP set proportion to proportion of chip heritability explained.

Threshold (%)	Subjects (n)	*Va/Vp*	SE
2.5	667	0.80	0.40
5	1540	0.60	0.20
10	2272	0.60	0.14
15	2628	0.62	0.12
20	2730	0.60	0.11
None	2928	0.48	0.10

Table 2.34: **Evaluation of genetic relatedness thresholds applied to the merged Hispanic dataset.** REML estimates of phenotypic variance explained by GWAS SNPs are given with standard error (SE) and number of subjects retained (n) for each relatedness threshold.

Subset	*Va/Vp*	SE	LRT	P-value	SNP Set Proportion	Proportion of *Chip* Heritability	Factor
MAF 1%							
L	0.02	0.05	0.12	0.36	0.26	0.03	0.1
A ∪ M ∪ AM	0.15	0.06	6.16	6.5×10^{-03}	0.16	0.24	1.6
AML	0.18	0.05	13.36	1.3×10^{-03}	0.09	0.28	3.1
Complement	0.29	0.07	16.46	2.5×10^{-05}	0.49	0.45	0.9
MAF 5%							
L	0.02	0.05	0.17	0.34	0.27	0.04	0.1
A ∪ M ∪ AM	0.16	0.06	8.0	2.3×10^{-03}	0.14	0.27	2.0
AML	0.18	0.05	13.32	1.3×10^{-03}	0.09	0.30	3.2
Complement	0.24	0.06	13.35	1.3×10^{-03}	0.50	0.39	0.8

Table 2.35: **Effect of MAF threshold on partition analysis of the WTCCC dataset.** REML estimates of phenotypic variance explained by the additive effect of SNPs from each subset in an IRPT-LCL partition analysis are given with standard error (SE), LRT statistic, p-value, SNP set proportion, and proportion of chip heritability explained. Factor relates SNP set proportion to proportion of chip heritability explained.

Subset	*Va/Vp*	SE	LRT	P-value	SNP Set Proportion	Proportion of *Chip* Heritability	Factor
				MAF 1%			
L	0.02	0.05	0.12	0.36	0.26	0.03	0.1
A ∪ M ∪ AM	0.15	0.06	6.16	6.5×10^{-03}	0.16	0.24	1.6
AML	0.18	0.05	13.36	1.3×10^{-03}	0.09	0.28	3.1
Complement	0.29	0.07	16.46	2.5×10^{-05}	0.49	0.45	0.9
				MAF 5%			
L	0.02	0.05	0.17	0.34	0.27	0.04	0.1
A ∪ M ∪ AM	0.16	0.06	8.0	2.3×10^{-03}	0.14	0.27	2.0
AML	0.18	0.05	13.32	1.3×10^{-03}	0.09	0.30	3.2
Complement	0.24	0.06	13.35	1.3×10^{-03}	0.50	0.39	0.8

Table 2.36: **Effect of MAF threshold on partition analysis of the merged Hispanic dataset.** REML estimates of phenotypic variance explained by the additive effect of SNPs from each subset in an IRPT-LCL partition analysis are given with standard error (SE), LRT statistic, p-value, SNP set proportion, and proportion of chip heritability explained. Factor relates SNP set proportion to proportion of chip heritability explained.

CHAPTER 3

TESTING FOR ASSOCIATION BETWEEN PREDICTED GENE EXPRESSION AND TYPE 2 DIABETES

3.1 Abstract

Most genes implicated by GWAS are done so by proximity to associated markers that mostly reside in non-coding genomic regions. However, this assumption is undermined by long-range interactions between eQTLs and promoters of target genes that are not necessarily the reported trait genes closest to associated regions. In order to incorporate regulatory genetic information to address the challenge of gene-mapping, we apply MetaXcan - an integrative approach that inputs summary statistics from GWAS to test for association between estimates of the *genetic component* of gene expression and disease. Leveraging 42 predictive models corresponding to a diverse set of primary human tissues and cell types from the DGN and GTEx projects, we performed a series of gene-based tests with summary data from the DIAGRAM trans-ethnic meta-analysis of T2D, including over 100K individuals. We find evidence that not only supports reported T2D genes at associated GWAS loci, but identified a set of novel T2D gene candidates that comprise the majority of significant MetaXcan associations. Moreover, we replicated associations in an independent cohort from the GERA study and found that of the 25 genes meeting genome-wide significance, 15 represent novel candidate disease genes. Moreover, we also found support for 4 genes that map to "unknown" T2D loci that have not been previously reported. Lastly, novel genes discovered in our analysis overlap with putative trait genes for traits related to T2D. This study shows that more insight about the genetic architecture of T2D can be gleaned by incorporating regulatory genetic information in gene mapping studies and represents an important step forward in the post-GWAS era.

3.2 Introduction

Type 2 diabetes (T2D) is a complex disease characterized by impaired glucose homeostasis resulting from dysfunction in insulin-secreting pancreatic islets and decreased insulin sensitivity in peripheral tissues [25]. In addition to environmental factors such as a sedentary lifestyle and poor diet, genetic susceptibility is an important contributor to the development of T2D [15]. Although few genes (e.g. *TCF7L2*, *PPARG*, *KCNJ11*) have been supported by pedigree-based linkage and candidate gene association studies [126, 5, 40, 55] genome-wide association studies (GWAS) have dramatically increased the number of loci that significantly associate with either T2D or glucose-related traits [157, 125, 15]. However, the majority of single nucleotide polymorphisms (SNPs) significantly associated with T2D reside in intronic and intergenic regions rather than protein-encoding regions [67, 154]. Therefore, results from GWAS suggest an important role for genetic variation that regulates gene expression rather than altering codon sequence [2] and have motivated efforts to map the regulatory landscape of the genome [92, 147, 12]. Indeed, sets of trait-associated SNPs are enriched for variants that associate with gene expression (i.e. expression quantitative trait loci or eQTLs) [111] and that occupy DNAse hypersensitivity sites (DHS) [99] - regions overrepresented for eQTLs *per se* [27]. Moreover, DHS explain a disproportionately high share of SNP heritability [78] across 11 complex traits [60] and sets enriched for eQTLs mapped in insulin-responsive peripheral tissues similarly "concentrate" SNP heritability estimates for T2D [150].

Importantly, efforts to elucidate the functional consequences of non-coding disease-associated variants have undermined the assumption that the nearest gene to an associated marker is always the relevant disease gene. For example, a non-coding SNP (rs12740374) at the 1p13 locus associated with myocardial infarction (MI) and low-density lipoprotein cholesterol (LDL-C) creates a C/EBP transcription factor binding site and alters the expression of *SORT1* in primary human hepatocytes [105]. Moreover, *Sort1* knockdown and overexpression studies in mice altered LCL-C and very low density lipoprotein (VLDL) levels [105].

Furthermore, in a study of the *FTO* locus harboring the strongest association with obesity, researchers observed a long-range interaction between the associated intronic region of *Fto* and the promoter of *Irx3* - a downstream transcription factor located $\sim$ 500 kb away - but not with the *Fto* promoter in the brains of adult mice [138]. Perturbing *Irx3* expression in hypothalamus also reduced body mass accumulation in the background of a high fat diet and improved measures of metabolic health [138]. Lastly, obesity-associated SNPs within the locus significantly associated with *IRX3* expression in human cerebellum but not with *FTO* expression [138]. These examples demonstrate that regulatory consequences of disease-associated variants may not solely target the putative causal gene reported from GWAS, if at all. Moreover, it is unclear to what extent regulatory genetic variation supports the putative causal gene at disease-associated loci. Therefore, we sought to address this problem systematically by applying a statistical method that leverages the wealth of genotype and expression data from large-scale eQTL mapping studies.

Experimental techniques that manipulate endogenous gene expression (e.g. gene silencing, conditional knockout) can delineate relevant disease genes but are generally not suitable for *in vivo* human studies [33, 89]. However, by testing for association between the *genetic component* of gene expression and disease, we exploit the fact that nature essentially perturbs gene expression through random genetic variation introduced during meiosis. This analytic approach - implemented in the program PrediXcan - allows for a gene-based test that reflects the mechanism of transcription and presents advantages over GWAS and other study designs [48]. Namely, it considerably reduces the multiple-testing burden, obviates causality issues encountered in differential gene expression studies, provides direction of effect for associated genes, and may resolve disease-relevant tissues [48]. Moreover, PrediXcan can corroborate reported disease genes as well as implicate novel genes as was the case for an analysis of type 1 diabetes based on predictors of gene expression in whole blood tissue [48]. In this study, we applied a recent adaptation of the PrediXcan method - MetaXcan - that inputs summary GWAS data [9] (Barbeira et al. 2016) to perform a systematic *in silico* evaluation of gene-

level associations at T2D loci. We applied MetaXcan using predictive models corresponding to more than 40 human tissues to summary data from a trans-ethnic GWAS meta-analysis representing over $100,000$ individuals and replicated results in an independent cohort.

3.3 Materials and Methods

3.3.1 Determining SNP predictors of gene expression

DGN whole blood model. SNP predictors of gene expression expression in whole blood tissue were determined as described in [161] with genome-wide genotype and RNA-seq data from the Depression Genes and Networks (DGN) cohort study [12] corresponding to 922 unrelated individuals ($\hat{\pi} < 0.05$) of European ancestry. In brief, imputation of 650K SNPs with minor allele frequency (MAF) > 0.05 and non-significant departure from Hardy-Weinberg equilibrium (HWE) were imputed to a 1000 Genomes (Phase 1, version 3) reference panel [145] with ShapeIt2 [28]. The full set of $\sim$1.9 M imputed SNPs with MAF > 0.05 and imputation $R^2 > 0.8$ were subsetted to SNPs included in HapMap Phase II [77]. HCP (hidden covariates with prior) normalized gene-level expression data was downloaded from the NIMH repository [161].

GTEx tissue models. RNA-seq gene expression from $8,555$ tissue samples (representing 53 unique tissue types) from 544 subjects and imputed genotypes (available for 450 subjects and imputed to a 1000 Genomes reference panel) was obtained from the Genotype Tissue Expression Project (GTEx) data release on 2014-06-13 [147]. Expression measures from the top 40 GTEx tissues with the largest available sample sizes [147, 161] and SNPs included in HapMap Phase II ($\sim$ 2.6 M) were carried forward in our model fitting procedure.

In order to delineate a set of informative SNPs for predicting tissue-level gene expression, we performed penalized regression with the Least Absolute Shrinkage and Selection Operator (LASSO) procedure [149]. This method leverages a shrinkage parameter that per-

forms feature selection while solving for the coefficient solutions to the regression of gene expression on SNP genotypes. Gene expression - as measured by reads per kilobase of transcript per million reads mapped (RPKM) - was adjusted for potential batch effects and unmeasured confounders by regressing out the first 15 PEER factors [141] in R [124]. For each gene expressed in a tissue, model fitting was performed by regressing PEER-adjusted gene expression on the set of SNPs located within 1 Mb of the transcription start site (TSS). Therefore, subsequent analyses pertain to estimates of genetic components of gene expression attributable to local regulatory variants. The SNP coefficients from this procedure are used as weights to estimate the genetic component of gene expression and are publicly available (`http://predictdb.org`).

Decomposition of cross-tissue gene expression. As described in Wheeler *et al.* (2016), a "cross-tissue" component of gene expression is decomposed from measured gene expression with a mixed linear model where the expression of a gene for individual i in tissue t, $Y_{i,t}$, is modeled as:

$$Y_{i,t} = Y_i^{CT} + \mathbf{Z}_i\boldsymbol{\beta} + \epsilon_t \tag{3.1}$$

where Y_i^{CT} is the random subject level intercept, $\mathbf{Z}_i$ represents covariates (i.e. overall intercept, gender, and PEER factors), and $\epsilon_{i,t}$ is the error term [161]. Moreover, Y_i^{CT} and ϵ are assumed to be independent of each other and normally distributed where $Y_i^{CT} \sim N(0, \sigma^2_{CT})$ and $\epsilon \sim N(0, \sigma^2_e)$. Model parameters were estimated with the `lme4` package [11] in R [124] and the cross-tissue components of gene expression were estimated from the posterior modes of the subject random intercepts.

3.3.2 Summary data from GWAS on type 2 diabetes

DIAGRAM summary data. Input GWAS summary data used in our MetaXcan-based association of predicted gene expression and T2D corresponded to the DIAGRAM trans-ethnic meta-analysis study [125] and was publicly available and downloaded from the DIA-

GRAM website (`http://diagram-consortium.org/`). This study involved a meta-analysis of 26,488 cases and 83,964 controls subjects from populations of European, east Asian, south Asian, and Mexican, and Mexican American ancestry. Although a majority of individuals were of European ancestry (12,171 cases and 56,862 controls) [104], the study included East Asian individuals from the AGEN-T2D Consortium (6,952 cases and 11,865 controls) [20], south Asian individuals from the SAT2D Consortium (5,561 cases and 14,458 controls) [84], and individuals of Mexican and Mexican American ancestry (1,804 cases and 779 controls) [117]. SNPs were lifted to NCBI build GRCh37 (UCSC hg19 assembly).

GWAS on T2D results from GERA study. Replication analyses were performed using summary GWAS data from an analysis on the Genetic Epidemiology on Adult Health and Aging (GERA) cohort (dbGaP phs000674.p1). GERA represents a large, multi-ethnic cohort of individuals of European, East Asian, African American, and Latino ancestry where each subgroup was genome-wide genotyped with arrays designed to maximize coverage of common and low-frequency variants in each constituent population [68, 69]. T2D case status was determined from ICD-9 codes available from electronic medical health records. SNPs meeting selection criteria for MAF ($\geq 1\%$), HWE departure ($p > 10^{-6}$), and call rate ($> 95\%$) were pre-phased with SHAPEITv2.5 [28] and imputed to a 1000 Genomes reference panel with IMPUTEv2.3 [72]. GWAS on T2D was performed on a set of 71,604 unrelated ($\hat{\pi} < 0.2$) subjects (9,747 cases and 61,857 controls) with SNPTESTv2.5 [96] and adjusted for principal components (PCs) to correct for population stratification.

3.3.3 Testing for association between predicted gene expression and T2D with MetaXcan

For this study, we used MetaXcan [9], an extension of the PrediXcan method [48], that takes as input summary statistics from GWAS. This approach improves computational efficiency over PrediXcan as it does not require individual-level genotype data to estimate genetic com-

ponents of gene expression for subsequent trait association testing. Rather, the PrediXcan Z-statistic (Z_g) is approximated by:

$$Z_g \approx \sum_{l \in \text{Model}_g} w_{l,g} \frac{\sigma_l}{\hat{\sigma_g}} \frac{\hat{\beta}_l}{\text{se}(\hat{\beta}_l)} \tag{3.2}$$

where $w_{l,g}$ represents the prediction model weight for SNP l on gene g, σ_l is the standard deviation for SNP l, $\hat{\sigma_g}$ is the standard deviation of predicted expression for gene g, $\hat{\beta}_l$ is the regression coefficient for the regression of expression on the allelic dosage of SNP l.

In our MetaXcan analyses of T2D, we use regression coefficients ($\hat{\beta}_l$) from results from the DIAGRAM trans-ethnic meta-analysis of GWAS and the GWAS on T2D from the GERA study. Values for $w_{l,g}$ were generated as described above and available from the PredictDB website (`http://predictdb.org`). $\hat{\sigma}_g^2$ is estimated as:

$$\begin{aligned} \hat{\sigma}_g^2 &= \text{Var}\left(\sum_{l \in \text{Model}_g} w_{l,g} \boldsymbol{X}_l \right) \\ &= \text{Var}(\mathbf{W}_g \mathbf{X}_g) \end{aligned} \tag{3.3}$$

Where $\mathbf{W}_g$ is the vector of $w_{l,g}$ for SNPs in the model of g and $\text{Var}(\mathbf{X}_g)$ is the covariance matrix of $\mathbf{X}_g$. We use SNP information from a 1000 Genome Project reference panel (European ancestry) to the compute the variances and covariances of the SNPs used to predict gene expression.

3.3.4 Locus analysis of T2D-associated regions from the DIAGRAM study

We downloaded annotated association results for the top $1,000$ most significant SNPs from the DIAGRAM trans-ethnic meta-analysis study from the Type 2 Diabetes Genetics Portal (`http://www.type2diabetesgenetics.org/`) (Accessed March 2016). This provided information for the reported T2D genes implicated by each associated SNP marker - corresponding to 89 unique genes. We then delineated genomic regions for each reported gene by

taking the set of all significantly-associated SNPs annotated to that gene and demarcating a window bounded by the SNPs most distal to each other. We then expanded the region by 1 Mb upstream and downstream of the "boundary" SNPs. This ensured that the reported gene was included within the genomic window corresponding to the T2D-associated locus. Some windows overlapped as several putative T2D genes are located within 1 Mb of each other (e.g. *KIF11* and *IDE*).

We then performed a genome-wide MetaXcan analysis of the DIAGRAM study to test for association between predicted expression for each gene with prediction $R^2 > 1\%$ in each of the 42 tissue models described above (including the "cross-tissue" model). When visualizing the MetaXcan results at each T2D locus we considered three significance thresholds: (1) significance correcting for multiple tests within each locus; (2) significant correcting for $10,000$ tests - a number exceeding the maximum number of tests performed for any given tissue model; (3) significance correcting for the total number of tests performed across all available tissue models.

3.3.5 Meta-analysis of association results from the MetaXcan analysis of DIAGRAM and GERA cohorts

We performed a sample-sized based meta-analysis [165] of the association results from our MetaXcan analyses of the DIAGRAM and GERA studies where the Z-score (Z) was given by:

$$Z = \frac{\sum_i Z_i w_i}{\sqrt{\sum_i w_i^2}} \tag{3.4}$$

where $Z_i = \Phi^{-1}(P_i/2) * \text{sign}(\Delta_i)$, P_i is the p-value for study i, $w_i = \sqrt{(}N_i)$, N_i refers to the sample size for study i, Δ_i is the direction of effect in study i, and the overall P-value is given by:

$$P = 2\Phi(|-Z|) \tag{3.5}$$

3.3.6 Gene set enrichment analysis of MetaXcan-significant gene sets

Gene set enrichment analysis. Gene set enrichment analyses (GSEAs) were performed by comparing sets of significant genes implicated by our MetaXcan analyses with the complement set of GENCODE v18 [63] genes (∼18K) for which we can predict in any tissue model with prediction $R^2 > 1\%$. We restricted analyses to test for enrichment of pathways designated as Gene Ontology Biological Process (GO:BP) [146]. Overrepresented p-values were obtained from a parametric Fishers exact test using the Wallenius approximation and a non-central hypergeometric distribution [170]. GSEA was performed with the `GOseq` package [170] in R [124] that applies a weighting scheme to control for selection bias introduced by differences in transcript length.

Text-mining analysis Text-mining for frequently occurring terms among the set of significantly enriched GO:BP pathways ($p < 0.05$) was performed with the `tm` package [42] in R [124]. To control for commonly occurring terms among the full set of $20,033$ GO:BP pathways (i.e. Regulation), we weighted the frequency of each term among the set of enriched pathways by its overall frequency. Frequently occurring terms among no more than the top 200 significantly enriched GO:BP pathways were visualized in text plots generated with the `wordcloud` (`http://research.cens.ucla.edu/`) package in R [124].

3.3.7 Cross-phenotype comparison of T2D gene enrichment

The full set of annotated single variant results from published GWAS listed on the National Human Genome Research Institute / European Bioinformatics Institute (NHGRI-EBI) online catalogue - corresponding to $1,362$ phenotypes - was downloaded from `https://www.ebi.ac.uk/gwas/` (Accessed April 2016). The set of reported genes for each trait was tested for enrichment of genes significantly associated with T2D in our MetaXcan analyses through a resampling procedure. An empirical p-value was determined by first taking the observed count of intersecting genes between reported genes for each trait and MetaXcan-

significant T2D genes. We then generated a null distribution of counts by randomly sampling $100,000$ gene sets from the set of all GENCODE v18 [63] with prediction $R^2 > 1\%$ in at least one tissue model. Each sample was matched for the number putative genes reported for each trait and the overlap with the set of MetaXcan-significant genes was recorded. The enrichment p-value was calculated as the number of instances a sampled count value equaled or exceeded the observed count between putative trait genes and MetaXcan-significant genes.

3.4 Results

3.4.1 MetaXcan analysis of T2D loci reveals novel gene associations in the DIAGRAM study

In order to determine if regulatory genetic variation supports reported genes at T2D loci, we evaluated ~ 2 Mb genomic windows flanking putative T2D genes for MetaXcan associations. We restricted this analysis to the 89 reported genes corresponding to the top $1,000$ SNP associations from the DIAGRAM trans-ethnic meta-analysis of T2D GWAS [125]. Moreover, we performed a MetaXcan analysis for each gene expression prediction model trained on genotype and expression data from whole blood tissue from the Depression Genes and Networks (DGN) study [12] and each of 40 human tissues from the Genotype-Tissue Expression Project (GTEx) [147]. We also applied a "cross-tissue" model [161] that estimates the genetic component of gene expression shared across tissues (see methods). Lastly, we considered results at three levels of stringency: (1) gene associations that are genome-wide significant correcting for multiple testing across all tissue models; (2) gene associations that are genome-wide significant in at least one tissue model; (3) gene associations that significant correcting for multiple testing at the locus.

Among the 89 genomic windows flanking reported T2D genes, 32 windows showed significant gene-level associations - at the highest level of stringency - between estimates of the genetic component of gene expression and T2D (Table 3.1, Supplementary Figure 3.5). Of

these regions, the reported gene was MetaXcan-significant in 9 loci whereas the reported gene was not significant in 23 loci. There were 5 loci ($\sim$ 16%) where only the reported gene was significant; *FTO*, *WFS1*, *PPARG*, *HMG20A*, and *APOC1*. In addition to directly implicating genes by eschewing proximity assumptions used in GWAS, this approach gives the direction of changes in gene expression that may promote T2D. For example, increased expression of *WFS1* in multiple tissue models significantly associated with T2D (Figure. 3.1A). Similarly, increased expression of *FTO* and *HMG20A* associated with T2D whereas decreased expression of *PPARG* associated with disease status (Figure 3.1B-D).

In addition to corroborating putative disease genes, this analysis implicated genes not reported as T2D genes *per se* at T2D loci. There were 4 loci ($\sim$ 13%) where multiple genes were associated in addition to the reported gene; *TCF7L2*, *HHEX*, *JAZF1*, *TP53INP1*. Notably, at the *TCF7L2* locus, decreased expression of *TCF7L2* as well as *GPAM*, *TECTB*, *HABP2*, and *NHLRC2* and increased expression of *CASP7* and *PLEKHS1* significantly associated with T2D (Figure 3.2A). All of these genes, with the exception of *TCF7L2* itself, have not been reported as a T2D gene from the DIAGRAM study or any published GWAS listed in the NHGRI-EBI catalogue. Similarly, associations involving increased expression of *CYP26C1* at the *HHEX* locus, decreased expression of *HOXA5* at the *JAZF1* locus, and decreased expression of *CCNE2* at the *TP53INP1* locus implicate potentially novel T2D genes (Figure 3.2B-D).

We expected that a majority of MetaXcan-significant associations would correspond to previously reported T2D genes. However, we found the opposite to be more often the case as a single non-reference gene yielded a significant association - at the highest stringency - at 15 genomic regions spanning T2D genes ($\sim$ 47%)(Table 3.1). For example, at the *CDKAL1* locus, decreased expression of an unreported gene - *SOX4* - significantly associated with T2D (Figure 3.3A). Similarly, at a region spanning the *IGF2BP2* and *ST6GAL1* loci, only decreased expression of unreported gene *DGKG* showed a significant MetaXcan association (Figure 3.3B). Moreover, decreased expression of unreported gene *GTSF1L* yielded a

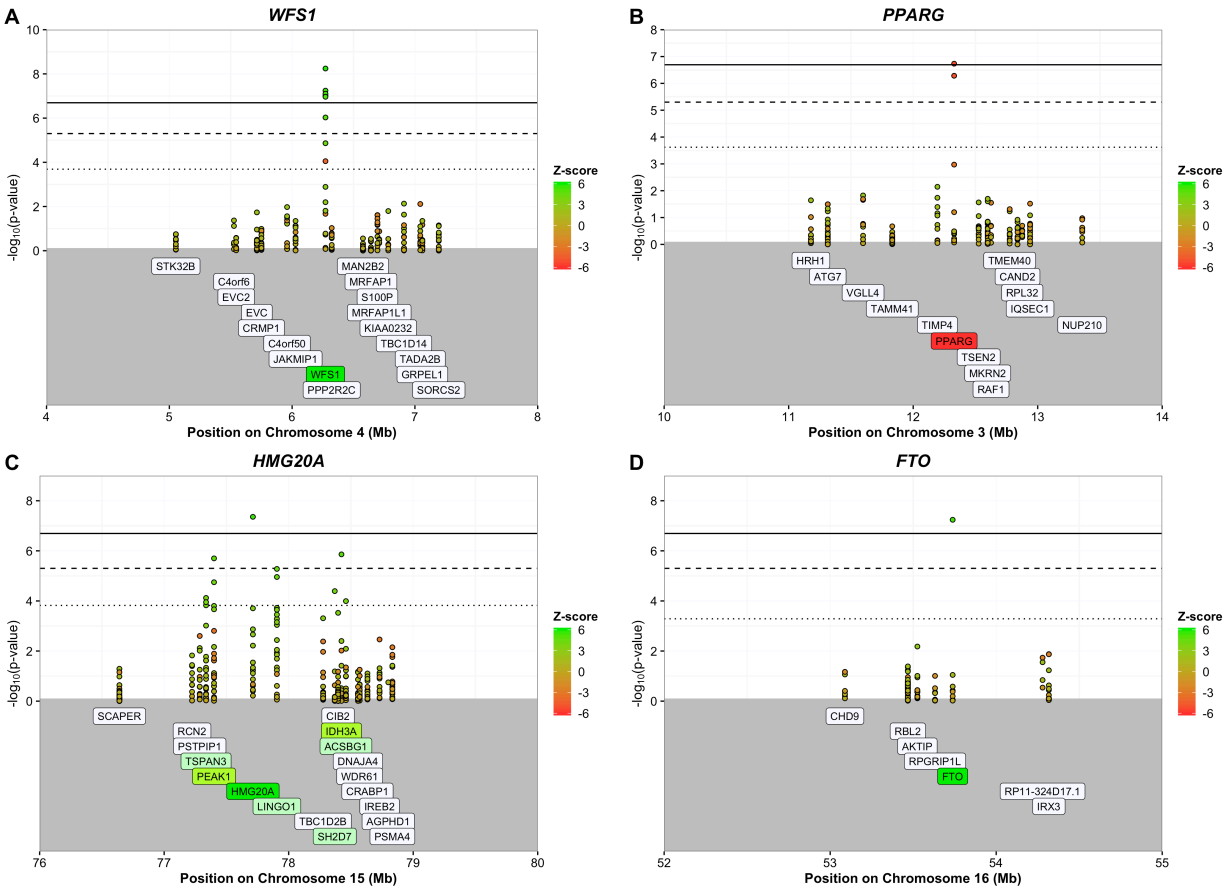

Figure 3.1: **Putative T2D genes supported by MetaXcan association in the DIAGRAM study.** We tested for association between predicted gene expression and T2D at ~ 2 Mb genomic regions encompassing putative T2D genes implicated by the top $1,000$ SNPs associated with T2D from the DIAGRAM trans-ethnic meta-analysis of GWAS using 42 tissue-level prediction models. Results are shown for regions where only the putative T2D gene is associated at the most stringent threshold: (A) *WFS1*, (B) *PPARG*, (C) *HMG20A*, and (D) *FTO*. The gene about which the genomic region is centered is indicated at the top of each panel. The solid, dashed, and dotted lines denote significance correcting for the total number of tests performed across all models, genome-wide significance in a single model ($10,000$ tests), and locus-wide significance, respectively. Genomic position (Mb) and significance ($-\log_{10}$(p-value)) for each predicted gene expression value (from a particular tissue model) are shown on the x- and y-axes, respectively. Gene labels are shown in the gray region and are positioned at the transcription start site (TSS). Moreover, the color of each point corresponds to the magnitude and sign of the Z-score where positive and negative Z-scores are colored green and red, respectively. Similarly, gene labels are colored according to the Z-score of the most significant association at a gene if it meets locus-wide (light), genome-wide significance in at least one model (medium), or significance correcting for the total number of tests performed across all models (dark).

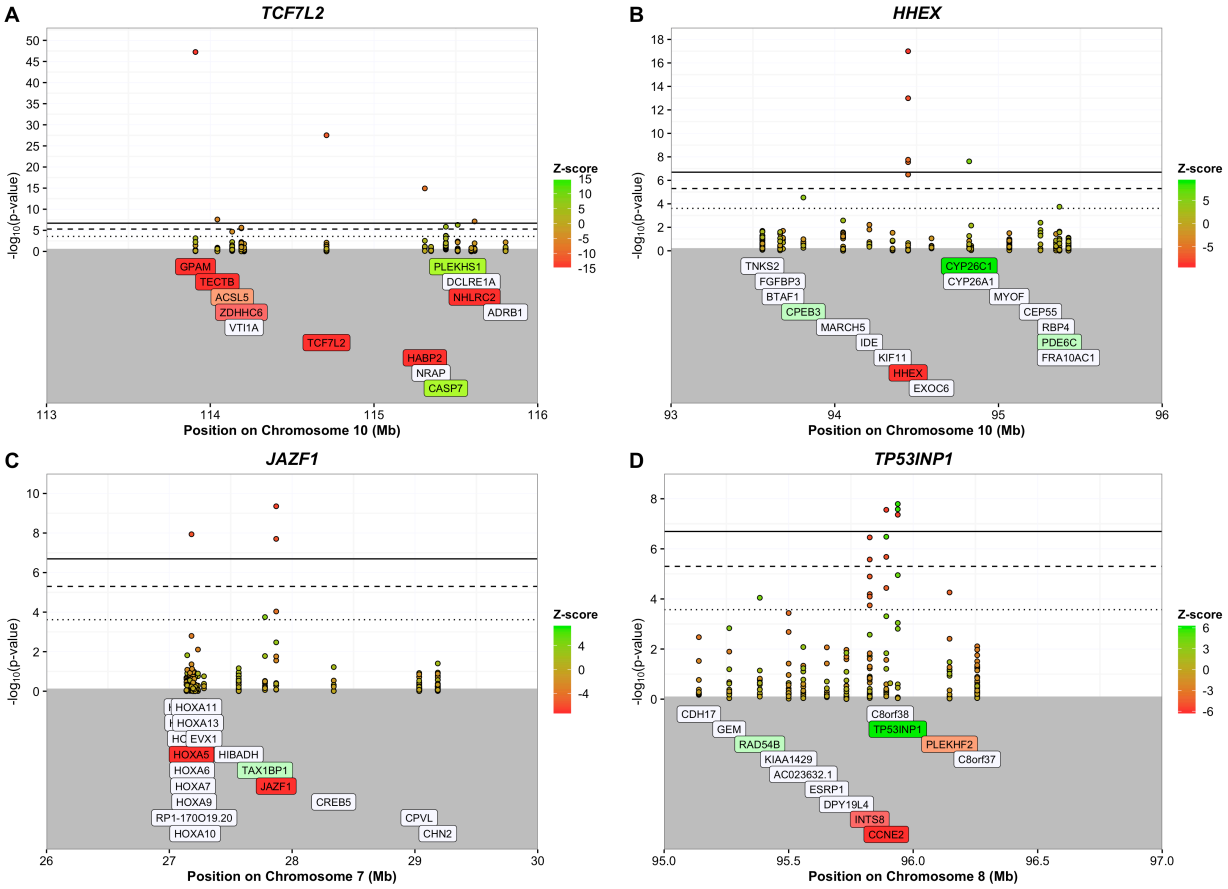

Figure 3.2: **Regions harboring MetaXcan associations with putative T2D genes and novel candidate T2D genes.** We tested for association between predicted gene expression and T2D at ~ 2 Mb genomic regions encompassing putative T2D genes implicated by the top $1,000$ SNPs associated with T2D from the DIAGRAM trans-ethnic meta-analysis of GWAS using 42 tissue-level prediction models. Results are shown for regions where both the putative T2D gene and unreported T2D genes are associated at the most stringent threshold: (A) *TCF7L2*, (B) *HHEX*, (C) *JAZF1*, and (D) *TP53INP1*. The gene about which the genomic region is centered is indicated at the top of each panel. The solid, dashed, and dotted lines denote significance correcting for the total number of tests performed across all models, genome-wide significance in a single model ($10,000$ tests), and locus-wide significance, respectively. Genomic position (Mb) and significance ($-\log_{10}$(p-value)) for each predicted gene expression value (from a particular tissue model) are shown on the x- and y-axes, respectively. Gene labels are shown in the gray region and are positioned at the transcription start site (TSS). Moreover, the color of each point corresponds to the magnitude and sign of the Z-score where positive and negative Z-scores are colored green and red, respectively. Similarly, gene labels are colored according to the Z-score of the most significant association at a gene if it meets locus-wide (light), genome-wide significance in at least one model (medium), or significance correcting for the total number of tests performed across all models (dark).

gene-level association within a region spanning reported genes *HNF4A*, *R3HDML*, *FITM2*, *GDAP1L1*, and *C20orf111* (Fig. 3.3C). Lastly, decreased expression of unreported gene *ZNF710* showed a significant association at the *C15orf38-AP3S2* locus (Figure 3.3D). However, some of these cases where a single non-reference (T2D) gene exhibits an association and not the reference gene are attributed to proximal T2D genes as was the case for increased expression of *HMG20A* at the region spanning *PEAK1* and *LINGO1* (Figure 3.1C), decreased expression of *PPARG* at a region spanning *TIMP4* and *TAMM41* (Figure 3.1B), and increased expression of *WFS1* at the *PPP2R2C* locus (Figure 3.1A).

There were also 8 loci (∼25%) where multiple non-reference genes (reported and novel alike) yielded significant MetaXcan associations. For example, decreased expression of *HHEX* (reported) and increased expression of *CYP26C1* (not previously reported) associated with T2D at the region spanning the *KIF11* and *IDE* loci (Figure 3.2B). At the *CDC123* locus, increased expression of unreported genes *CAMK1D* and *NUDT5* strongly associated with disease (Figure3.3E). Moreover, at a region spanning the *PRC1* and *VPS33B* loci downstream of the *AP3S2* locus , decreased expression of *ZNF710* and, notably, increased expression of *RCCCD1* (not previously reported) was consistently and significantly associated with T2D in multiple tissue models (Figure 3.3F).

Although the reference genes in the previous example do not associate with T2D at the highest stringency level, both *PRC1* and *VPS33B* yield genome-wide associations in tests based on a single tissue model. We therefore expanded our set of results to include gene-level associations that meet genome-wide significance in at least one tissue model to determine if we could account for more reported T2D genes. Indeed, we observed that the number of genomic windows spanning putative T2D genes and harboring MetaXcan associations increased from 32 to 48 (Table 3.1, Supplementary Figure 1). Of these instances, the reported gene yielded gene-level associations at genome-wide significance in 22 loci, including the loci corresponding to *KCNJ11*, *KLHL42*, *CPNE4*, and *ANK1*. However, the majority of MetaXcan-significant genes are potentially novel genes as 30 out of 52 genes were unreported

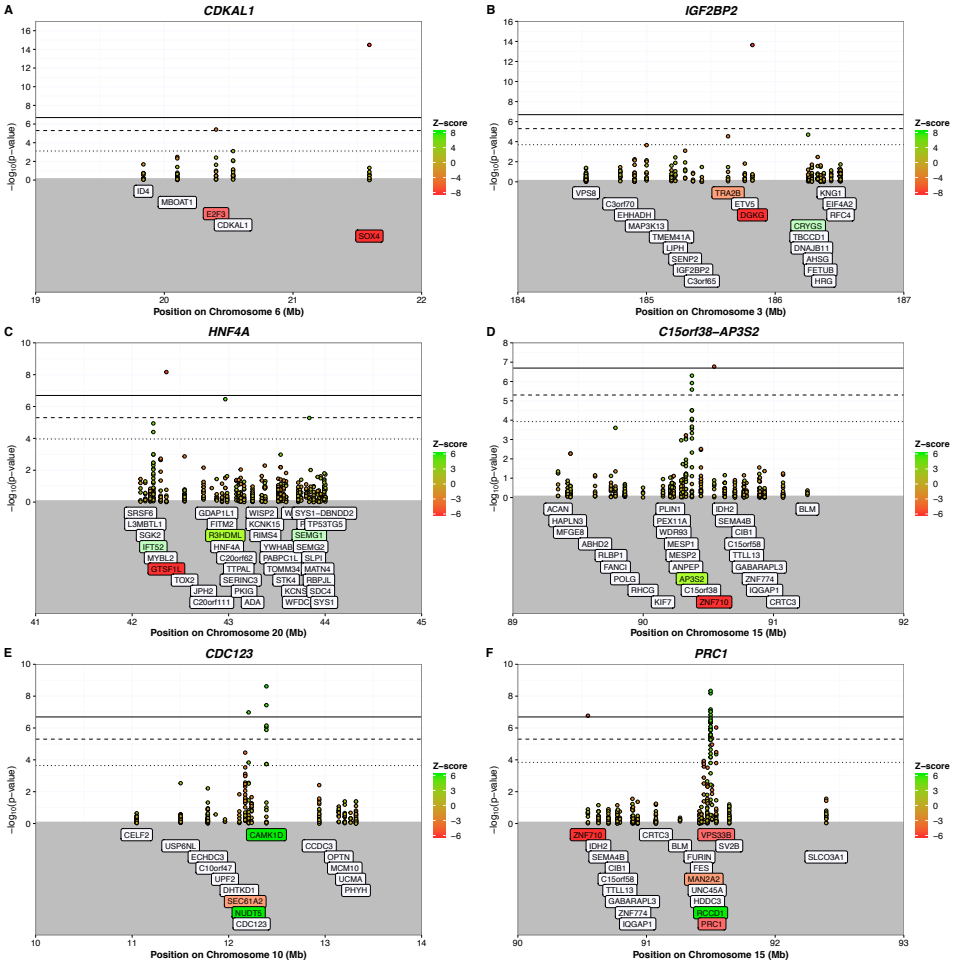

Figure 3.3: **Regions where novel T2D genes show strong MetaXcan association with T2D.** Results are shown for regions where unreported T2D genes are associated at the most stringent threshold: (A) *CDKAL1*, (B) *IGF2BP2*, (C) *HNF4A*, (D) *C15orf38-AP3S2*, (E) *CDC123*, and (F) *PRC1*. The solid, dashed, and dotted lines denote significance correcting for the total number of tests performed across all models, genome-wide significance in a single model (10,000 tests), and locus-wide significance, respectively. Genomic position (Mb) and significance ($-\log_{10}$(p-value)) for each predicted gene expression value (from a particular tissue model) are shown on the x- and y-axes, respectively. Gene labels are shown in the gray region and are positioned at the transcription start site (TSS). Moreover, the color of each point corresponds to the magnitude and sign of the Z-score where positive and negative Z-scores are colored green and red, respectively

in the DIAGRAM study at this significance level and 14 out of 23 genes were unreported at the highest stringency level (Table 3.1).

Lastly, we found the disparity between reported and unreported genes to be more pronounced when we considered the set of all gene associations that met locus-wide significance (i.e. correcting for the number of tests performed at a single locus). Under this criterion, 75 out of the 89 genomic windows harbored significant MetaXcan associations and implicated 109 genes (Supplementary Figure 3.5). Of these genes, however, only 34 ($\sim$ 31%) corresponded to putative T2D genes. Although these results provided additional support for reported genes such as *ABCC8*, *TSPAN8*, and *MC4R*, they further showed that most of the discovered associations between predicted gene expression and T2D correspond to novel T2D genes.

3.4.2 Novel T2D genes at putative T2D loci replicate in the GERA-T2D study

In our MetaXcan analysis of T2D loci in the DIAGRAM study, we observed a total of 102 significant associations between predicted gene expression and T2D that were genome-wide significant in at least one tissue - implicating a total of 52 genes (Table 3.1). In order to delineate a more robust set of T2D genes (including potentially novel disease genes), we performed a separate analysis of these associations in an independent cohort - the Resource for Genetic Epidemiology Research on Adult Health and Aging (GERA) study. This genetic study (a collaboration between the Kaiser Permanente Research Program on Genes, Environment, and Health and the UCSF Institute for Human Genetics) represents a multiethnic cohort of 100K individuals from Northern California with available electronic medical records (EMRs). We performed MetaXcan analyses using summary statistics from a GWAS of T2D cases and matched controls from the GERA study.

Of the 102 genome-wide significant associations observed in our DIAGRAM analysis, 63 associations ($\sim$ 62%) and 25 genes ($\sim$48%) replicated in the GERA-T2D analysis (Table

3.1). Of these 25 genes, 10 were reported T2D genes and 15 were putative novel genes. The reported T2D genes that replicated in GERA were *PPARG*, *WFS1*, *JAZF1*, *TCF7L2*, *HHEX*, *KCNJ11*, *AP3S2*, *PEAK1*, *PRC1*, and *HMG20A* and the putative novel genes were *DGKG*, *E2F3*, *SOX4*, *HOXA5*, *NHLRC2*, *GPAM*, *HABP2*, *TECTB*, *PLEKHS1*, *CAMK1D*, *CYP26C1*, *CLPB*, *KLHDC5*, *RCCD1*, and *GTSF1L*. Among the set of reported T2D genes, the direction of association was mostly consistent across all models for which each gene was significant. The only exception being *KCNJ11* where increased expression of *KCNJ11* in the testis and DGN whole blood models associated with T2D whereas decreased expression of *KCNJ11* in the esophagus (mucosa) and skin (sun exposed) models associated with T2D (Table 3.1). Similarly, the direction of association was consistent across models for which a putative T2D gene was significantly associated with T2D. Notably, increased expression of *RCCD1* was highly significant across 24 tissue models in both DIAGRAM and GERA studies. However, *KLHDC5* was an exception as increased expression of *KLHDC5* in the DGN whole blood study associated with T2D in DIAGRAM whereas decreased expression associated with T2D in GERA.

We also considered MetaXcan associations that were significant in the DIAGRAM study in at least one model at a false discovery rate (FDR) $< 5\%$ (Supplementary Table 3.2) . This corresponded to 150 associations (at T2D-associated regions; there were 184 genome-wide associations) and 76 genes, of which 80 associations and 32 genes replicated at this threshold, including two additional reported genes (*LINGO1* and *NCR3LG1*) and 5 additional novel genes (*CDKN2C*, *CYP21A2*, *NKX6-3*, *CDKN2A*, *P2RX4*). Lastly, we performed a meta-analysis of results from our MetaXcan analyses of the DIAGRAM and GERA studies and observed that 22 of the 25 replicated genes (i.e. Bonferroni-significant in at least one model) more strongly associated with T2D - the exceptions being *E2F3*, *HABP2*, and *KLHDC5* (Table 3.1).

Chr	Gene	Reported Genes within 1Mb Locus	Model	DIAGRAM Z–score	DIAGRAM P–value	GERA Z–score	GERA P–value	Meta–analysis Z–score	Meta–analysis P–value
1	*UTS2*	none reported	WholeBloodDGN	5	6.3e–07	–0.49	0.62	3.6	0.00038
3	*CPNE4*	*CPNE4*	CrossTissue	4.8	1.4e–06	1.5	0.14	4.3	1.5e–05
3	*CPNE4*	*CPNE4*	Adipose Subcutaneous	4.8	1.8e–06	1.5	0.13	4.4	1.3e–05
3	*PPARG*	*PPARG,TIMP4,TAMM41,*	Cells EBV transformedlymphocytes	–5	5.2e–07	–3.7	0.00026	–5.7	1.1e–08
3	*PPARG*	*PPARG,TIMP4,TAMM41,*	Heart LeftVentricle	–5.2	1.8e–07	0.14	0.89	–4	6.7e–05
3	*UBE2E1*	*UBE2E2*	Artery Coronary	4.6	3.5e–06	1.1	0.26	4.2	3.2e–05
3	*DGKG*	*IGF2BP2,ST6GAL1*	Ovary	–7.6	2.3e–14	–6.3	2.8e–10	–8.8	1.3e–18
3	*NR1D2*	*UBE2E2*	Skin SunExposed Lowerleg	4.8	1.5e–06	1.8	0.072	4.5	8.6e–06
4	*WFS1*	*WFS1,PPP2R2C*	CrossTissue	5.3	1.1e–07	4.8	1.5e–06	6.3	2.7e–10
4	*WFS1*	*WFS1,PPP2R2C*	Brain Cerebellum	5.3	9.6e–08	4	5.8e–05	6	2.3e–09
4	*WFS1*	*WFS1,PPP2R2C*	Nerve Tibial	4.9	9.4e–07	4.6	4.3e–06	5.9	3e–09
4	*WFS1*	*WFS1,PPP2R2C*	Skin NotSunExposed Suprapubic	5.8	5.7e–09	4.6	4.7e–06	6.7	1.7e–11
4	*WFS1*	*WFS1,PPP2R2C*	Skin SunExposed Lowerleg	5.4	5.7e–08	4.7	2.9e–06	6.3	3e–10
4	*WFS1*	*WFS1,PPP2R2C*	Thyroid	5.4	7.9e–08	4.6	3.9e–06	6.5	8e–11
6	*E2F3*	*CDKAL1*	Brain Cerebellum	–4.6	3.9e–06	–2.2	0.03	–4.5	5.8e–06
6	*HCG27*	*CCHCR1,POU5F1,HLA–C, TCF19*	Heart LeftVentricle	4.5	5.6e–06	1.2	0.23	4.1	3.5e–05
6	*SOX4*	*CDKAL1*	Muscle Skeletal	–7.9	3.3e–15	–4.9	8.5e–07	–8.4	3.7e–17
6	*HLA–A*	none reported	Pituitary	6.1	1.2e–09	–0.77	0.44	4.3	2e–05
6	*ZNRD1*	none reported	WholeBloodDGN	–5.5	3.6e–08	0.3	0.77	–4	5.2e–05
7	*JAZF1*	*JAZF1*	Muscle Skeletal	–6.2	4.4e–10	–2.5	0.014	–6	2.5e–09
7	*JAZF1*	*JAZF1*	Pancreas	–5.6	2e–08	–2.9	0.0044	–5.6	1.8e–08
7	*HOXA5*	*JAZF1*	WholeBlood	–5.7	1.2e–08	–6.1	1e–09	–7.4	1.8e–13
8	*TP53INP1*	*TP53INP1,DPY19L4,INTS8, NDUFAF6*	Esophagus Muscularis	5.6	1.6e–08	0.82	0.41	4.7	2.1e–06
8	*TP53INP1*	*TP53INP1,DPY19L4,INTS8, NDUFAF6*	Testis	–5.5	4.3e–08	–0.51	0.61	–4.5	7.8e–06
8	*TP53INP1*	*TP53INP1,DPY19L4,INTS8, NDUFAF6*	WholeBloodDGN	5.6	2.6e–08	1.3	0.21	4.9	1.2e–06
8	*INTS8*	*TP53INP1,DPY19L4,INTS8, NDUFAF6*	Artery Tibial	–4.7	2.7e–06	NA	NA	NA	NA
8	*INTS8*	*TP53INP1,DPY19L4,INTS8, NDUFAF6*	WholeBloodDGN	–5.1	3.5e–07	–1.7	0.083	–4.7	2.2e–06
8	*NDUFAF6*	*NDUFAF6*	Brain Cerebellum	–5	4.7e–07	1	0.32	–3.3	0.00093
8	*ANK1*	*ANK1*	CrossTissue	5.1	2.8e–07	–0.8	0.42	3.5	0.00052
8	*VDAC3*	*ANK1*	AdrenalGland	–5	4.6e–07	–1.9	0.057	–5	5.1e–07

Table 3.1: **MetaXcan associations with T2D.** Results for genes and corresponding models that meet genome-wide significance *in at least one model* from the DIAGRAM analysis are shown with nearby genes and results from the GERA replication study and meta-analysis of DIAGRAM and GERA Metaxcan associations. Blue shading denotes genes not implicated by the top 1,000 SNPs from the DIAGRAM trans-ethnic meta-analysis of GWAS. Pink and red shading denote genome-wide significance in one model and across all models, respectively, for the DIAGRAM and meta-analysis. Replication in the GERA study is denoted by a pink outline.

Chr	Gene	Reported Genes within 1Mb Locus	Model	DIAGRAM Z–score	DIAGRAM P–value	GERA Z–score	GERA P–value	Meta–analysis Z–score	Meta–analysis P–value
8	*CCNE2*	*TP53INP1,DPY19L4,INTS8, NDUFAF6*	Cells EBV transformedlymphocytes	5.1	3.3e–07	–0.71	0.48	3.6	0.00032
8	*CCNE2*	*TP53INP1,DPY19L4,INTS8, NDUFAF6*	Lung	–4.7	2.1e–06	–0.1	0.92	–3.7	0.00023
8	*CCNE2*	*TP53INP1,DPY19L4,INTS8, NDUFAF6*	WholeBloodDGN	–5.6	2.8e–08	–0.88	0.38	–4.7	3.1e–06
10	*TCF7L2*	*TCF7L2*	Artery Aorta	–11	3e–28	–9.2	2.8e–20	–13	3.3e–40
10	*HHEX*	*HHEX,KIF11,IDE,*	CrossTissue	–5.6	1.8e–08	–0.69	0.49	–4.6	5e–06
10	*HHEX*	*HHEX,KIF11,IDE,*	Breast MammaryTissue	–8.6	1e–17	–3.6	0.00034	–8.4	5.2e–17
10	*HHEX*	*HHEX,KIF11,IDE,*	Cells Transformedfibroblasts	–5.1	3.2e–07	–3.9	1e–04	–5.9	4.2e–09
10	*HHEX*	*HHEX,KIF11,IDE,*	Testis	–5.6	2.8e–08	–2.9	0.0034	–5.7	1e–08
10	*HHEX*	*HHEX,KIF11,IDE,*	WholeBloodDGN	–7.4	1e–13	–3.2	0.0015	–7.2	5.1e–13
10	*NHLRC2*	*TCF7L2*	AdrenalGland	–5.4	7.7e–08	–3.2	0.0013	–6	2.5e–09
10	*CASP7*	*TCF7L2*	Artery Aorta	4.8	1.4e–06	0.95	0.34	4.2	2.2e–05
10	*GPAM*	*TCF7L2*	Artery Tibial	–15	5.7e–48	–10	1.5e–25	–16	2.1e–60
10	*ZDHHC6*	*TCF7L2*	Brain Anteriorcingulatecortex BA24	–4.6	3.4e–06	0.12	0.9	–3.6	0.00036
10	*ZDHHC6*	*TCF7L2*	Muscle Skeletal	–4.7	2.3e–06	0.79	0.43	–3.2	0.0013
10	*HABP2*	*TCF7L2*	Brain CerebellarHemisphere	–8	1.2e–15	–13	2e–39	–13	9.6e–37
10	*TECTB*	*TCF7L2*	Brain Putamen basalganglia	–5.6	2.7e–08	–5.2	2.2e–07	–6.9	5.6e–12
10	*PLEKHS1*	*TCF7L2*	Esophagus Mucosa	5	5e–07	6.4	1.1e–10	7.2	4.8e–13
10	*CAMK1D*	*CDC123*	Heart LeftVentricle	4.9	9e–07	3.8	0.00013	5.7	9.6e–09
10	*CAMK1D*	*CDC123*	Lung	6	2.4e–09	1.2	0.25	5.1	2.8e–07
10	*CAMK1D*	*CDC123*	Spleen	5.5	3.7e–08	2.4	0.015	5.4	5.7e–08
10	*CAMK1D*	*CDC123*	WholeBlood	4.8	1.3e–06	1.6	0.12	4.5	7.8e–06
10	*CAMK1D*	*CDC123*	WholeBloodDGN	5	7e–07	2.7	0.0072	5.1	3.5e–07
10	*CYP26C1*	*HHEX,KIF11,IDE,*	Pancreas	5.6	2.4e–08	3.8	0.00013	6.1	1.2e–09
10	*NUDT5*	*CDC123*	WholeBloodDGN	5.3	1e–07	1.5	0.14	4.8	1.8e–06
11	*KCNJ11*	*KCNJ11,ABCC8,NCR3LG1,*	Esophagus Mucosa	–4.7	2.2e–06	–1.2	0.22	–4.4	1.4e–05
11	*KCNJ11*	*KCNJ11,ABCC8,NCR3LG1,*	Skin SunExposed Lowerleg	–4.9	8e–07	–2.7	0.0073	–5	6.6e–07
11	*KCNJ11*	*KCNJ11,ABCC8,NCR3LG1,*	Testis	4.7	2.4e–06	2	0.044	4.6	3.6e–06
11	*KCNJ11*	*KCNJ11,ABCC8,NCR3LG1,*	WholeBloodDGN	4.7	3.3e–06	2.1	0.037	4.6	4.9e–06

Table 3.1 (Continued): **MetaXcan associations with T2D.** Results for genes and corresponding models that meet genome-wide significance *in at least one model* from the DIAGRAM analysis are shown with nearby genes and results from the GERA replication study and meta-analysis of DIAGRAM and GERA Metaxcan associations. Blue shading denotes genes not implicated by the top $1,000$ SNPs from the DIAGRAM trans-ethnic meta-analysis of GWAS. Pink and red shading denote genome-wide significance in one model and across all models, respectively, for the DIAGRAM and meta-analysis. Replication in the GERA study is denoted by a pink outline.

Chr	Gene	Reported Genes within 1Mb Locus	Model	DIAGRAM Z–score	DIAGRAM P–value	GERA Z–score	GERA P–value	Meta–analysis Z–score	Meta–analysis P–value
11	*CLPB*	*ARAP1,STARD10,FCHSD2,*	Artery Coronary	−4.7	2.8e−06	−3.8	0.00017	−5.5	3.5e−08
11	*SYT8*	*KCNQ1*	Esophagus GastroesophagealJunction	4.9	1.2e−06	1.1	0.27	4.3	1.5e−05
12	*KLHL42*	*KLHL42*	Artery Tibial	−4.7	2.8e−06	NA	NA	NA	NA
12	*MPHOSPH9*	*MPHOSPH9*	Heart AtrialAppendage	−4.4	8.7e−06	NA	NA	NA	NA
12	*C12orf65*	*MPHOSPH9*	Nerve Tibial	4.7	2.1e−06	NA	NA	NA	NA
12	*KLHDC5*	*KLHL42*	WholeBloodDGN	5	5.9e−07	−2.1	0.036	2.8	0.0055
15	*AP3S2*	*C15orf38−AP3S2,AP3S2*	CrossTissue	4.7	2.6e−06	3.2	0.0013	5.1	3.8e−07
15	*AP3S2*	*C15orf38−AP3S2,AP3S2*	AdrenalGland	5	5e−07	3.5	0.00041	5.9	4.8e−09
15	*AP3S2*	*C15orf38−AP3S2,AP3S2*	Stomach	4.9	1.2e−06	3.5	0.00045	5.5	3.2e−08
15	*LINGO1*	*HMG20A,PEAK1,LINGO1,*	Brain Cortex	4.6	5.3e−06	−0.25	0.81	3.4	0.00072
15	*PEAK1*	*HMG20A,PEAK1,LINGO1,*	Spleen	4.8	2e−06	3.2	0.0016	5.2	1.9e−07
15	*PRC1*	*PRC1,VPS33B*	Pancreas	−4.6	4.5e−06	−3.3	0.00089	−5.1	3.7e−07
15	*HMG20A*	*HMG20A,PEAK1,LINGO1,*	Pituitary	5.5	4.4e−08	3.4	0.00065	5.9	4.5e−09
15	*VPS33B*	*PRC1,VPS33B*	Testis	−4.9	9.4e−07	−1.1	0.28	−4.3	1.6e−05
15	*RCCD1*	*PRC1,VPS33B*	Adipose Subcutaneous	4.6	4.9e−06	2.9	0.0033	4.9	9e−07
15	*RCCD1*	*PRC1,VPS33B*	Artery Aorta	5.3	1.5e−07	3	0.0023	5.6	1.7e−08
15	*RCCD1*	*PRC1,VPS33B*	Artery Coronary	4.9	1.1e−06	2.5	0.012	5	4.7e−07
15	*RCCD1*	*PRC1,VPS33B*	Artery Tibial	5.1	3.3e−07	3	0.0028	5.4	6.1e−08
15	*RCCD1*	*PRC1,VPS33B*	Brain Anteriorcingulatecortex BA24	5.8	6.6e−09	3.4	0.00067	6.2	4.5e−10
15	*RCCD1*	*PRC1,VPS33B*	Brain Caudate basalganglia	5.2	2.3e−07	2.8	0.0048	5.4	8.6e−08
15	*RCCD1*	*PRC1,VPS33B*	Brain CerebellarHemisphere	4.7	2.5e−06	3.1	0.0017	5.2	2.3e−07
15	*RCCD1*	*PRC1,VPS33B*	Brain Cerebellum	4.6	3.7e−06	3.1	0.0018	5	5.6e−07
15	*RCCD1*	*PRC1,VPS33B*	Brain Cortex	4.6	3.7e−06	3.1	0.0018	5.1	3.3e−07
15	*RCCD1*	*PRC1,VPS33B*	Brain FrontalCortex BA9	5	4.9e−07	3.1	0.002	5.4	6.3e−08
15	*RCCD1*	*PRC1,VPS33B*	Breast MammaryTissue	4.8	1.4e−06	3	0.003	5.2	2.1e−07
15	*RCCD1*	*PRC1,VPS33B*	Cells Transformedfibroblasts	5.4	6.9e−08	3.5	0.00053	5.9	3.9e−09
15	*RCCD1*	*PRC1,VPS33B*	Colon Sigmoid	4.8	1.3e−06	3.1	0.0019	5.5	4.2e−08
15	*RCCD1*	*PRC1,VPS33B*	Esophagus GastroesophagealJunction	4.9	8e−07	2.8	0.0044	5.3	1.4e−07

Table 3.1 (Continued): **MetaXcan associations with T2D.** Results for genes and corresponding models that meet genome-wide significance *in at least one model* from the DIAGRAM analysis are shown with nearby genes and results from the GERA replication study and meta-analysis of DIAGRAM and GERA Metaxcan associations. Blue shading denotes genes not implicated by the top 1,000 SNPs from the DIAGRAM trans-ethnic meta-analysis of GWAS. Pink and red shading denote genome-wide significance in one model and across all models, respectively, for the DIAGRAM and meta-analysis. Replication in the GERA study is denoted by a pink outline.

Chr	Gene	Reported Genes within 1Mb Locus	Model	DIAGRAM Z–score	DIAGRAM P–value	GERA Z–score	GERA P–value	Meta–analysis Z–score	Meta–analysis P–value
15	*RCCD1*	*PRC1,VPS33B*	Heart LeftVentricle	5.9	4.7e–09	2.8	0.0056	5.9	2.8e–09
15	*RCCD1*	*PRC1,VPS33B*	Muscle Skeletal	5	4.6e–07	3.1	0.002	5.4	8.3e–08
15	*RCCD1*	*PRC1,VPS33B*	Nerve Tibial	5.3	9.1e–08	3.2	0.0013	5.6	2.3e–08
15	*RCCD1*	*PRC1,VPS33B*	Pancreas	5	4.5e–07	3	0.003	5.3	1.5e–07
15	*RCCD1*	*PRC1,VPS33B*	Skin NotSunExposed Suprapubic	4.7	3e–06	3.2	0.0016	5.1	2.7e–07
15	*RCCD1*	*PRC1,VPS33B*	Skin SunExposed Lowerleg	4.7	2.7e–06	3	0.0027	4.9	7.6e–07
15	*RCCD1*	*PRC1,VPS33B*	SmallIntestine TerminalIleum	4.5	5.9e–06	3.1	0.0021	5.3	1.3e–07
15	*RCCD1*	*PRC1,VPS33B*	Stomach	5.1	3.9e–07	3.1	0.0022	5.5	4.4e–08
15	*RCCD1*	*PRC1,VPS33B*	Testis	5.3	9.5e–08	3.2	0.0016	5.7	1.4e–08
15	*RCCD1*	*PRC1,VPS33B*	Thyroid	5.1	3.4e–07	3	0.0027	5.5	4.4e–08
15	*ZNF710*	*PRC1,VPS33B,C15orf38–AP3S2, AP3S2*	Brain Caudate basalganglia	–5.2	1.7e–07	NA	NA	NA	NA
15	*IDH3A*	*HMG20A,PEAK1,LINGO1,*	Brain Hippocampus	4.8	1.4e–06	1.6	0.11	4.6	4.3e–06
15	*ZFAND6*	none reported	Brain CerebellarHemisphere	4.5	7.1e–06	–1.4	0.16	2.8	0.006
15	*FAH*	none reported	Brain Cerebellum	4.5	6.3e–06	–1.2	0.21	2.8	0.0051
16	*FTO*	*FTO*	Muscle Skeletal	5.4	5.8e–08	1.7	0.085	5	6.3e–07
19	*APOC1*	*APOC1*	Esophagus Mucosa	5.3	1.2e–07	NA	NA	NA	NA
19	*KLC3*	*APOC1*	Brain Cerebellum	4.5	6.4e–06	1.9	0.055	4.3	1.5e–05
19	*CLASRP*	*APOC1*	Heart LeftVentricle	4.6	3.8e–06	0.5	0.62	3.8	0.00012
20	*R3HDML*	*HNF4A,R3HDML,FITM2, GDAP1L1,C20orf111*	Skin NotSunExposed Suprapubic	5.1	3.5e–07	0.54	0.59	4.2	3e–05
20	*SEMG1*	*HNF4A,R3HDML,FITM2, GDAP1L1,C20orf111*	Artery Coronary	4.6	5.2e–06	0.14	0.89	3.6	0.00032
20	*GTSF1L*	*HNF4A,R3HDML,FITM2, GDAP1L1,C20orf111*	Cells Transformedfibroblasts	–5.8	6.9e–09	–2.9	0.0034	–5.9	3e–09

Table 3.1 (Continued): **MetaXcan associations with T2D.** Results for genes and corresponding models that meet genome-wide significance *in at least one model* from the DIAGRAM analysis are shown with nearby genes and results from the GERA replication study and meta-analysis of DIAGRAM and GERA Metaxcan associations. Blue shading denotes genes not implicated by the top $1,000$ SNPs from the DIAGRAM trans-ethnic meta-analysis of GWAS. Pink and red shading denote genome-wide significance in one model and across all models, respectively, for the DIAGRAM and meta-analysis. Replication in the GERA study is denoted by a pink outline.

3.4.3 Genome-wide MetaXcan analysis reveals novel T2D loci in the DIAGRAM study

We found significant MetaXcan associations for both reported and novel T2D genes within genomic regions encompassing the top 89 genes implicated from GWAS in the DIAGRAM trans-ethnic study. However, we considered that this approach may support genes that reside in regions that extend beyond the genomic windows evaluated in the T2D locus-based analyses (i.e. novel T2D loci). We therefore surveyed all genome-wide significant associations (in at least one tissue) and observed that the majority of gene-level associations mapped to regions spanning known T2D loci (Figure 3.4). However, there were five associations that corresponded to genes falling outside putative T2D loci; increased expression of *UTS2*, *HLA-A*, and *ZFAND6* and decreased expression of *ZNRD1* and *FAH* (Table 3.1). With the exception of *ZFAND6*, these genes were not reported as putative T2D genes in any published GWAS study listed in the NHGRI-EBI catalogue [157].

In order to glean insight into relevant biological pathways suggested by our gene sets, we performed gene set enrichment analyses (GSEAs) with our MetaXcan results. Among the complete set of genes that met genome-wide significance in our analysis of the DIAGRAM study, there was a significant enrichment for gene ontology (GO) pathways (i.e. GO: Biological Process) pertaining to the endocrine pancreas and metabolism (e.g. negative regulation of type B pancreatic cell apoptotic process, regulation of insulin secretion, glucose homeostasis) (Supplementary Table 3.3). Similar results were observed among significant genes from the meta-analysis of MetaXcan results from the DIAGRAM and GERA analyses (Supplementary Table 3.4). On the other hand, the GO pathways most enriched from the set of "unknown" T2D loci genes related to circadian rhythm and kidney function (e.g. positive regulation of circadian sleep/wake cycle, REM sleep, negative regulation of glomerular filtration) (Supplementary Table 3.5). Moreover, we text-mined the complete set of significantly enriched GO pathways - adjusting for baseline frequency (see methods) - for frequently occurring terms. Among pathways enriched from the full set of significantly-associated genes,

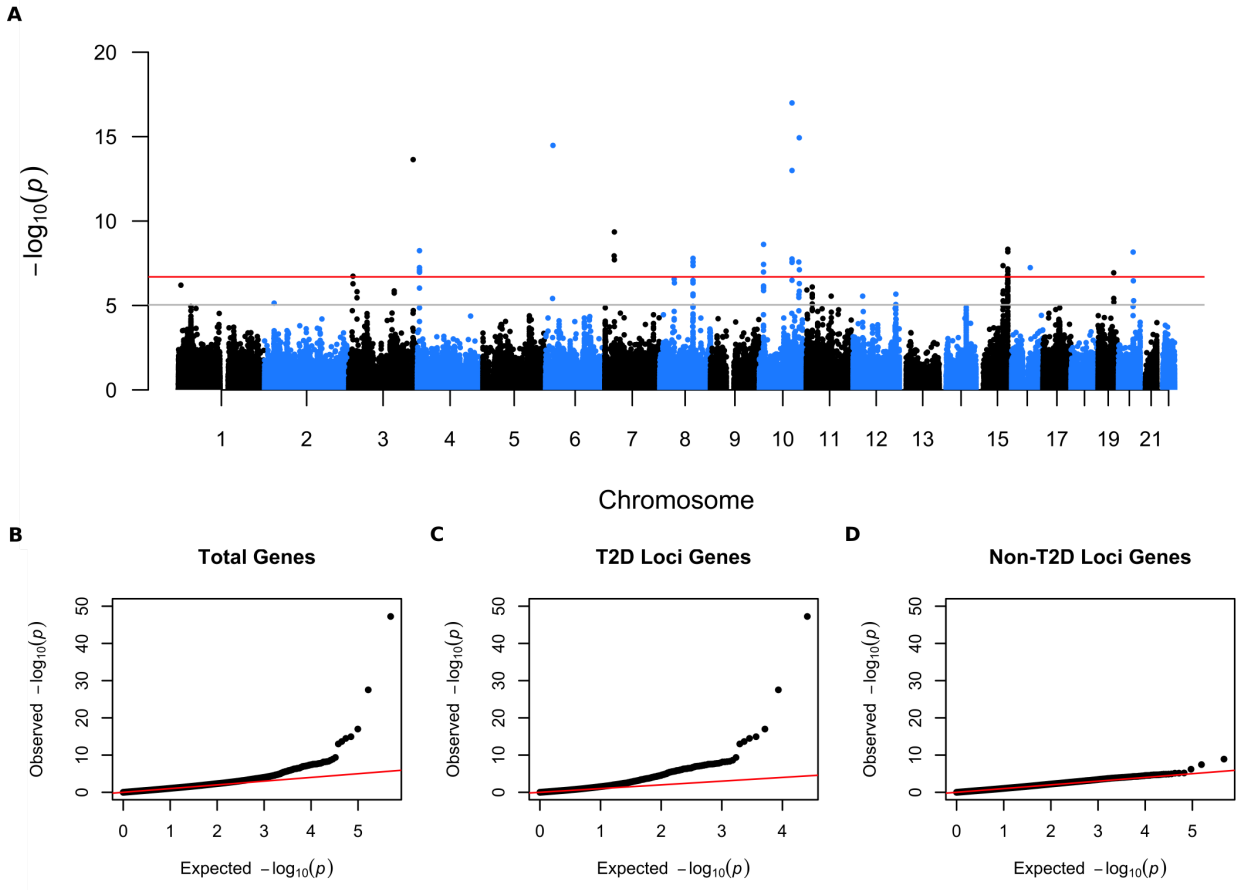

Figure 3.4: **Genome-wide MetaXcan analyses shows gene associations constrained to putative T2D loci.** We used MetaXcan to estimate gene expression and test for T2D association in the DIAGRAM trans-ethnic meta-analysis summary dataset for each of 42 predictive models (i.e. 42 separate association tests). (A) Manhattan plot showing significance ($-\log_{10}$(p-value)) of associations between predicted expression and T2D. Results across all tissue-level analyses are shown. The solid red line denotes significance threshold corrected for the total number of individual tests performed across all models and the solid gray line denotes the suggestive line corresponding to significance in at least one model (i.e. $10,000$ tests). (B) QQ-plot showing the expected versus observed distribution of $-\log_{10}$(p-value) for each predicted gene expression trait across all 42 association tests. (C) QQ-plot showing the expected versus observed distribution of $-\log_{10}$(p-value) across all 42 association tests for all genes within 1 Mb of T2D-associated SNPs. (D) QQ-plot showing the expected versus observed distribution of $-\log_{10}$(p-value) across all 42 association tests for all genes exceeding 1 Mb of T2D-associated SNPs.

the most frequently occurring terms corresponded WNT-signaling (e.g. beta-catenin-tcf7l2), diabetes treatment (i.e. metformin), and sugar metabolism (e.g d-ribose) (Supplementary Figure 3.6A). Conversely, "wakefulness" was the most frequent term among pathways enriched from "unknown" loci genes (Supplementary Figure 3.6B). A similar enrichment profile was observed among genes that were genome-wide significant from the meta-analysis of DIAGRAM and GERA results (Supplementary Figure 3.6C). However, this was not the case for the "unknown" loci genes as they did not replicate in the GERA study (Table 3.1, Supplementary Figure 3.6D).

3.4.4 MetaXcan-significant genes support shared genetic etiology across multiple complex traits

A key aspect of MetaXcan is that genes are implicated based on integrating information from SNP-level GWAS associations and ability to predict gene expression. This implies that the strength of GWAS signals is reflected in the significance of gene-level associations. Therefore, an instance where a reported gene from GWAS corresponds with a significant MetaXcan association strengthens evidence that the reported gene - via a regulatory process - is the likely causal gene. We observed that although ∼58-61% of MetaXcan-significant genes mapping to regions encompassing T2D loci in the DIAGRAM study are putative novel T2D genes, this study provides additional support for ∼40% of reported T2D genes. Indeed, the set of MetaXcan-significant genes from DIAGRAM was highly enriched for T2D genes reported from all published GWAS on T2D listed in the NHGRI-EBI catalogue ($p < 1\text{x}10^{-5}$) (Supplementary Table 3.6). However, enrichment of reported genes corresponding to other traits among the set of significant T2D genes implicated in this study may support shared genetic etiology with T2D. For example, we also observed a significant enrichment for putative genes related to quantitative traits relevant to T2D; BMI ($p = 3.5\text{x}10^{-4}$), fasting insulin-related traits ($p = 0.0017$), glycated hemoglobin levels ($p = 0.0024$), and proinsulin levels ($p = 0.034$) (Supplementary Table 3.6). Moreover, we observed enrichment

for additional traits related to T2D via glucose homeostasis; two-hour glucose challenge ($p = 0.016$) and fasting glucose related traits ($p = 0.05$) (Supplementary Table 5). An advantage of this approach is that relationships between T2D and other complex traits can be corroborated via novel T2D genes not implicated by GWAS. For example, a shared genetic etiology between T2D and both total cholesterol ($p = 0.033$) and LDL-cholesterol ($p = 0.031$) is corroborated by our enrichment analysis. These results are attributable to the inclusion of putative cholesterol genes *APOC1* (also a reported T2D gene) and *GPAM* - a potentially novel T2D gene suggested by our MetaXcan results. Moreover, *DGKG* - a novel candidate gene replicated in our analyses - has been previously implicated with weight and body mass index (BMI) through proximity to GWAS associations. Additional candidate genes that have been implicated with other diseases include *SOX4* (bone mineral density) and *HLA-A* (IgE levels, vitiligo, and nasopharyngeal carcinoma, among others) (Supplementary Table 3.6).

3.5 Discussion

We found that 75 out of 89 T2D genes implicated by GWAS are within 1 Mb of a significant association between predicted gene expression and disease. Moreover, 32 of these genomic regions harbor MetaXcan associations meeting significance correcting for the number of genome-wide tests performed across 42 tissue models. Although we provided corroborating support for established T2D genes (i.e. *PPARG*, *JAZF1*, *TCF7L2*), the majority ($\sim$ 60%) of genes implicated by our analyses are novel disease genes. Of the 109 genome-wide significant associations discovered in our analysis of the DIAGRAM dataset, 63 replicated in an independent cohort from the GERA study. These corresponded to 25 T2D genes - 15 of which are novel candidates. Despite the fact that most of our discoveries were novel, the set of MetaXcan-significant genes were enriched for pathways directly related to T2D pathobiology (e.g. glucose metabolism and regulation of insulin secretion).

We also found evidence for genes located in genomic regions greater than 1 Mb away

from the 89 T2D genes implicated by the the top SNPs associations in the DIAGRAM trans-ethnic meta-analysis study - four of which had not been previously reported from GWAS. Moreover, we found this gene set to be enriched for pathways regulating circadian rhythm - an important hormonal process whose dysregulation has been linked to insulin resistance in peripheral tissues and increased risk for T2D [17, 29, 36]. We also observed that novel T2D genes discovered with MetaXcan were shared between our results and putative genes for related traits such as body mass index (*DGKG*) and LDL cholesterol (*GPAM*).

Although we expected most significant associations between predicted gene expression and T2D to correspond to models trained on tissues directly involved in glucose homeostasis (i.e. pancreas, liver, skeletal muscle, adipose, hypothalamus), we observed that this was not the case for many associations meeting genome-wide significance. For example, the association with decreased expression of *TCF7L2* was observed in our MetaXcan analysis with predictors trained in aortic artery whereas Shu *et al.* (2009) observed decreased gene and protein expression of *TCF7L2* in pancreatic islets from diabetic *db/db* mice [135]. On the other hand, Savic *et al.* (2011) observed that mice harboring a null *TCF7L2* allele showed improved glucose tolerance and lower insulin levels [132]. While it may be possible that decreased *TCF7L2* expression in aortic artery may indeed increase disease risk *per se*, it may be the case that our results are largely attributable to regulatory genetic variation shared across tissues. Therefore, we may be able to detect gene associations in cases where we have limited power to predict expression for a particular T2D gene in the tissue most relevant to disease but are suitably powered to do so in a separate tissue that shares eQTLs for that gene.

Despite the fact that significant SNP associations with T2D at the *TCF7L2* locus are non-coding and imply a regulatory mechanism in insulin-responsive peripheral tissues, significant eQTLs for *TCF7L2* were mapped only in aortic artery by the GTEx Project [147]. Our observed MetaXcan association with *TCF7L2* may therefore reflect the fact that the aortic artery sample afforded us the greatest power to predict *TCF7L2* in any model. More-

over, similarity in regulatory genetic architecture shared across tissues enabled us to detect a relationship to T2D. This suggests that although we trained models from tissue samples from the DGN study and GTEx Project - representing the most comprehensive set of primary human tissues for eQTL studies to date - more work will need to be done to generate informative predictive models in order to resolve the most relevant tissues. For example, we could delineate predictive models from cell-specific data such islet cells from the endocrine pancreas or incorporate information from "response" eQTLs that associate with gene expression in the context of physiological perturbations (i.e. insulin response or hyperglycemia) or at different development stages. Moreover, additional information on chromatin structure and epigenetic state from experimental assays (e.g. ATAC-seq [18], Capture-C [74] performed on cells from metabolic tissues can be incorporated into predictive models of gene expression.

An important caveat of this study is that we used average expression over a gene when generating predictive models and may therefore miss the consequences of regulatory variants that impact splicing and the expression of isoforms at T2D loci. Although this would not diminish our positive findings (i.e. this does not create false positives), this may explain why we failed to detect genome-wide significant associations at regions encompassing putative T2D genes.

Additionally, the predictive models employed in this study were trained from local variants within 1 Mb of each gene. Although most eQTLs mapped in human tissues to date are local eQTLs, this is influenced by the fact that the greater number of genetic variants, smaller haplotype structure, and relative smaller sample sizes associated with human studies considerably reduces power to detect distal eQTLs that regulate target genes through a non allele-specific mechanism (i.e. *trans* eQTLs) [2]. However, a few studies indicate that distal-acting eQTLs mapped in human adipose and skeletal muscle tissue may account for some of the genetic architecture of T2D [38, 150]. Therefore, we may expand the number of predicted gene associations to include target genes regulated distal-acting, *trans* eQTLs by extending the training set of variants when generating predictive models.

In our study, we applied MetaXcan to explicitly integrate regulatory genetic information to improve disease gene mapping and overcome key limitations of GWAS and differential gene expression studies [48]. This approach, along with similar approaches adopted by Gusev *et al.*(2015) and Zhu *et al.*(2016), directly addresses the importance of eQTLs in complex human traits and advances genetic studies beyond GWAS [59, 174]. Importantly, we provide information about the direction of gene expression that associates with disease, that was predominantly consistent across the most significant associations discovered in this study and replicated in an independent cohort. This immediately suggests potential therapeutic targets where the increased expression of genes - many of which were not previously reported from GWAS - significantly relates to increased disease risk. Moreover, these results establish a basis for subsequent experiments (e.g. gene editing) to interrogate the cellular and physiological consequences of dysregulation of novel candidate genes. Therefore, this investigation represents an important step forward in elucidating the genetic basis of T2D and other complex diseases.

3.6 Supplementary Figures and Tables

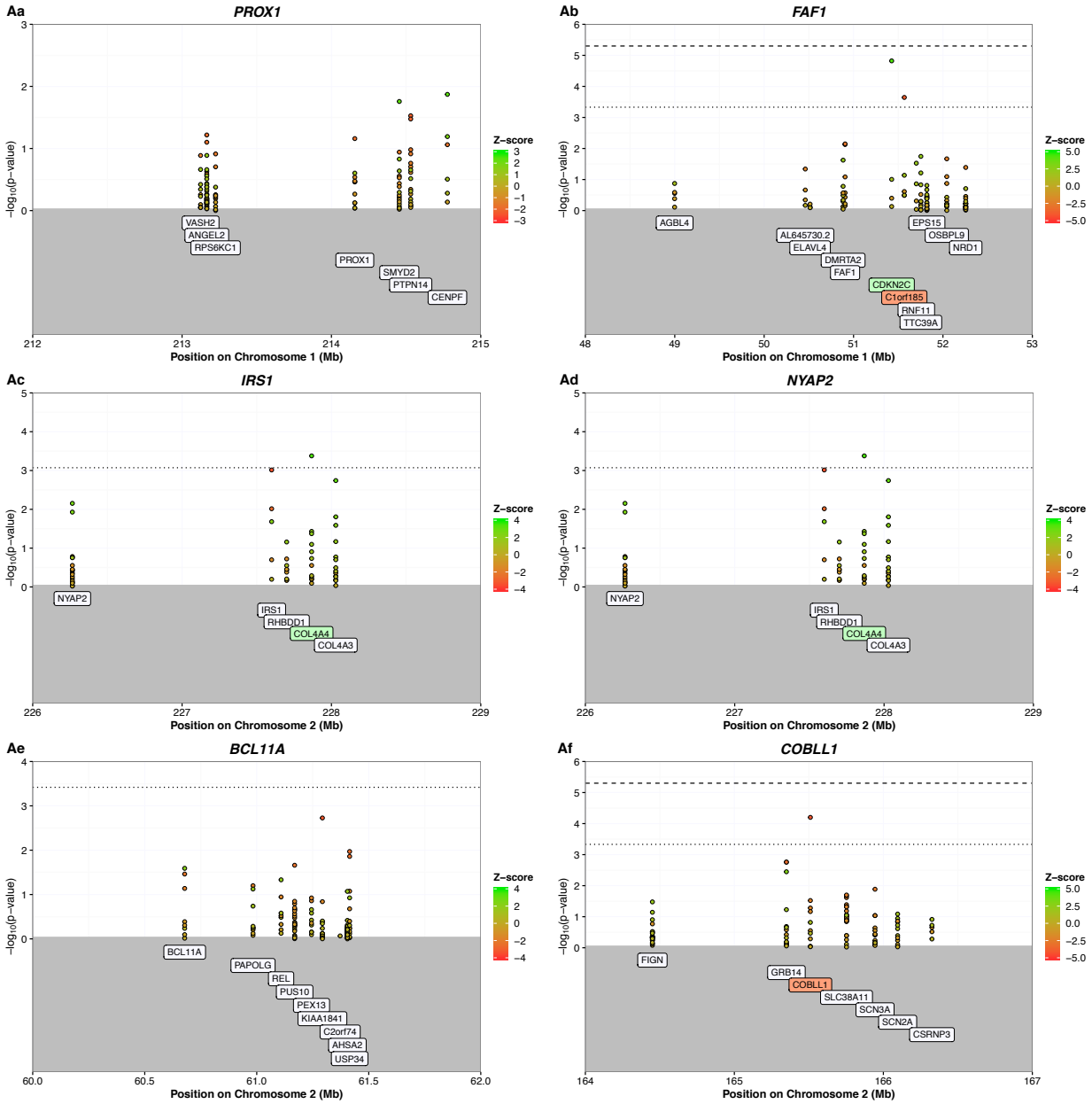

Figure 3.5: **MetaXcan profiles at T2D-associated loci** The solid, dashed, and dotted lines denote significance correcting for the total number of tests performed across all models, genome-wide significance in a single model (10,000 tests), and locus-wide significance, respectively. Genomic position (Mb) and significance ($-\log_{10}$(p-value)) for each predicted gene expression value (from a particular tissue model) are shown on the x- and y-axes, respectively. Gene labels are shown in the gray region and are positioned at the transcription start site (TSS). Moreover, the color of each point corresponds to the magnitude and sign of the Z-score where positive and negative Z-scores are colored green and red, respectively.

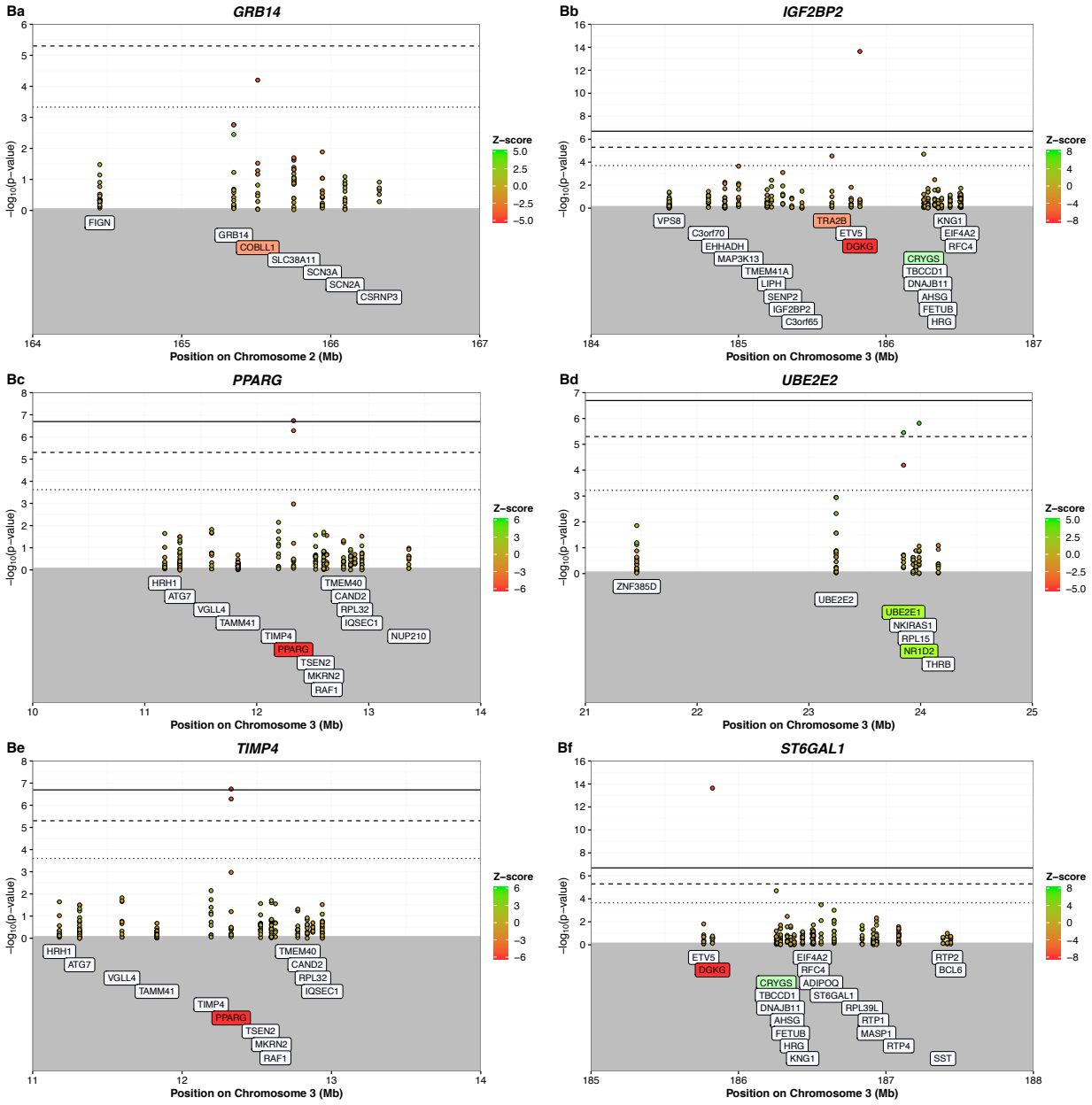

Figure 3.5 (Continued): **MetaXcan profiles at T2D-associated loci** The solid, dashed, and dotted lines denote significance correcting for the total number of tests performed across all models, genome-wide significance in a single model (10,000 tests), and locus-wide significance, respectively. Genomic position (Mb) and significance ($-\log_{10}$(p-value)) for each predicted gene expression value (from a particular tissue model) are shown on the x- and y-axes, respectively. Gene labels are shown in the gray region and are positioned at the transcription start site (TSS). Moreover, color corresponds to the magnitude and sign of the Z-score where positive and negative Z-scores are colored green and red, respectively.

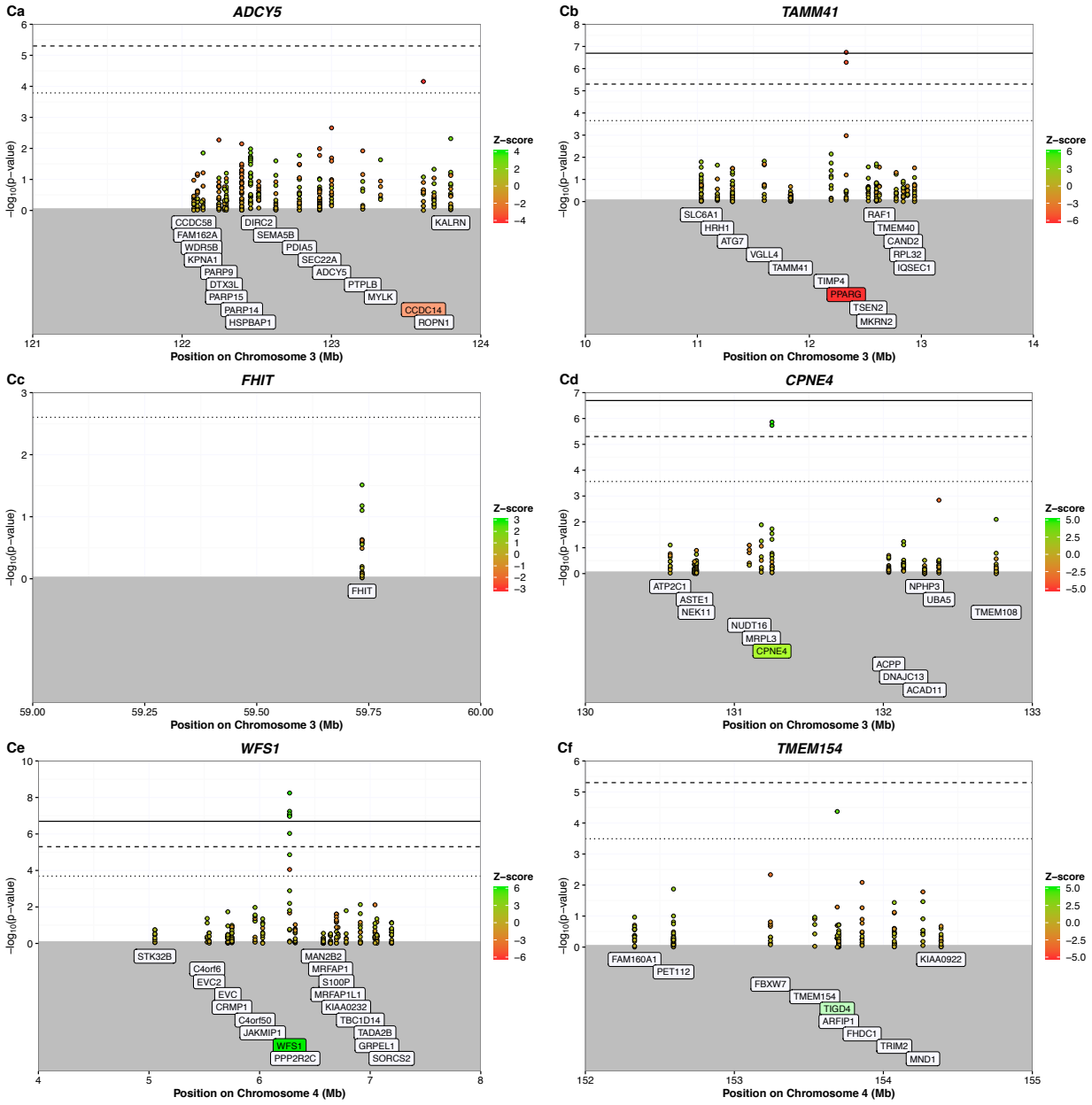

Figure 3.5 (Continued): **MetaXcan profiles at T2D-associated loci** The solid, dashed, and dotted lines denote significance correcting for the total number of tests performed across all models, genome-wide significance in a single model (10,000 tests), and locus-wide significance, respectively. Genomic position (Mb) and significance ($-\log_{10}$(p-value)) for each predicted gene expression value (from a particular tissue model) are shown on the x- and y-axes, respectively. Gene labels are shown in the gray region and are positioned at the transcription start site (TSS). Moreover, color corresponds to the magnitude and sign of the Z-score where positive and negative Z-scores are colored green and red, respectively.

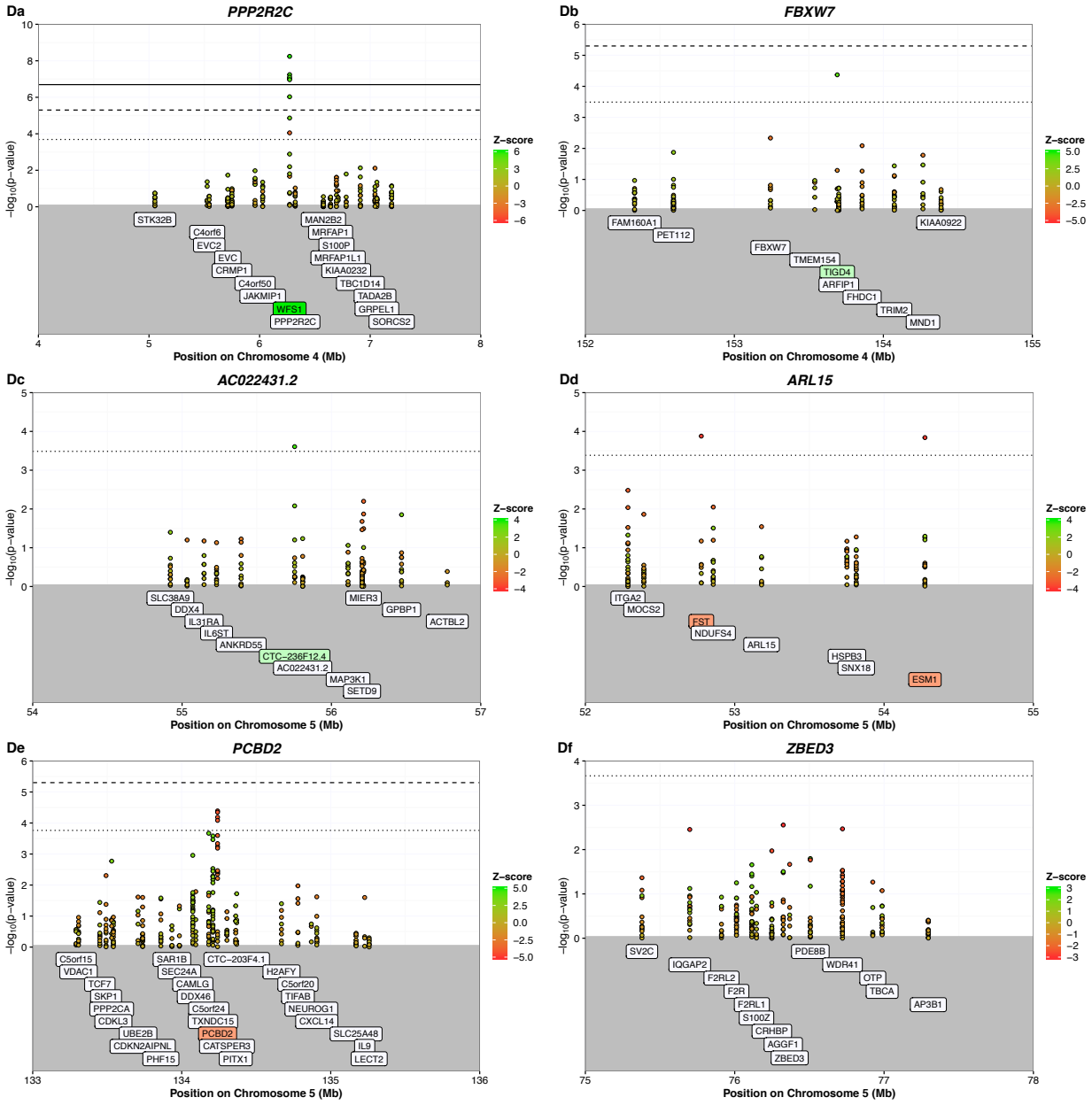

Figure 3.5 (Continued): **MetaXcan profiles at T2D-associated loci** The solid, dashed, and dotted lines denote significance correcting for the total number of tests performed across all models, genome-wide significance in a single model (10,000 tests), and locus-wide significance, respectively. Genomic position (Mb) and significance ($-\log_{10}$(p-value)) for each predicted gene expression value (from a particular tissue model) are shown on the x- and y-axes, respectively. Gene labels are shown in the gray region and are positioned at the transcription start site (TSS). Moreover, color corresponds to the magnitude and sign of the Z-score where positive and negative Z-scores are colored green and red, respectively.

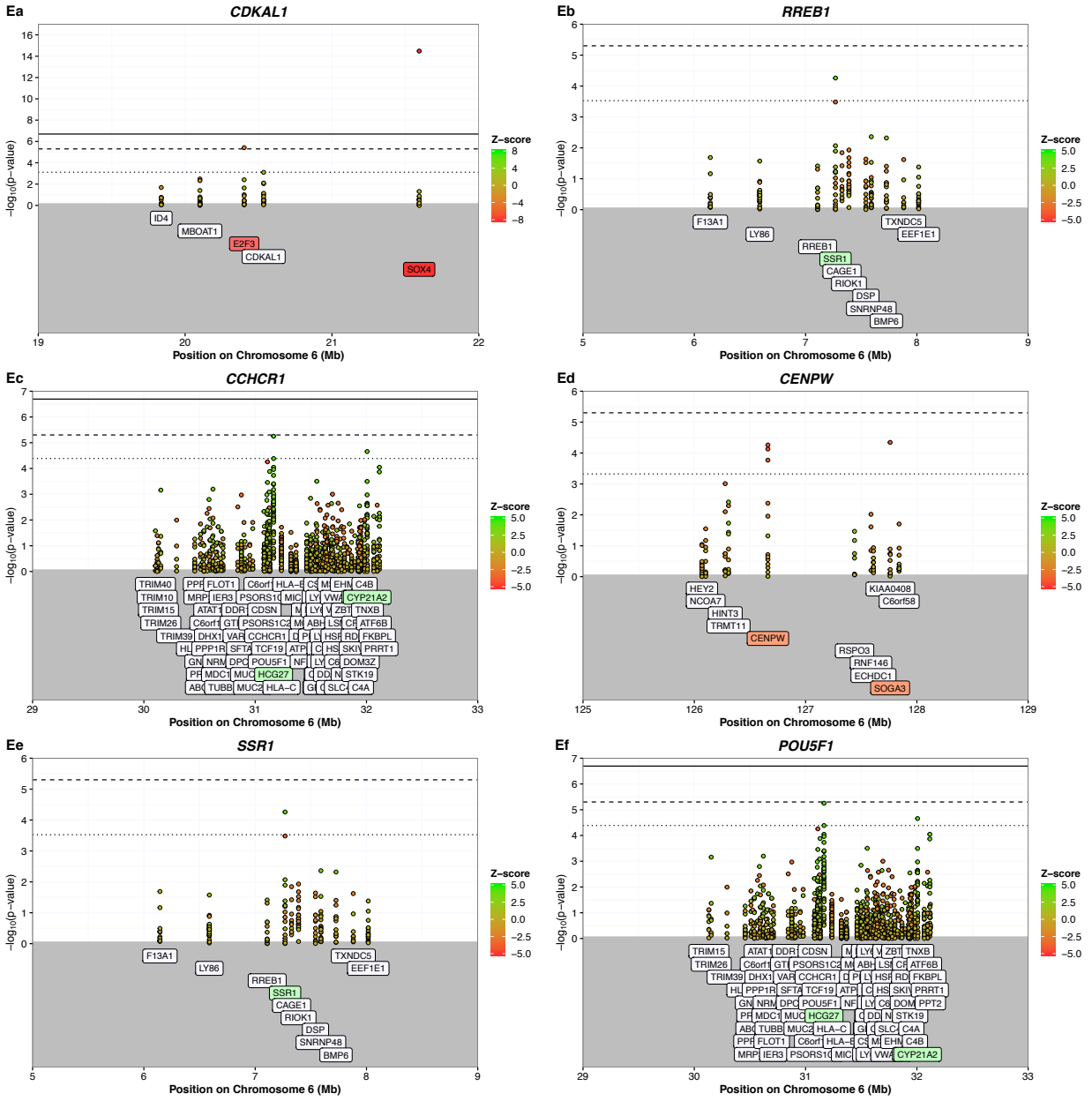

Figure 3.5 (Continued): **MetaXcan profiles at T2D-associated loci** The solid, dashed, and dotted lines denote significance correcting for the total number of tests performed across all models, genome-wide significance in a single model (10,000 tests), and locus-wide significance, respectively. Genomic position (Mb) and significance ($-\log_{10}$(p-value)) for each predicted gene expression value (from a particular tissue model) are shown on the x- and y-axes, respectively. Gene labels are shown in the gray region and are positioned at the transcription start site (TSS). Moreover, color corresponds to the magnitude and sign of the Z-score where positive and negative Z-scores are colored green and red, respectively.

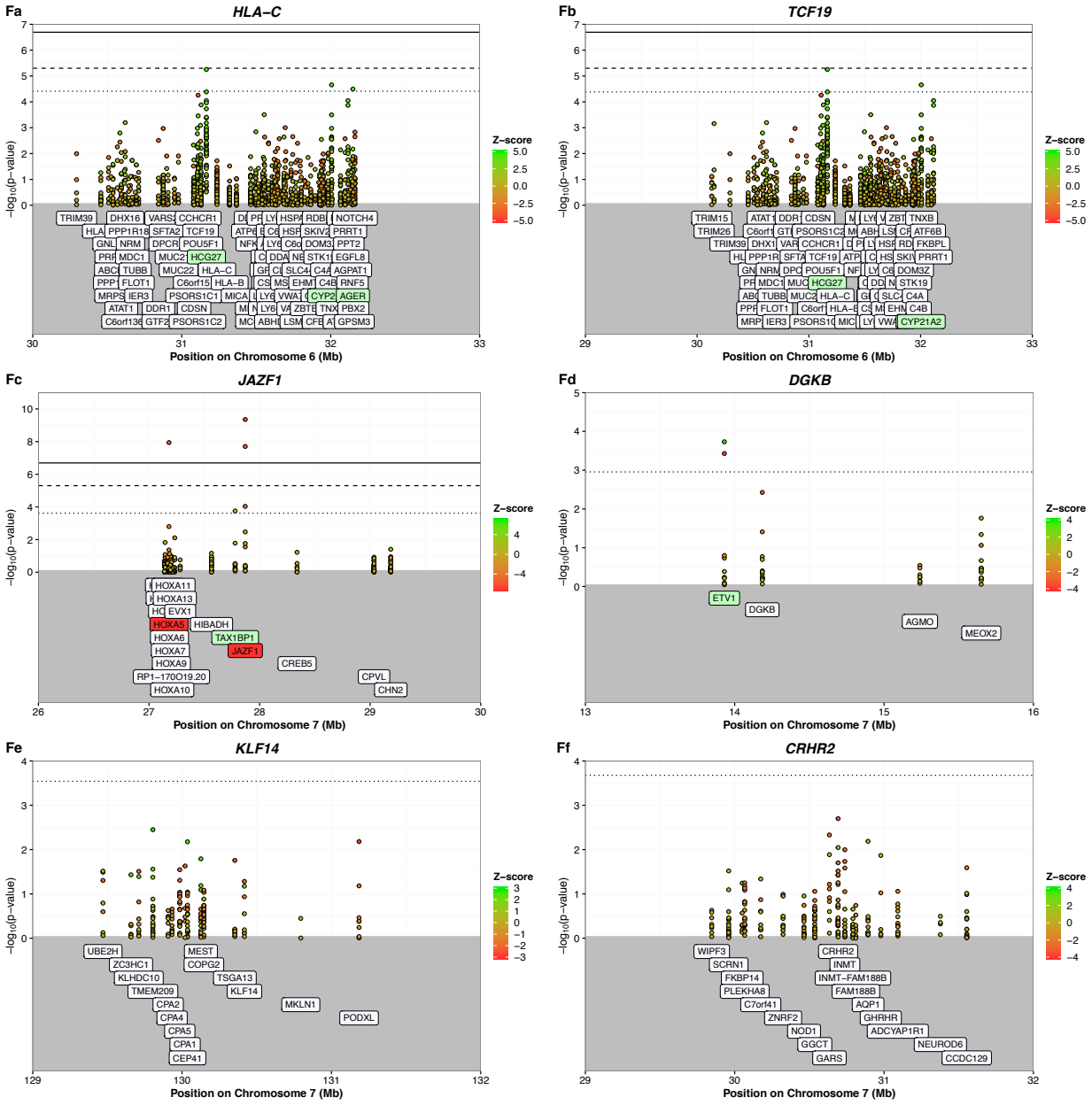

Figure 3.5 (Continued): **MetaXcan profiles at T2D-associated loci** The solid, dashed, and dotted lines denote significance correcting for the total number of tests performed across all models, genome-wide significance in a single model (10,000 tests), and locus-wide significance, respectively. Genomic position (Mb) and significance ($-\log_{10}$(p-value)) for each predicted gene expression value (from a particular tissue model) are shown on the x- and y-axes, respectively. Gene labels are shown in the gray region and are positioned at the transcription start site (TSS). Moreover, color corresponds to the magnitude and sign of the Z-score where positive and negative Z-scores are colored green and red, respectively.

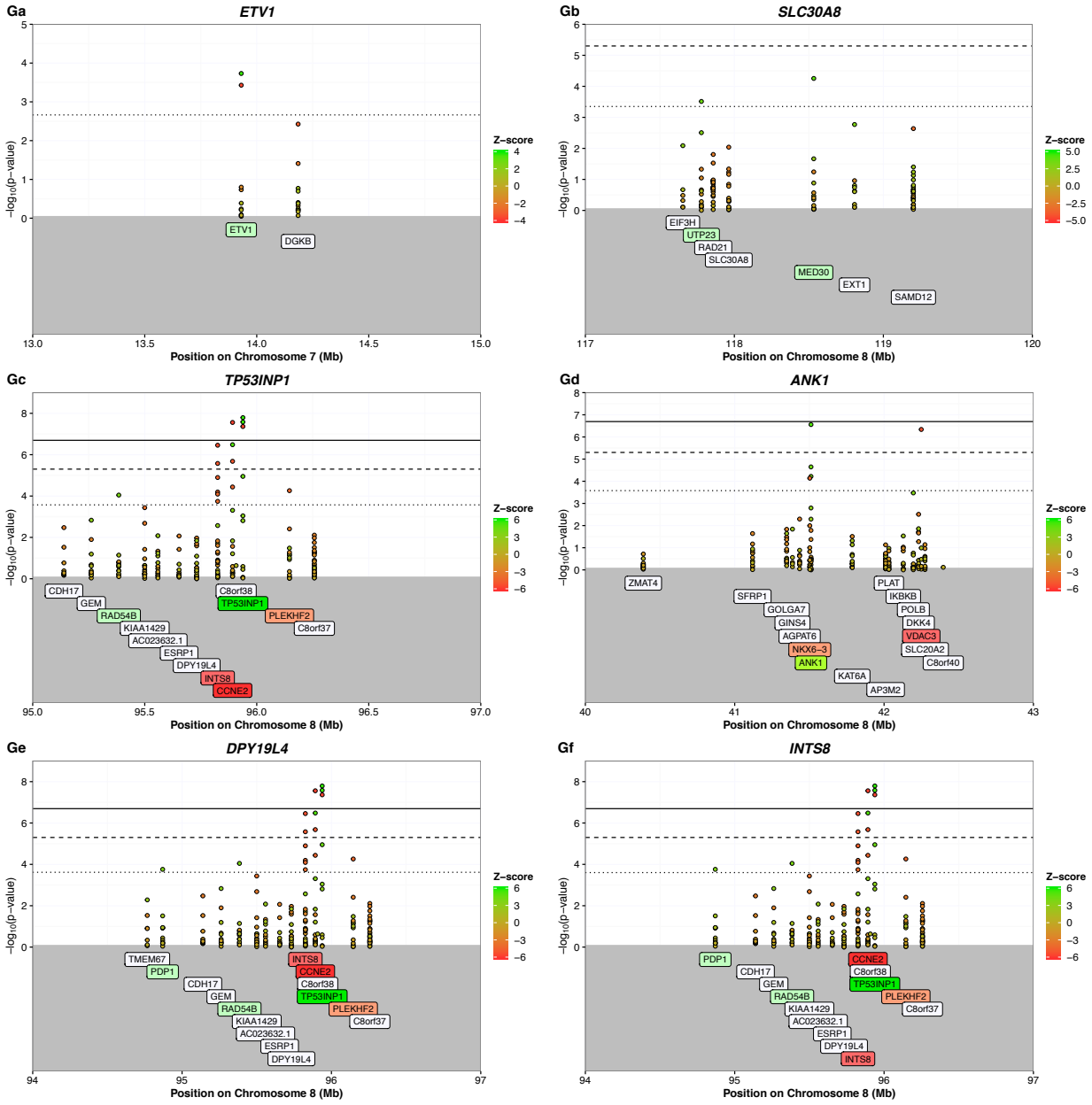

Figure 3.5 (Continued): **MetaXcan profiles at T2D-associated loci** The solid, dashed, and dotted lines denote significance correcting for the total number of tests performed across all models, genome-wide significance in a single model (10,000 tests), and locus-wide significance, respectively. Genomic position (Mb) and significance ($-\log_{10}$(p-value)) for each predicted gene expression value (from a particular tissue model) are shown on the x- and y-axes, respectively. Gene labels are shown in the gray region and are positioned at the transcription start site (TSS). Moreover, color corresponds to the magnitude and sign of the Z-score where positive and negative Z-scores are colored green and red, respectively.

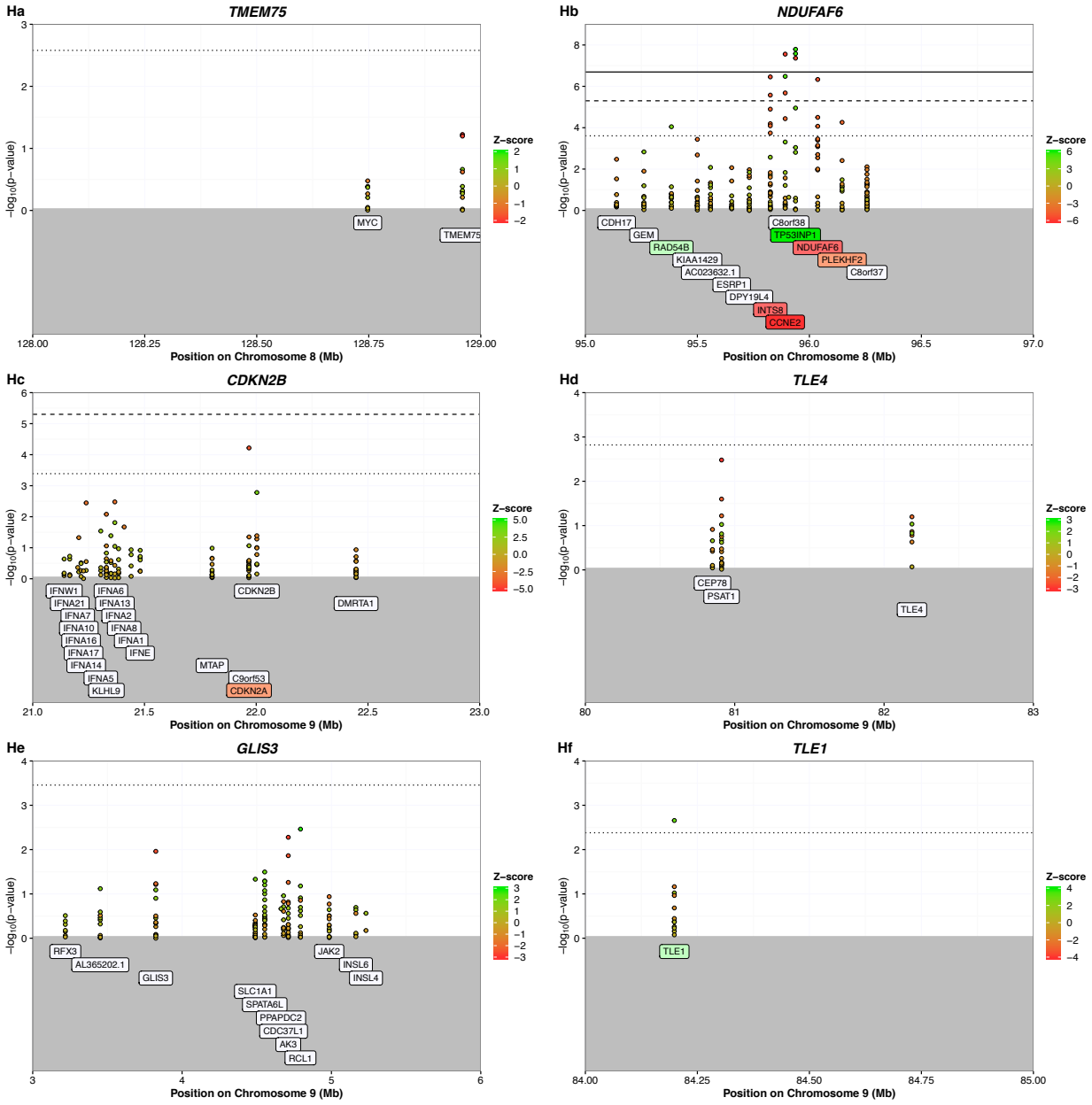

Figure 3.5 (Continued): **MetaXcan profiles at T2D-associated loci** The solid, dashed, and dotted lines denote significance correcting for the total number of tests performed across all models, genome-wide significance in a single model (10,000 tests), and locus-wide significance, respectively. Genomic position (Mb) and significance ($-\log_{10}$(p-value)) for each predicted gene expression value (from a particular tissue model) are shown on the x- and y-axes, respectively. Gene labels are shown in the gray region and are positioned at the transcription start site (TSS). Moreover, color corresponds to the magnitude and sign of the Z-score where positive and negative Z-scores are colored green and red, respectively.

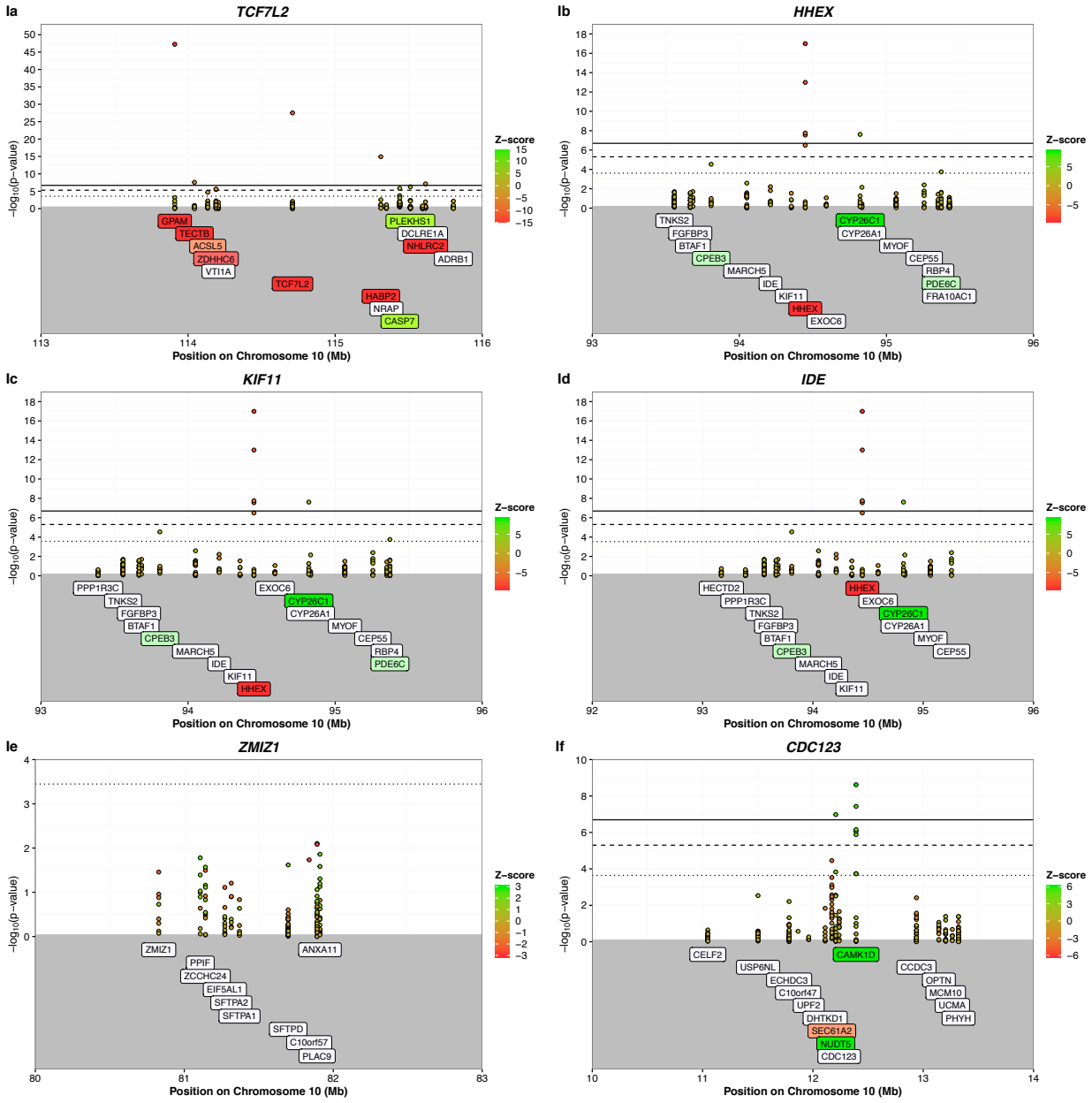

Figure 3.5 (Continued): **MetaXcan profiles at T2D-associated loci** The solid, dashed, and dotted lines denote significance correcting for the total number of tests performed across all models, genome-wide significance in a single model (10,000 tests), and locus-wide significance, respectively. Genomic position (Mb) and significance ($-\log_{10}$(p-value)) for each predicted gene expression value (from a particular tissue model) are shown on the x- and y-axes, respectively. Gene labels are shown in the gray region and are positioned at the transcription start site (TSS). Moreover, color corresponds to the magnitude and sign of the Z-score where positive and negative Z-scores are colored green and red, respectively.

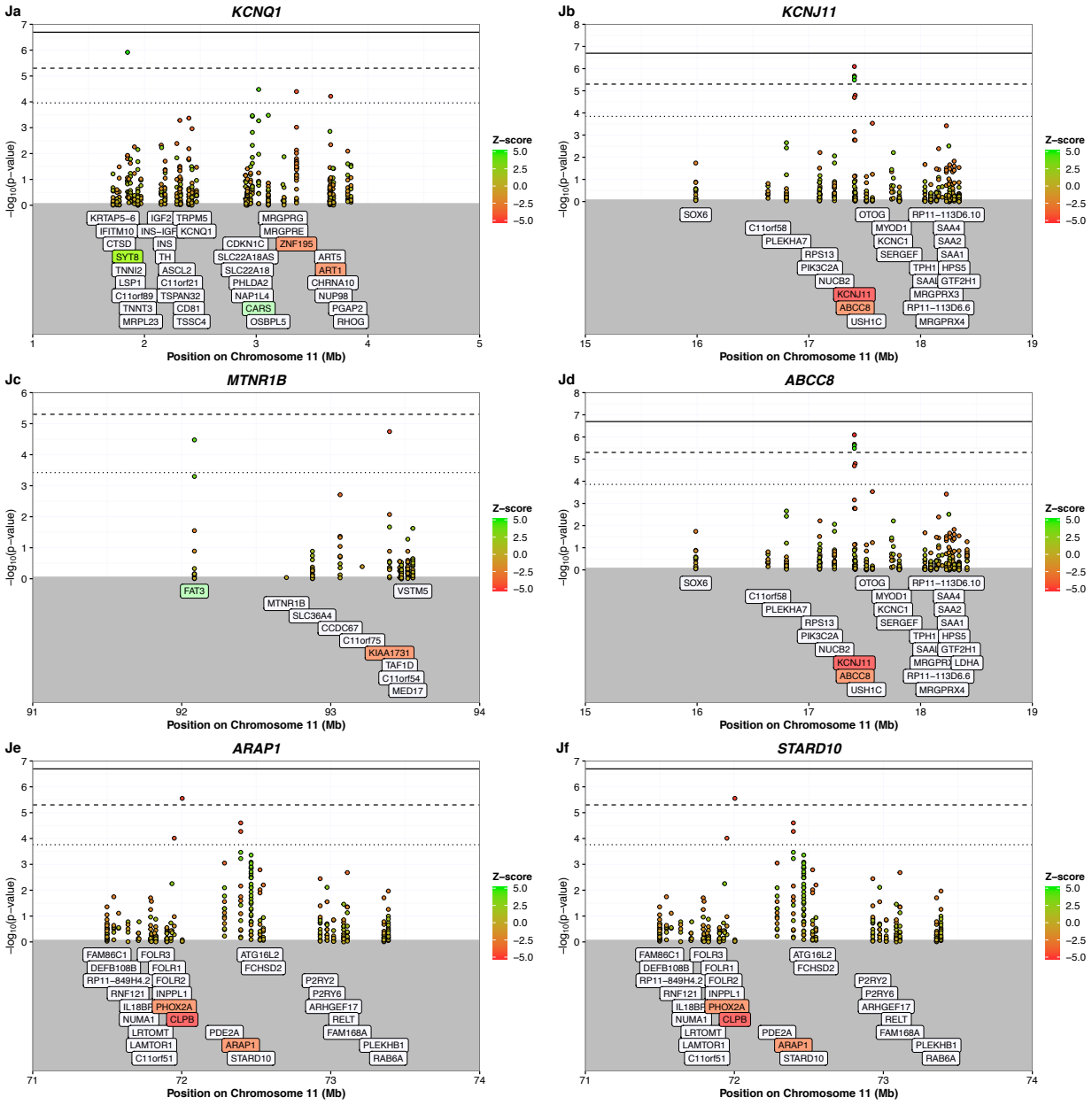

Figure 3.5 (Continued): **MetaXcan profiles at T2D-associated loci** The solid, dashed, and dotted lines denote significance correcting for the total number of tests performed across all models, genome-wide significance in a single model (10,000 tests), and locus-wide significance, respectively. Genomic position (Mb) and significance ($-\log_{10}$(p-value)) for each predicted gene expression value (from a particular tissue model) are shown on the x- and y-axes, respectively. Gene labels are shown in the gray region and are positioned at the transcription start site (TSS). Moreover, color corresponds to the magnitude and sign of the Z-score where positive and negative Z-scores are colored green and red, respectively.

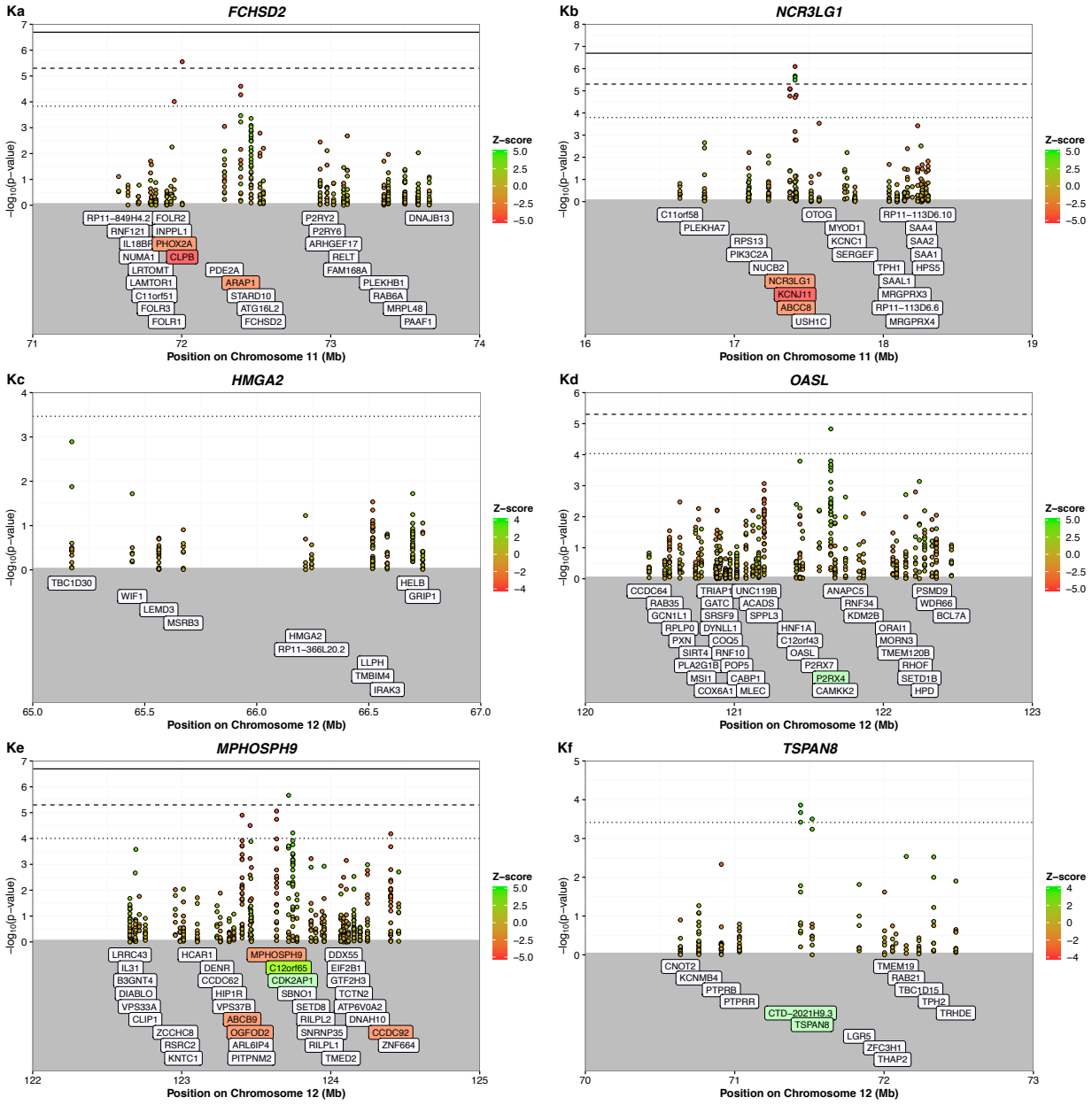

Figure 3.5 (Continued): **MetaXcan profiles at T2D-associated loci** The solid, dashed, and dotted lines denote significance correcting for the total number of tests performed across all models, genome-wide significance in a single model (10,000 tests), and locus-wide significance, respectively. Genomic position (Mb) and significance ($-\log_{10}$(p-value)) for each predicted gene expression value (from a particular tissue model) are shown on the x- and y-axes, respectively. Gene labels are shown in the gray region and are positioned at the transcription start site (TSS). Moreover, color corresponds to the magnitude and sign of the Z-score where positive and negative Z-scores are colored green and red, respectively.

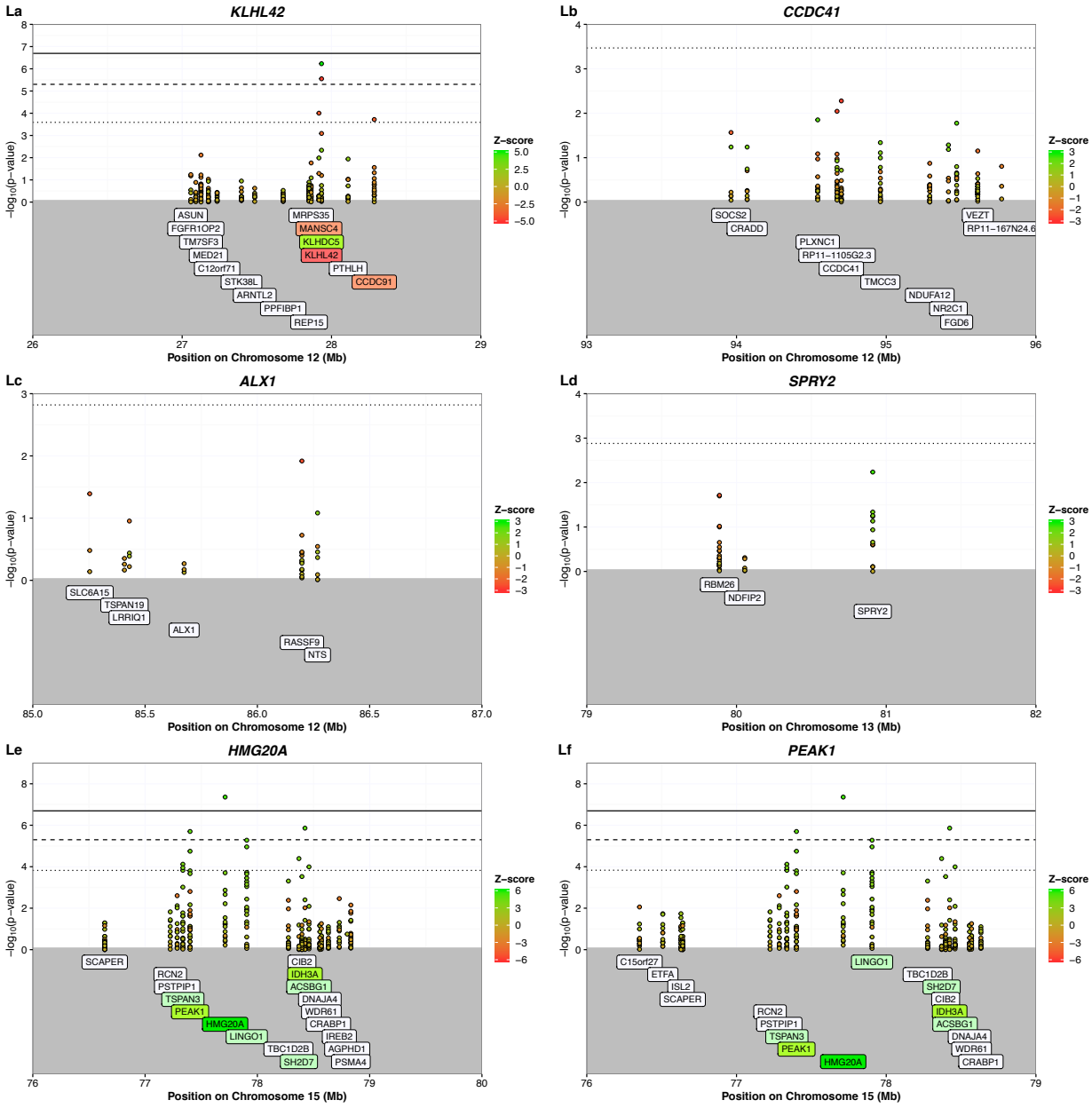

Figure 3.5 (Continued): **MetaXcan profiles at T2D-associated loci** The solid, dashed, and dotted lines denote significance correcting for the total number of tests performed across all models, genome-wide significance in a single model (10,000 tests), and locus-wide significance, respectively. Genomic position (Mb) and significance ($-\log_{10}$(p-value)) for each predicted gene expression value (from a particular tissue model) are shown on the x- and y-axes, respectively. Gene labels are shown in the gray region and are positioned at the transcription start site (TSS). Moreover, color corresponds to the magnitude and sign of the Z-score where positive and negative Z-scores are colored green and red, respectively.

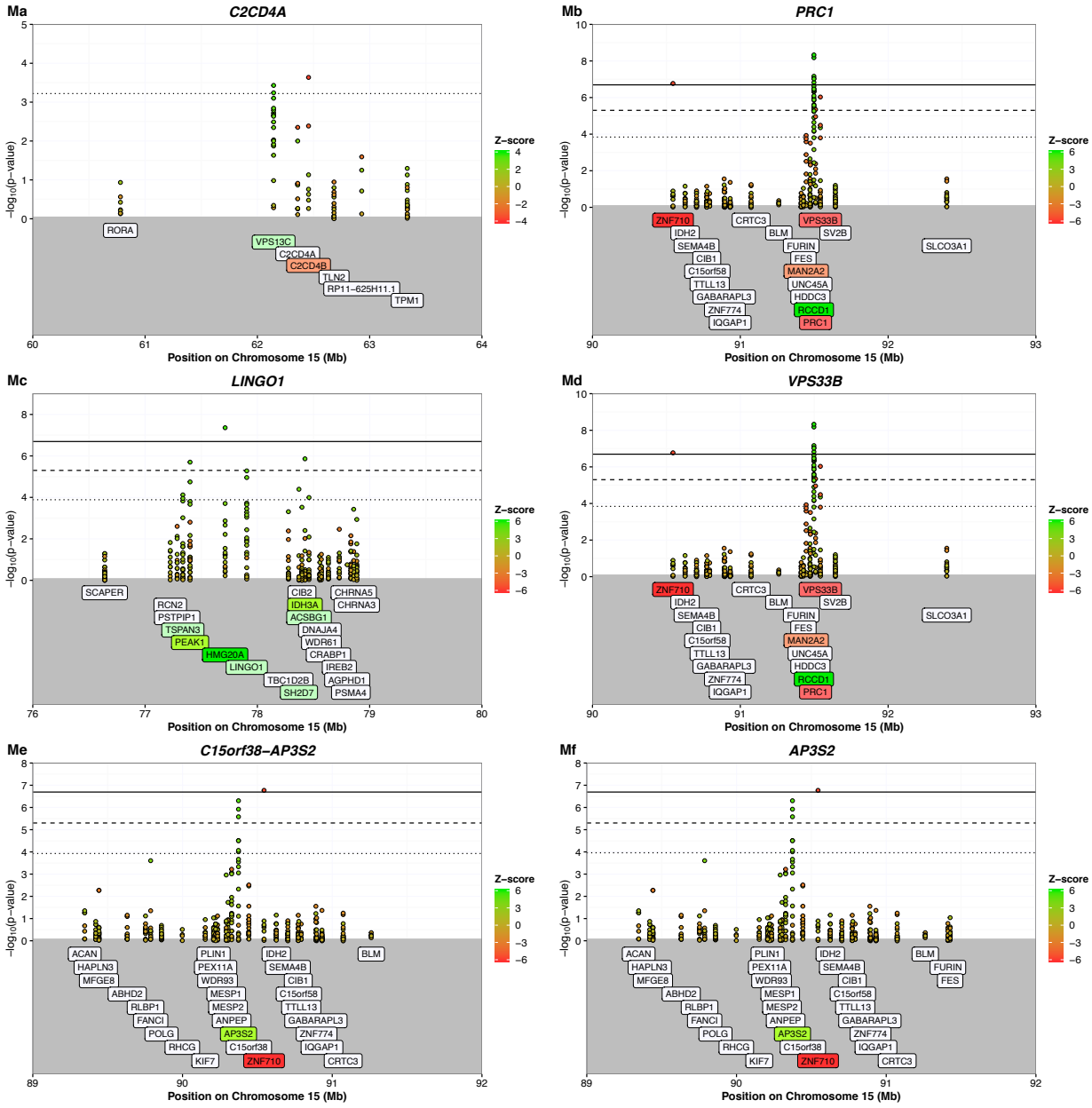

Figure 3.5 (Continued): **MetaXcan profiles at T2D-associated loci** The solid, dashed, and dotted lines denote significance correcting for the total number of tests performed across all models, genome-wide significance in a single model (10,000 tests), and locus-wide significance, respectively. Genomic position (Mb) and significance ($-\log_{10}$(p-value)) for each predicted gene expression value (from a particular tissue model) are shown on the x- and y-axes, respectively. Gene labels are shown in the gray region and are positioned at the transcription start site (TSS). Moreover, color corresponds to the magnitude and sign of the Z-score where positive and negative Z-scores are colored green and red, respectively.

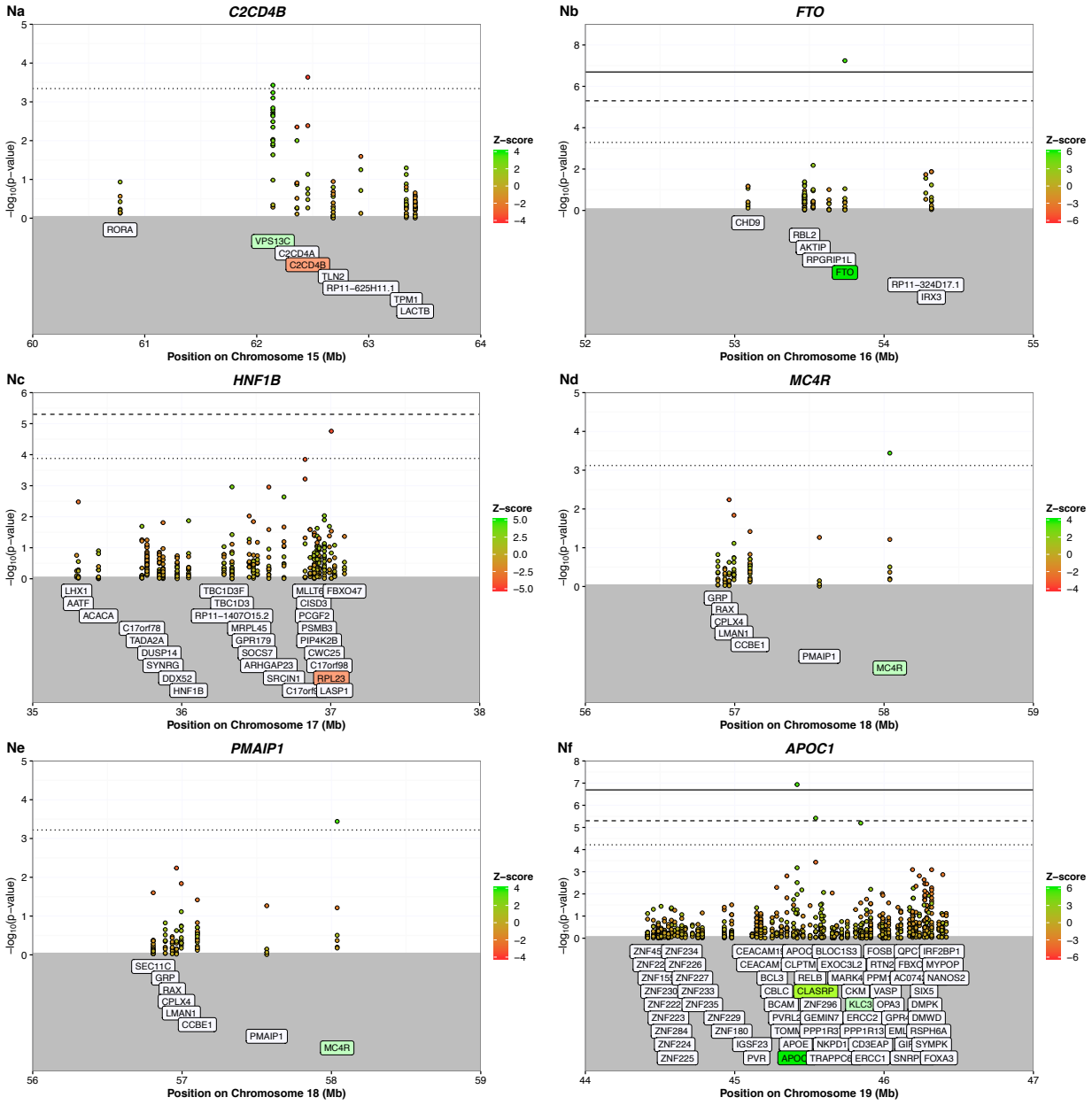

Figure 3.5 (Continued): **MetaXcan profiles at T2D-associated loci** The solid, dashed, and dotted lines denote significance correcting for the total number of tests performed across all models, genome-wide significance in a single model (10,000 tests), and locus-wide significance, respectively. Genomic position (Mb) and significance ($-\log_{10}$(p-value)) for each predicted gene expression value (from a particular tissue model) are shown on the x- and y-axes, respectively. Gene labels are shown in the gray region and are positioned at the transcription start site (TSS). Moreover, color corresponds to the magnitude and sign of the Z-score where positive and negative Z-scores are colored green and red, respectively.

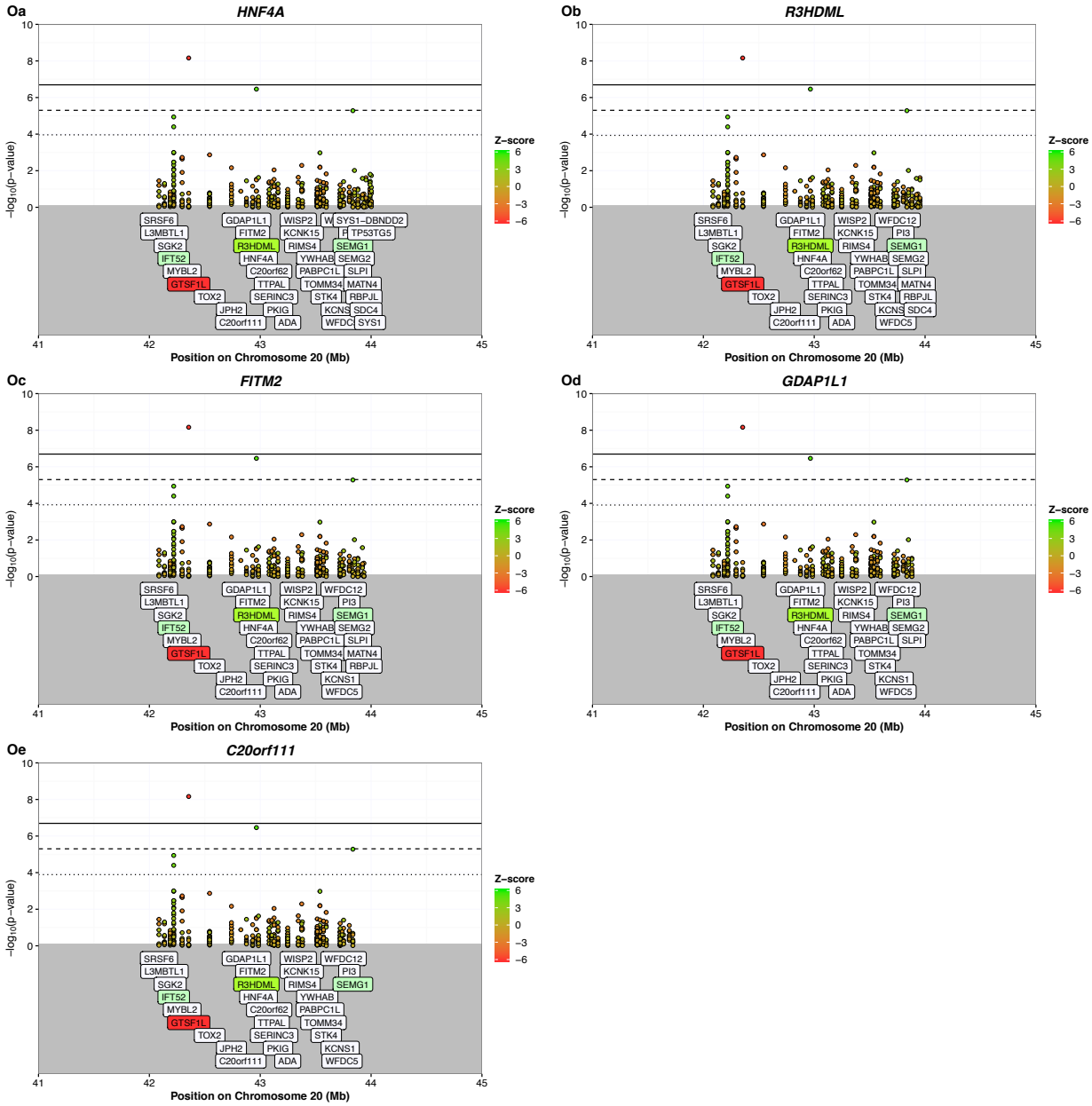

Figure 3.5 (Continued): **MetaXcan profiles at T2D-associated loci** The solid, dashed, and dotted lines denote significance correcting for the total number of tests performed across all models, genome-wide significance in a single model (10,000 tests), and locus-wide significance, respectively. Genomic position (Mb) and significance ($-\log_{10}$(p-value)) for each predicted gene expression value (from a particular tissue model) are shown on the x- and y-axes, respectively. Gene labels are shown in the gray region and are positioned at the transcription start site (TSS). Moreover, color corresponds to the magnitude and sign of the Z-score where positive and negative Z-scores are colored green and red, respectively.

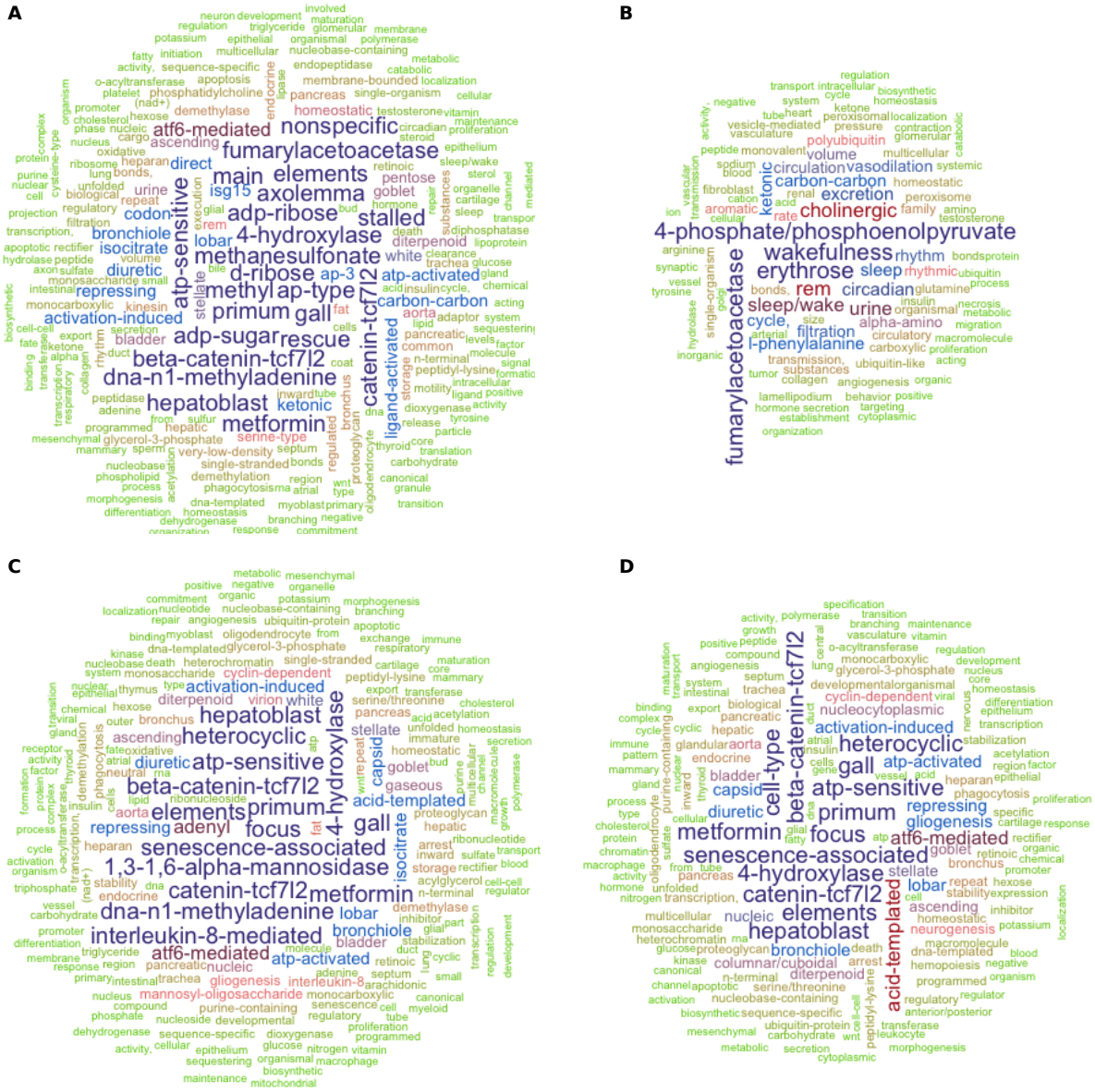

Figure 3.6: **Text-mining analysis of frequently occurring terms among enriched pathways from MetaXcan-significant gene sets.** We text-mined the set of Gene Ontology Biological Process (GO:BP) pathways that were enriched (overrepresented p-value ≤ 0.05) among sets of genes meeting genome-wide significance in at least on tissue from our MetaXcan analyses. Wordlcouds show the most frequently occurring terms (larger font and centrally located) - limited to no more than the top 200 terms. Results are shown for: (A) all genes meeting genome-wide significance in at least one tissue from the DIAGRAM analyses; (B) the subset of genes in A that are not located within 1 Mb of T2D-associated SNPs; (C) genome-wide significant genes from the meta-analysis of MetaXcan results from the DIAGRAM and GERA analyses; (D) all genes meeting genome-wide significance in at least one tissue discovered in the DIAGRAM analysis that replicate in the GERA analysis.

Chr	Gene	Reported Genes within 1Mb Locus	Model	DIAGRAM Z–score	DIAGRAM P–value	GERA Z–score	GERA P–value	Meta–analysis Z–score	Meta–analysis P–value
1	*CDKN2C*	*FAF1*	Esophagus GastroesophagealJunction	4.3	1.5e–05	3.4	7e–04	5.1	4.1e–07
1	*RP11–268J15.5*	none reported	AdrenalGland	4.3	2e–05	2.2	0.025	4.6	4.7e–06
1	*RP11–268J15.5*	none reported	Lung	4.1	4.3e–05	2.4	0.016	4.3	1.6e–05
1	*RP11–268J15.5*	none reported	Stomach	4.3	2.1e–05	2.6	0.0086	4.6	3.8e–06
1	*GDAP2*	none reported	Brain Cerebellum	–4.2	3e–05	–1.3	0.19	–3.8	0.00016
1	*SH3D21*	none reported	Heart AtrialAppendage	–4.4	1.1e–05	–2.5	0.013	–4.7	3.2e–06
1	*UTS2*	none reported	WholeBloodDGN	5	6.3e–07	–0.49	0.62	3.6	0.00038
2	*NRBP1*	none reported	WholeBlood	–4.5	7.3e–06	–4.2	2.9e–05	–5.5	4.2e–08
3	*CPNE4*	*CPNE4*	CrossTissue	4.8	1.4e–06	1.5	0.14	4.3	1.5e–05
3	*CPNE4*	*CPNE4*	Adipose Subcutaneous	4.8	1.8e–06	1.5	0.13	4.4	1.3e–05
3	*PPARG*	*PPARG,TIMP4,TAMM41,*	Cells EBV transformedlymphocytes	–5	5.2e–07	–3.7	0.00026	–5.7	1.1e–08
3	*PPARG*	*PPARG,TIMP4,TAMM41,*	Heart LeftVentricle	–5.2	1.8e–07	0.14	0.89	–4	6.7e–05
3	*UBE2E1*	*UBE2E2*	Artery Coronary	4.6	3.5e–06	1.1	0.26	4.2	3.2e–05
3	*CRYGS*	*IGF2BP2,ST6GAL1*	Artery Tibial	4.3	2e–05	NA	NA	NA	NA
3	*TRA2B*	*IGF2BP2*	Brain Anteriorcingulatecortex BA24	–4.2	2.9e–05	–1.4	0.15	–4	6.7e–05
3	*DGKG*	*IGF2BP2,ST6GAL1*	Ovary	–7.6	2.3e–14	–6.3	2.8e–10	–8.8	1.3e–18
3	*NR1D2*	*UBE2E2*	Skin SunExposed Lowerleg	4.8	1.5e–06	1.8	0.072	4.5	8.6e–06
3	*ZNF148*	none reported	Pituitary	–4.2	2.8e–05	–0.65	0.52	–3.5	0.00043
3	*ATP2B2*	none reported	Skin SunExposed Lowerleg	4.3	2e–05	1.1	0.28	3.7	0.00022
3	*MUSTN1*	none reported	WholeBloodDGN	4.2	3e–05	0.97	0.33	3.7	0.00026
3	*SFMBT1*	none reported	CrossTissue	4.1	4.5e–05	1.4	0.16	3.7	0.00018
4	*WFS1*	*WFS1,PPP2R2C*	CrossTissue	5.3	1.1e–07	4.8	1.5e–06	6.3	2.7e–10
4	*WFS1*	*WFS1,PPP2R2C*	Brain CerebellarHemisphere	4.3	1.4e–05	4.7	2.5e–06	5.7	1.4e–08
4	*WFS1*	*WFS1,PPP2R2C*	Brain Cerebellum	5.3	9.6e–08	4	5.8e–05	6	2.3e–09
4	*WFS1*	*WFS1,PPP2R2C*	Nerve Tibial	4.9	9.4e–07	4.6	4.3e–06	5.9	3e–09
4	*WFS1*	*WFS1,PPP2R2C*	Skin NotSunExposed Suprapubic	5.8	5.7e–09	4.6	4.7e–06	6.7	1.7e–11
4	*WFS1*	*WFS1,PPP2R2C*	Skin SunExposed Lowerleg	5.4	5.7e–08	4.7	2.9e–06	6.3	3e–10
4	*WFS1*	*WFS1,PPP2R2C*	Thyroid	5.4	7.9e–08	4.6	3.9e–06	6.5	8e–11
4	*TIGD4*	*TMEM154,FBXW7*	Brain CerebellarHemisphere	4.1	4.3e–05	–0.66	0.51	2.8	0.0048
5	*PCBD2*	*PCBD2*	Lung	–4.1	4.1e–05	–1.3	0.19	–3.8	0.00015

Table 3.2: **MetaXcan associations with T2D.** Results for genes and corresponding models that meet genome-wide significance *in at least one model* from the DIAGRAM analysis are shown with nearby genes and results from the GERA replication study and meta-analysis of DIAGRAM and GERA Metaxcan associations. Blue shading denotes genes not implicated by the top 1,000 SNPs from the DIAGRAM trans-ethnic meta-analysis of GWAS. Pink and red shading denote genome-wide significance in one model and across all models, respectively, for the DIAGRAM and meta-analysis. Replication in the GERA study is denoted by a pink outline.

Chr	Gene	Reported Genes within 1Mb Locus	Model	DIAGRAM Z–score	DIAGRAM P–value	GERA Z–score	GERA P–value	Meta–analysis Z–score	Meta–analysis P–value
5	*PCBD2*	*PCBD2*	Pancreas	–4.1	4.5e–05	NA	NA	NA	NA
5	*PCBD2*	*PCBD2*	Skin SunExposed Lowerleg	–4	6.5e–05	1.4	0.15	–2.3	0.022
6	*E2F3*	*CDKAL1*	Brain Cerebellum	–4.6	3.9e–06	–2.2	0.03	–4.5	5.8e–06
6	*HCG27*	*CCHCR1,POU5F1,HLA–C, TCF19*	Heart LeftVentricle	4.5	5.6e–06	1.2	0.23	4.1	3.5e–05
6	*SOX4*	*CDKAL1*	Muscle Skeletal	–7.9	3.3e–15	–4.9	8.5e–07	–8.4	3.7e–17
6	*CYP21A2*	*CCHCR1,POU5F1,HLA–C, TCF19*	Nerve Tibial	4.2	2.2e–05	2.4	0.018	4.3	1.4e–05
6	*AGER*	*HLA–C*	CrossTissue	4.2	3.2e–05	1.6	0.12	3.9	0.00011
6	*KCNK17*	none reported	Pancreas	3.9	8.1e–05	0.13	0.9	3	0.0024
6	*HLA–A*	none reported	Pituitary	6.1	1.2e–09	–0.77	0.44	4.3	2e–05
6	*ZNRD1*	none reported	WholeBloodDGN	–5.5	3.6e–08	0.3	0.77	–4	5.2e–05
7	*JAZF1*	*JAZF1*	Muscle Skeletal	–6.2	4.4e–10	–2.5	0.014	–6	2.5e–09
7	*JAZF1*	*JAZF1*	Pancreas	–5.6	2e–08	–2.9	0.0044	–5.6	1.8e–08
7	*HOXA5*	*JAZF1*	WholeBlood	–5.7	1.2e–08	–6.1	1e–09	–7.4	1.8e–13
7	*NACAD*	none reported	Brain Anteriorcingulatecortex BA24	4.2	2.9e–05	–0.84	0.4	2.8	0.0044
7	*ZNF713*	none reported	Brain Cerebellum	4	5.4e–05	–0.57	0.57	2.8	0.0057
7	*NOM1*	none reported	Muscle Skeletal	4	5.4e–05	1	0.31	3.6	0.00035
7	*MAFK*	none reported	WholeBloodDGN	4.3	1.4e–05	–0.25	0.8	3.2	0.0014
8	*TP53INP1*	*TP53INP1,DPY19L4,INTS8, NDUFAF6*	Artery Aorta	4.4	1.1e–05	1.8	0.073	4.3	1.4e–05
8	*TP53INP1*	*TP53INP1,DPY19L4,INTS8, NDUFAF6*	Esophagus Muscularis	5.6	1.6e–08	0.82	0.41	4.7	2.1e–06
8	*TP53INP1*	*TP53INP1,DPY19L4,INTS8, NDUFAF6*	Testis	–5.5	4.3e–08	–0.51	0.61	–4.5	7.8e–06
8	*TP53INP1*	*TP53INP1,DPY19L4,INTS8, NDUFAF6*	WholeBloodDGN	5.6	2.6e–08	1.3	0.21	4.9	1.2e–06
8	*INTS8*	*TP53INP1,DPY19L4,INTS8, NDUFAF6*	Artery Tibial	–4.7	2.7e–06	NA	NA	NA	NA
8	*INTS8*	*TP53INP1,DPY19L4,INTS8, NDUFAF6*	Lung	–4.4	1.3e–05	–1.1	0.25	–3.9	9.3e–05
8	*INTS8*	*TP53INP1,DPY19L4,INTS8, NDUFAF6*	WholeBloodDGN	–5.1	3.5e–07	–1.7	0.083	–4.7	2.2e–06
8	*NDUFAF6*	*NDUFAF6*	Brain CerebellarHemisphere	–4.2	3.2e–05	–0.78	0.44	–3.6	0.00034
8	*NDUFAF6*	*NDUFAF6*	Brain Cerebellum	–5	4.7e–07	1	0.32	–3.3	0.00093
8	*ANK1*	*ANK1*	CrossTissue	5.1	2.8e–07	–0.8	0.42	3.5	0.00052
8	*ANK1*	*ANK1*	Skin SunExposed Lowerleg	4.2	2.3e–05	–0.19	0.85	3.1	0.0021
8	*VDAC3*	*ANK1*	AdrenalGland	–5	4.6e–07	–1.9	0.057	–5	5.1e–07
8	*CCNE2*	*TP53INP1,DPY19L4,INTS8, NDUFAF6*	Cells EBV transformedlymphocytes	5.1	3.3e–07	–0.71	0.48	3.6	0.00032

Table 3.2 (Continued): **MetaXcan associations with T2D.** Results for genes and corresponding models that meet genome-wide significance *in at least one model* from the DIAGRAM analysis are shown with nearby genes and results from the GERA replication study and meta-analysis of DIAGRAM and GERA Metaxcan associations. Blue shading denotes genes not implicated by the top 1,000 SNPs from the DIAGRAM trans-ethnic meta-analysis of GWAS. Pink and red shading denote genome-wide significance in one model and across all models, respectively, for the DIAGRAM and meta-analysis. Replication in the GERA study is denoted by a pink outline.

Chr	Gene	Reported Genes within 1Mb Locus	Model	DIAGRAM Z–score	DIAGRAM P–value	GERA Z–score	GERA P–value	Meta–analysis Z–score	Meta–analysis P–value
8	*CCNE2*	*TP53INP1,DPY19L4,INTS8, NDUFAF6*	Heart LeftVentricle	−4.1	3.7e−05	−1.3	0.21	−3.8	0.00012
8	*CCNE2*	*TP53INP1,DPY19L4,INTS8, NDUFAF6*	Lung	−4.7	2.1e−06	−0.1	0.92	−3.7	0.00023
8	*CCNE2*	*TP53INP1,DPY19L4,INTS8, NDUFAF6*	WholeBloodDGN	−5.6	2.8e−08	−0.88	0.38	−4.7	3.1e−06
8	*PLEKHF2*	*TP53INP1,DPY19L4,INTS8, NDUFAF6*	Heart LeftVentricle	−4	5.5e−05	−0.65	0.52	−3.5	0.00054
8	*NKX6−3*	*ANK1*	Pancreas	−4	7.3e−05	−2	0.041	−4	6.6e−05
9	*CDKN2A*	*CDKN2B*	Brain CerebellarHemisphere	−4	6.1e−05	−4.5	5.8e−06	−5.3	9.9e−08
10	*TCF7L2*	*TCF7L2*	Artery Aorta	−11	3e−28	−9.2	2.8e−20	−13	3.3e−40
10	*HHEX*	*HHEX,KIF11,IDE,*	CrossTissue	−5.6	1.8e−08	−0.69	0.49	−4.6	5e−06
10	*HHEX*	*HHEX,KIF11,IDE,*	Breast MammaryTissue	−8.6	1e−17	−3.6	0.00034	−8.4	5.2e−17
10	*HHEX*	*HHEX,KIF11,IDE,*	Cells Transformedfibroblasts	−5.1	3.2e−07	−3.9	1e−04	−5.9	4.2e−09
10	*HHEX*	*HHEX,KIF11,IDE,*	Testis	−5.6	2.8e−08	−2.9	0.0034	−5.7	1e−08
10	*HHEX*	*HHEX,KIF11,IDE,*	WholeBloodDGN	−7.4	1e−13	−3.2	0.0015	−7.2	5.1e−13
10	*NHLRC2*	*TCF7L2*	AdrenalGland	−5.4	7.7e−08	−3.2	0.0013	−6	2.5e−09
10	*CASP7*	*TCF7L2*	Artery Aorta	4.8	1.4e−06	0.95	0.34	4.2	2.2e−05
10	*GPAM*	*TCF7L2*	Artery Tibial	−15	5.7e−48	−10	1.5e−25	−16	2.1e−60
10	*ZDHHC6*	*TCF7L2*	Brain Anteriorcingulatecortex BA24	−4.6	3.4e−06	0.12	0.9	−3.6	0.00036
10	*ZDHHC6*	*TCF7L2*	Muscle Skeletal	−4.7	2.3e−06	0.79	0.43	−3.2	0.0013
10	*HABP2*	*TCF7L2*	Brain CerebellarHemisphere	−8	1.2e−15	−13	2e−39	−13	9.6e−37
10	*TECTB*	*TCF7L2*	Brain Putamen basalganglia	−5.6	2.7e−08	−5.2	2.2e−07	−6.9	5.6e−12
10	*PLEKHS1*	*TCF7L2*	Esophagus Mucosa	5	5e−07	6.4	1.1e−10	7.2	4.8e−13
10	*CAMK1D*	*CDC123*	Heart LeftVentricle	4.9	9e−07	3.8	0.00013	5.7	9.6e−09
10	*CAMK1D*	*CDC123*	Lung	6	2.4e−09	1.2	0.25	5.1	2.8e−07
10	*CAMK1D*	*CDC123*	Spleen	5.5	3.7e−08	2.4	0.015	5.4	5.7e−08
10	*CAMK1D*	*CDC123*	WholeBlood	4.8	1.3e−06	1.6	0.12	4.5	7.8e−06
10	*CAMK1D*	*CDC123*	WholeBloodDGN	5	7e−07	2.7	0.0072	5.1	3.5e−07
10	*CPEB3*	*HHEX,KIF11,IDE,*	Heart LeftVentricle	4.2	2.9e−05	1.1	0.28	3.8	0.00015
10	*CYP26C1*	*HHEX,KIF11,IDE,*	Pancreas	5.6	2.4e−08	3.8	0.00013	6.1	1.2e−09
10	*SEC61A2*	*CDC123*	Pituitary	−4.1	3.5e−05	−1.1	0.26	−3.7	2e−04
10	*NUDT5*	*CDC123*	WholeBloodDGN	5.3	1e−07	1.5	0.14	4.8	1.8e−06
11	*ABCC8*	*KCNJ11,ABCC8,NCR3LG1,*	Brain Cortex	−4.3	1.6e−05	−1.6	0.1	−4.1	3.5e−05

Table 3.2 (Continued): **MetaXcan associations with T2D.** Results for genes and corresponding models that meet genome-wide significance *in at least one model* from the DIAGRAM analysis are shown with nearby genes and results from the GERA replication study and meta-analysis of DIAGRAM and GERA Metaxcan associations. Blue shading denotes genes not implicated by the top 1,000 SNPs from the DIAGRAM trans-ethnic meta-analysis of GWAS. Pink and red shading denote genome-wide significance in one model and across all models, respectively, for the DIAGRAM and meta-analysis. Replication in the GERA study is denoted by a pink outline.

Chr	Gene	Reported Genes within 1Mb Locus	Model	DIAGRAM Z–score	DIAGRAM P–value	GERA Z–score	GERA P–value	Meta–analysis Z–score	Meta–analysis P–value
11	*KCNJ11*	*KCNJ11,ABCC8,NCR3LG1,*	Esophagus Mucosa	−4.7	2.2e−06	−1.2	0.22	−4.4	1.4e−05
11	*KCNJ11*	*KCNJ11,ABCC8,NCR3LG1,*	Skin NotSunExposed Suprapubic	−4.3	2e−05	−0.58	0.56	−3.6	0.00037
11	*KCNJ11*	*KCNJ11,ABCC8,NCR3LG1,*	Skin SunExposed Lowerleg	−4.9	8e−07	−2.7	0.0073	−5	6.6e−07
11	*KCNJ11*	*KCNJ11,ABCC8,NCR3LG1,*	Testis	4.7	2.4e−06	2	0.044	4.6	3.6e−06
11	*KCNJ11*	*KCNJ11,ABCC8,NCR3LG1,*	WholeBloodDGN	4.7	3.3e−06	2.1	0.037	4.6	4.9e−06
11	*NCR3LG1*	*NCR3LG1*	CrossTissue	−4.3	1.7e−05	−1.9	0.062	−4.1	3.7e−05
11	*NCR3LG1*	*NCR3LG1*	Esophagus Mucosa	−4.4	8.6e−06	−2.2	0.029	−4.6	4e−06
11	*NCR3LG1*	*NCR3LG1*	Nerve Tibial	−4.5	8.1e−06	−2.1	0.034	−4.4	1.1e−05
11	*ARAP1*	*ARAP1,STARD10,FCHSD2,*	Esophagus Mucosa	−4	5.3e−05	−2	0.05	−4.2	3e−05
11	*ARAP1*	*ARAP1,STARD10,FCHSD2,*	Skin NotSunExposed Suprapubic	−4.2	2.5e−05	−1.5	0.12	−4	6.5e−05
11	*CLPB*	*ARAP1,STARD10,FCHSD2,*	Artery Coronary	−4.7	2.8e−06	−3.8	0.00017	−5.5	3.5e−08
11	*SYT8*	*KCNQ1*	Esophagus GastroesophagealJunction	4.9	1.2e−06	1.1	0.27	4.3	1.5e−05
11	*PHOX2A*	*ARAP1,STARD10,FCHSD2,*	Pancreas	−3.9	9.7e−05	−1.7	0.087	−3.8	0.00016
11	*ZNF195*	*KCNQ1*	Testis	−4.1	4e−05	0.39	0.7	−3	0.003
11	*CTD−2210P24.4*	none reported	Brain Nucleusaccumbens basalganglia	4.3	1.6e−05	1.5	0.13	4.2	3.1e−05
11	*TCP11L1*	none reported	Muscle Skeletal	4.2	2.3e−05	−0.39	0.7	3	0.0024
11	*TCP11L1*	none reported	Pancreas	3.9	0.00011	−0.33	0.74	2.8	0.0058
11	*NFRKB*	none reported	Pancreas	−4	7.3e−05	−1.5	0.15	−3.7	0.00021
12	*KLHL42*	*KLHL42*	Artery Tibial	−4.7	2.8e−06	NA	NA	NA	NA
12	*MPHOSPH9*	*MPHOSPH9*	CrossTissue	−4.3	1.8e−05	NA	NA	NA	NA
12	*MPHOSPH9*	*MPHOSPH9*	Heart AtrialAppendage	−4.4	8.7e−06	NA	NA	NA	NA
12	*C12orf65*	*MPHOSPH9*	Nerve Tibial	4.7	2.1e−06	NA	NA	NA	NA
12	*CDK2AP1*	*MPHOSPH9*	Skin SunExposed Lowerleg	4	6.1e−05	−0.25	0.8	2.9	0.0041
12	*P2RX4*	*OASL*	WholeBlood	4.3	1.5e−05	4.3	2.1e−05	5.4	6.6e−08
12	*ABCB9*	*MPHOSPH9*	WholeBloodDGN	−4.4	1.3e−05	0.6	0.55	−3	0.0024
12	*OGFOD2*	*MPHOSPH9*	WholeBloodDGN	−4.2	3.1e−05	NA	NA	NA	NA
12	*KLHDC5*	*KLHL42*	WholeBloodDGN	5	5.9e−07	−2.1	0.036	2.8	0.0055
14	*SAMD15*	none reported	Artery Coronary	−4.1	3.8e−05	1.1	0.28	−2.7	0.008
14	*VIPAS39*	none reported	Breast MammaryTissue	−4.3	2e−05	−0.15	0.88	−3.4	0.00078
14	*VIPAS39*	none reported	Nerve Tibial	−4.3	1.8e−05	0.15	0.88	−3.2	0.0016

Table 3.2 (Continued): **MetaXcan associations with T2D.** Results for genes and corresponding models that meet genome-wide significance *in at least one model* from the DIAGRAM analysis are shown with nearby genes and results from the GERA replication study and meta-analysis of DIAGRAM and GERA Metaxcan associations. Blue shading denotes genes not implicated by the top $1,000$ SNPs from the DIAGRAM trans-ethnic meta-analysis of GWAS. Pink and red shading denote genome-wide significance in one model and across all models, respectively, for the DIAGRAM and meta-analysis. Replication in the GERA study is denoted by a pink outline.

Chr	Gene	Reported Genes within 1Mb Locus	Model	DIAGRAM Z–score	DIAGRAM P–value	GERA Z–score	GERA P–value	Meta–analysis Z–score	Meta–analysis P–value
14	*VIPAS39*	none reported	WholeBlood	–4.4	1.3e–05	–0.35	0.73	–3.5	0.00047
14	*NOXRED1*	none reported	CrossTissue	4.1	4.4e–05	0.42	0.67	3.3	0.0011
15	*AP3S2*	*C15orf38–AP3S2,AP3S2*	CrossTissue	4.7	2.6e–06	3.2	0.0013	5.1	3.8e–07
15	*AP3S2*	*C15orf38–AP3S2,AP3S2*	AdrenalGland	5	5e–07	3.5	0.00041	5.9	4.8e–09
15	*AP3S2*	*C15orf38–AP3S2,AP3S2*	Artery Aorta	4.2	3.1e–05	3.1	0.0019	4.8	1.4e–06
15	*AP3S2*	*C15orf38–AP3S2,AP3S2*	Cells Transformedfibroblasts	4.2	3.1e–05	3.5	0.00039	5	6.4e–07
15	*AP3S2*	*C15orf38–AP3S2,AP3S2*	Stomach	4.9	1.2e–06	3.5	0.00045	5.5	3.2e–08
15	*LINGO1*	*HMG20A,PEAK1,LINGO1,*	Brain Cortex	4.6	5.3e–06	–0.25	0.81	3.4	0.00072
15	*LINGO1*	*HMG20A,PEAK1,LINGO1,*	Heart LeftVentricle	4.4	1.1e–05	2.4	0.016	4.6	3.7e–06
15	*PEAK1*	*HMG20A,PEAK1,LINGO1,*	Esophagus Mucosa	4.3	1.8e–05	3.2	0.0014	5	5.8e–07
15	*PEAK1*	*HMG20A,PEAK1,LINGO1,*	Spleen	4.8	2e–06	3.2	0.0016	5.2	1.9e–07
15	*PRC1*	*PRC1,VPS33B*	Pancreas	–4.6	4.5e–06	–3.3	0.00089	–5.1	3.7e–07
15	*PRC1*	*PRC1,VPS33B*	WholeBloodDGN	–4.4	1.1e–05	–2.2	0.03	–4.4	1e–05
15	*HMG20A*	*HMG20A,PEAK1,LINGO1,*	Pituitary	5.5	4.4e–08	3.4	0.00065	5.9	4.5e–09
15	*VPS33B*	*PRC1,VPS33B*	Pituitary	4.1	4.4e–05	1.7	0.096	3.9	8.1e–05
15	*VPS33B*	*PRC1,VPS33B*	Testis	–4.9	9.4e–07	–1.1	0.28	–4.3	1.6e–05
15	*VPS33B*	*PRC1,VPS33B*	WholeBloodDGN	–4.2	3.3e–05	–1.7	0.093	–4	6.7e–05
15	*RCCD1*	*PRC1,VPS33B*	CrossTissue	4.5	5.4e–06	2.7	0.0071	4.7	2.4e–06
15	*RCCD1*	*PRC1,VPS33B*	Adipose Subcutaneous	4.6	4.9e–06	2.9	0.0033	4.9	9e–07
15	*RCCD1*	*PRC1,VPS33B*	Artery Aorta	5.3	1.5e–07	3	0.0023	5.6	1.7e–08
15	*RCCD1*	*PRC1,VPS33B*	Artery Coronary	4.9	1.1e–06	2.5	0.012	5	4.7e–07
15	*RCCD1*	*PRC1,VPS33B*	Artery Tibial	5.1	3.3e–07	3	0.0028	5.4	6.1e–08
15	*RCCD1*	*PRC1,VPS33B*	Brain Anteriorcingulatecortex BA24	5.8	6.6e–09	3.4	0.00067	6.2	4.5e–10
15	*RCCD1*	*PRC1,VPS33B*	Brain Caudate basalganglia	5.2	2.3e–07	2.8	0.0048	5.4	8.6e–08
15	*RCCD1*	*PRC1,VPS33B*	Brain CerebellarHemisphere	4.7	2.5e–06	3.1	0.0017	5.2	2.3e–07
15	*RCCD1*	*PRC1,VPS33B*	Brain Cerebellum	4.6	3.7e–06	3.1	0.0018	5	5.6e–07
15	*RCCD1*	*PRC1,VPS33B*	Brain Cortex	4.6	3.7e–06	3.1	0.0018	5.1	3.3e–07
15	*RCCD1*	*PRC1,VPS33B*	Brain FrontalCortex BA9	5	4.9e–07	3.1	0.002	5.4	6.3e–08
15	*RCCD1*	*PRC1,VPS33B*	Brain Nucleusaccumbens basalganglia	4.4	1.3e–05	3.1	0.002	5	5.5e–07
15	*RCCD1*	*PRC1,VPS33B*	Breast MammaryTissue	4.8	1.4e–06	3	0.003	5.2	2.1e–07

Table 3.2 (Continued): **MetaXcan associations with T2D.** Results for genes and corresponding models that meet genome-wide significance *in at least one model* from the DIAGRAM analysis are shown with nearby genes and results from the GERA replication study and meta-analysis of DIAGRAM and GERA Metaxcan associations. Blue shading denotes genes not implicated by the top $1,000$ SNPs from the DIAGRAM trans-ethnic meta-analysis of GWAS. Pink and red shading denote genome-wide significance in one model and across all models, respectively, for the DIAGRAM and meta-analysis. Replication in the GERA study is denoted by a pink outline.

Chr	Gene	Reported Genes within 1Mb Locus	Model	DIAGRAM Z–score	DIAGRAM P–value	GERA Z–score	GERA P–value	Meta–analysis Z–score	Meta–analysis P–value
15	*RCCD1*	*PRC1,VPS33B*	Cells Transformedfibroblasts	5.4	6.9e–08	3.5	0.00053	5.9	3.9e–09
15	*RCCD1*	*PRC1,VPS33B*	Colon Sigmoid	4.8	1.3e–06	3.1	0.0019	5.5	4.2e–08
15	*RCCD1*	*PRC1,VPS33B*	Esophagus GastroesophagealJunction	4.9	8e–07	2.8	0.0044	5.3	1.4e–07
15	*RCCD1*	*PRC1,VPS33B*	Esophagus Mucosa	4.2	2.3e–05	3.1	0.002	4.9	9.4e–07
15	*RCCD1*	*PRC1,VPS33B*	Heart LeftVentricle	5.9	4.7e–09	2.8	0.0056	5.9	2.8e–09
15	*RCCD1*	*PRC1,VPS33B*	Muscle Skeletal	5	4.6e–07	3.1	0.002	5.4	8.3e–08
15	*RCCD1*	*PRC1,VPS33B*	Nerve Tibial	5.3	9.1e–08	3.2	0.0013	5.6	2.3e–08
15	*RCCD1*	*PRC1,VPS33B*	Pancreas	5	4.5e–07	3	0.003	5.3	1.5e–07
15	*RCCD1*	*PRC1,VPS33B*	Skin NotSunExposed Suprapubic	4.7	3e–06	3.2	0.0016	5.1	2.7e–07
15	*RCCD1*	*PRC1,VPS33B*	Skin SunExposed Lowerleg	4.7	2.7e–06	3	0.0027	4.9	7.6e–07
15	*RCCD1*	*PRC1,VPS33B*	SmallIntestine TerminalIleum	4.5	5.9e–06	3.1	0.0021	5.3	1.3e–07
15	*RCCD1*	*PRC1,VPS33B*	Stomach	5.1	3.9e–07	3.1	0.0022	5.5	4.4e–08
15	*RCCD1*	*PRC1,VPS33B*	Testis	5.3	9.5e–08	3.2	0.0016	5.7	1.4e–08
15	*RCCD1*	*PRC1,VPS33B*	Thyroid	5.1	3.4e–07	3	0.0027	5.5	4.4e–08
15	*RCCD1*	*PRC1,VPS33B*	WholeBloodDGN	4.1	4e–05	3.1	0.0022	4.6	3.7e–06
15	*ZNF710*	*PRC1,VPS33B,C15orf38–AP3S2, AP3S2*	Brain Caudate basalganglia	–5.2	1.7e–07	NA	NA	NA	NA
15	*IDH3A*	*HMG20A,PEAK1,LINGO1,*	Brain Hippocampus	4.8	1.4e–06	1.6	0.11	4.6	4.3e–06
15	*SH2D7*	*HMG20A,PEAK1,LINGO1,*	Pancreas	4.1	4.1e–05	0.81	0.42	3.5	0.00048
15	*ZFAND6*	none reported	Brain CerebellarHemisphere	4.5	7.1e–06	–1.4	0.16	2.8	0.006
15	*FAH*	none reported	Brain Cerebellum	4.5	6.3e–06	–1.2	0.21	2.8	0.0051
16	*FTO*	*FTO*	Muscle Skeletal	5.4	5.8e–08	1.7	0.085	5	6.3e–07
17	*RPL23*	*HNF1B*	Brain Cortex	–4.3	1.7e–05	–1.3	0.21	–3.9	8.6e–05
17	*PEMT*	none reported	Brain Anteriorcingulatecortex BA24	4.2	2.9e–05	–1.2	0.25	2.7	0.0073
17	*SPATA20*	none reported	Esophagus GastroesophagealJunction	4.3	1.4e–05	1.6	0.1	4.2	2.7e–05
17	*SNX11*	none reported	Pancreas	3.9	1e–04	0.36	0.72	3.1	0.0019
19	*APOC1*	*APOC1*	Esophagus Mucosa	5.3	1.2e–07	NA	NA	NA	NA
19	*KLC3*	*APOC1*	Brain Cerebellum	4.5	6.4e–06	1.9	0.055	4.3	1.5e–05
19	*CLASRP*	*APOC1*	Heart LeftVentricle	4.6	3.8e–06	0.5	0.62	3.8	0.00012
19	*JUND*	none reported	Brain CerebellarHemisphere	4	5.2e–05	–0.69	0.49	2.8	0.0056
19	*ELANE*	none reported	Brain Cortex	–4.1	4e–05	–0.36	0.72	–3.3	0.00085
20	*R3HDML*	*HNF4A,R3HDML,FITM2, GDAP1L1,C20orf111*	Skin NotSunExposed Suprapubic	5.1	3.5e–07	0.54	0.59	4.2	3e–05
20	*SEMG1*	*HNF4A,R3HDML,FITM2, GDAP1L1,C20orf111*	Artery Coronary	4.6	5.2e–06	0.14	0.89	3.6	0.00032
20	*IFT52*	*HNF4A,R3HDML,FITM2, GDAP1L1,C20orf111*	Cells EBV transformedlymphocytes	4.4	1.2e–05	1.3	0.18	4.1	4.7e–05
20	*GTSF1L*	*HNF4A,R3HDML,FITM2, GDAP1L1,C20orf111*	Cells Transformedfibroblasts	–5.8	6.9e–09	–2.9	0.0034	–5.9	3e–09

Table 3.2 (Continued): **MetaXcan associations with T2D.** Results for genes and corresponding models that meet genome-wide significance *in at least one model* from the DIAGRAM analysis are shown with nearby genes and results from the GERA replication study and meta-analysis of DIAGRAM and GERA Metaxcan associations. Blue shading denotes genes not implicated by the top $1,000$ SNPs from the DIAGRAM trans-ethnic meta-analysis of GWAS. Pink and red shading denote genome-wide significance in one model and across all models, respectively, for the DIAGRAM and meta-analysis. Replication in the GERA study is denoted by a pink outline.

Rank	GO:BP Pathway	P-value
1	negative regulation of type B pancreatic cell apoptotic process	0.0001
2	AP-3 adaptor complex	0.0002
3	carbohydrate homeostasis	0.0003
4	glucose homeostasis	0.0003
5	regulation of type B pancreatic cell apoptotic process	0.0004
6	fat cell proliferation	0.0004
7	regulation of fat cell proliferation	0.0004
8	type B pancreatic cell apoptotic process	0.0004
9	monocarboxylic acid binding	0.001
10	core promoter binding	0.001
11	axolemma	0.001
12	multicellular organism growth	0.001
13	thyroid gland development	0.001
14	response to glucose	0.002
15	regulation of receptor biosynthetic process	0.002
16	response to hexose	0.002
17	negative regulation of cellular process	0.002
18	response to monosaccharide	0.002
19	receptor biosynthetic process	0.002
20	core promoter sequence-specific DNA binding	0.002
21	regulation of insulin secretion	0.002
22	response to carbohydrate	0.003
23	regulation of circadian rhythm	0.003
24	pancreas development	0.003
25	response to methyl methanesulfonate	0.003
26	cellular response to methyl methanesulfonate	0.003
27	regulation of N-terminal peptidyl-lysine acetylation	0.003
28	positive regulation of N-terminal peptidyl-lysine acetylation	0.003
29	negative regulation of ATF6-mediated unfolded protein response	0.003
30	response to metformin	0.003
31	negative regulation of pancreatic stellate cell proliferation	0.003
32	fumarylacetoacetase activity	0.003
33	hepatic duct development	0.003
34	hepatoblast differentiation	0.003
35	oxidative DNA demethylase activity	0.003
36	negative regulation of phosphatidylcholine catabolic process	0.003
37	respiratory system process	0.003
38	regulation of oligodendrocyte differentiation	0.003
39	fatty acid binding	0.003
40	epithelial tube branching involved in lung morphogenesis	0.004
41	regulation of collagen biosynthetic process	0.004
42	regulation of protein export from nucleus	0.004
43	regulation of cell proliferation	0.004
44	regulation of peptide hormone secretion	0.004
45	negative regulation of epithelial cell apoptotic process	0.004
46	response to testosterone	0.004
47	regulation of peptide secretion	0.004
48	insulin secretion	0.004
49	regulation of collagen metabolic process	0.005
50	regulation of multicellular organismal metabolic process	0.005

Table 3.3: **Biological pathways enriched among genes associated with T2D from MetaXcan analysis of DIAGRAM dataset.** The top 50 Gene Ontology Biological Process (GO:BP) pathways enriched among the set of MetaXcan-significant genes are shown with overrepresented p-value.

Rank	GO:BP Pathway	P-value
1	regulation of protein export from nucleus	0.0001
2	negative regulation of type B pancreatic cell apoptotic process	0.0001
3	carbohydrate homeostasis	0.0001
4	glucose homeostasis	0.0001
5	regulation of cell proliferation	0.0002
6	regulation of type B pancreatic cell apoptotic process	0.0002
7	fat cell proliferation	0.0002
8	regulation of fat cell proliferation	0.0002
9	type B pancreatic cell apoptotic process	0.0003
10	protein export from nucleus	0.0004
11	cyclin-dependent protein serine/threonine kinase inhibitor activity	0.0005
12	nuclear export	0.0005
13	multicellular organism growth	0.001
14	response to glucose	0.001
15	glial cell differentiation	0.001
16	oligodendrocyte differentiation	0.001
17	response to hexose	0.001
18	small molecule binding	0.001
19	thyroid gland development	0.001
20	response to monosaccharide	0.001
21	regulation of receptor biosynthetic process	0.001
22	regulation of cyclin-dependent protein serine/threonine kinase activity	0.002
23	response to carbohydrate	0.002
24	receptor biosynthetic process	0.002
25	pancreas development	0.002
26	heterocyclic compound binding	0.002
27	cell proliferation	0.002
28	regulation of sequence-specific DNA binding transcription factor activity	0.002
29	gliogenesis	0.002
30	negative regulation of transcription from RNA polymerase II promoter	0.002
31	growth	0.002
32	cyclin-dependent protein serine/threonine kinase regulator activity	0.002
33	organic cyclic compound binding	0.002
34	regulation of oligodendrocyte differentiation	0.002
35	respiratory system process	0.002
36	transcription from RNA polymerase II promoter	0.002
37	epithelial tube branching involved in lung morphogenesis	0.003
38	protein serine/threonine kinase inhibitor activity	0.003
39	viral capsid	0.003
40	regulation of N-terminal peptidyl-lysine acetylation	0.003
41	positive regulation of N-terminal peptidyl-lysine acetylation	0.003
42	negative regulation of ATF6-mediated unfolded protein response	0.003
43	oxidative DNA demethylase activity	0.003
44	hepatic duct development	0.003
45	hepatoblast differentiation	0.003
46	response to metformin	0.003
47	negative regulation of pancreatic stellate cell proliferation	0.003
48	protein stabilization	0.003
49	nucleotide binding	0.003
50	nucleoside phosphate binding	0.003

Table 3.4: **Biological pathways enriched among genes associated with T2D from meta-analysis of results from MetaXcan analyses of DIAGRAM and GERA datasets.** The top 50 Gene Ontology Biological Process (GO:BP) pathways enriched among the set of MetaXcan-significant genes are shown with overrepresented p-value.

Rank	GO:BP Pathway	P-value
1	fumarylacetoacetase activity	0.0003
2	hydrolase activity, acting on acid carbon-carbon bonds	0.001
3	hydrolase activity, acting on acid carbon-carbon bonds, in ketonic substances	0.001
4	positive regulation of circadian sleep/wake cycle, REM sleep	0.001
5	negative regulation of glomerular filtration	0.001
6	negative regulation of urine volume	0.001
7	negative regulation of renal sodium excretion	0.001
8	tyrosine catabolic process	0.001
9	regulation of circadian sleep/wake cycle, wakefulness	0.001
10	positive regulation of circadian sleep/wake cycle, wakefulness	0.001
11	circadian sleep/wake cycle, wakefulness	0.001
12	regulation of circadian sleep/wake cycle, REM sleep	0.001
13	positive regulation of circadian sleep/wake cycle, sleep	0.002
14	positive regulation of synaptic transmission, cholinergic	0.002
15	circadian sleep/wake cycle, REM sleep	0.002
16	arginine catabolic process	0.002
17	negative regulation of heart rate	0.002
18	tyrosine metabolic process	0.002
19	L-phenylalanine catabolic process	0.003
20	erythrose 4-phosphate/phosphoenolpyruvate family amino acid catabolic process	0.003
21	regulation of synaptic transmission, cholinergic	0.003
22	positive regulation of fibroblast migration	0.003
23	regulation of glomerular filtration	0.003
24	L-phenylalanine metabolic process	0.003
25	erythrose 4-phosphate/phosphoenolpyruvate family amino acid metabolic process	0.003
26	glomerular filtration	0.004
27	protein targeting to peroxisome	0.004
28	peroxisomal transport	0.004
29	protein localization to peroxisome	0.004
30	establishment of protein localization to peroxisome	0.004
31	positive regulation of circadian rhythm	0.004
32	renal system process involved in regulation of blood volume	0.004
33	renal filtration	0.004
34	arginine metabolic process	0.004
35	regulation of circadian sleep/wake cycle, sleep	0.004
36	regulation of urine volume	0.005
37	negative regulation of heart contraction	0.005
38	renal sodium excretion	0.005
39	regulation of renal sodium excretion	0.005
40	regulation of circadian sleep/wake cycle	0.005
41	positive regulation of collagen metabolic process	0.005
42	positive regulation of collagen biosynthetic process	0.005
43	positive regulation of multicellular organismal metabolic process	0.005
44	aromatic amino acid family catabolic process	0.005
45	circadian sleep/wake cycle, sleep	0.005
46	positive regulation of heart rate	0.005
47	positive regulation of vasodilation	0.006
48	renal system process involved in regulation of systemic arterial blood pressure	0.006
49	circadian sleep/wake cycle process	0.006
50	glutamine family amino acid catabolic process	0.006

Table 3.5: **Biological pathways enriched among "unknown" loci genes associated with T2D from MetaXcan analysis of DIAGRAM dataset.** The top 50 Gene Ontology Biological Process (GO:BP) pathways enriched among the set of MetaXcan-significant genes located beyond 1 Mb of the top 89 putative T2D genes reported from the DIAGRAM meta-analysis of GWAS are shown with overrepresented p-value.

Trait	P-value	Reported T2D Genes	Novel Genes
Type 2 diabetes	0.00001	*TCF7L2, HHEX, FTO JAZF1, HMG20A, WFS1 PPARG, TP53INP1, R3HDML PRC1, AP3S2, INTS8 MPHOSPH9, KLHDC5, CAMK1D KCNJ11, ANK1, AP3S2, ZFAND6*	
Type 2 diabetes and other traits	0.0002	*TCF7L2, WFS1*	
Body mass index	0.0004	*TCF7L2, FTO, KCNJ11 PPARG, CPNE4*	*DGKG*
HIV-1 control	0.001		*ZNRD1, HLA-A*
IgE levels	0.001		*HCG27, HLA-A*
Fasting insulin-related traits (interaction with BMI)	0.002	*TCF7L2, PPARG*	
Weight	0.002	*FTO*	*DGKG*
Glycated hemoglobin levels	0.002	*TCF7L2, ANK1*	
Sasang constitutional medicine type (So-Eum)	0.003	*FTO*	
Body mass in chronic obstructive pulmonary disease	0.003	*FTO*	
Drug-induced liver injury (amoxicillin-clavulanate)	0.003	*HLA-A*	
Drug-induced liver injury	0.003	*PPARG*	
Breast cancer	0.005	*TCF7L2, FTO, PRC1*	
Free thyroxine concentration	0.006	*JAZF1*	
Height adjusted BMI	0.006	*FTO*	
Essential tremor	0.006	*LINGO1*	
Metabolic syndrome	0.006	*TCF7L2, FTO*	
Breast Cancer in BRCA1 mutation carriers	0.007	*TCF7L2*	
Change in intraocular pressure in response to steroid treatment (triamcinolone acetonide)	0.009		*HLA-A*
Apolipoprotein Levels	0.009	*APOC1*	
Vitiligo	0.012		*CASP7, HLA-A*
Biomedical quantitative traits	0.012	*FTO*	
Vincristine-induced peripheral neuropathy in acute lymphoblastic leukemia	0.015	*NDUFAF6*	
Educational attainment	0.015	*MPHOSPH9*	*C12orf65*
Two-hour glucose challenge	0.016	*TCF7L2*	
Beta-2 microglubulin plasma levels	0.021		*HLA-A*
Triglycerides	0.021	*FTO, APOC1*	
Dietary macronutrient intake	0.021	*FTO*	
Complement C3 and C4 levels	0.022		*HLA-A*
Adiposity	0.023	*FTO*	
Nasopharyngeal carcinoma	0.025		*HLA-A*
Bone mineral density	0.026	*KLHDC5*	*SOX4*
Osteoarthritis	0.026	*FTO*	
Bone mineral density (paediatric, lower limb)	0.027	*KLHDC5*	
Bipolar disorder (body mass index interaction)	0.027	*TCF7L2*	
Plasminogen activator inhibitor type 1 levels (PAI-1)	0.028	*PPARG*	
Dehydroepiandrosterone sulphate levels	0.028	*HHEX*	
LDL cholesterol	0.030	*APOC1*	*GPAM*
Retinopathy in non-diabetics	0.031	*KLHDC5*	
Cholesterol, total	0.034	*APOC1*	*GPAM*
QT interval (interaction)	0.034	*CAMK1D*	
Proinsulin levels	0.034	*TCF7L2*	
Bone mineral density (paediatric, total body less head)	0.039	*KLHDC5*	
Obesity (extreme)	0.040	*FTO*	
Pulmonary function decline	0.042	*ANK1*	
AIDS progression	0.042		*ZNRD1*
Alzheimer's disease (age of onset)	0.046	*APOC1*	
Fasting glucose-related traits	0.050	*TCF7L2*	

Table 3.6: **MetaXcan-significant genes overlap with putative trait genes implicated by GWAS across multiple complex traits.** We performed a sampling study to to test for enrichment of putative trait genes among the set of T2D genes implicated by our MetaXcan analysis of the DIAGRAM trans-ethnic study. The enrichment p-value is shown for each trait that significantly shares putative genes (implicated by GWAS) in common with the set of MetaXcan-signficant T2D genes. Moreover, the table shows the shared genes for each trait and indicates whether the gene is a putative T2D gene (i.e. either in the set of 89 genes implicated by the top $1,000$ SNPs from the DIAGRAM trans-ethnic study or reported in the NHGRI-EBI catalogue for Type 2 diabetes) or is a novel candidate T2D gene implicated by MetaXcan.

CHAPTER 4

CONCLUSIONS

The work presented in this dissertation builds upon previous research that established a strong relationship between genetic variation that regulates gene expression (i.e. eQTLs) and loci associated with complex diseases. Nicolae et al. (2010) showed that eQTLs were highly enriched among sets of complex trait-associated SNPs and Below et al. (2011) found that eQTLs mapped in human adipose and skeletal muscle were significantly overrepresented among SNPs that associated with T2D in the Starr County Mexican American cohort [14, 111]. However, the set of SNPs achieving genome-wide significance from GWAS - though numerous and greatly expanded over the loci discovered from linkage and candidate gene studies - explain only a small fraction of the heritability estimated from family studies [100, 15]. We hypothesized that eQTLs mapped in insulin-responsive peripheral tissues would disproportionately account for much of the overall genetic component of T2D risk. In order to address this hypothesis, we partitioned the estimate of phenotypic variance explained by common SNPs interrogated in GWAS (i.e. h^2_{chip}) by subsets comprised of eQTLs.

We first estimated h^2_{chip} in GWAS datasets representing populations of European and Mexican ancestry and found that these estimates not only exceeded the phenotypic variance accounted for by GWAS-significant SNPs (Figure 2.2), but were also consistent with h^2 estimates garnered from family-based studies [81, 4]. Our estimate of h^2_{chip} in the WTCCC dataset was also comparable to other estimates of chip heritability in European populations [140, 104]. Importantly, these results indicated that much of the "missing" heritability in these datasets can be accounted for by variants in LD with common SNPs represented in GWAS.

Next, we performed a series of heritability partitioning analyses whereby we jointly estimated the proportions of h^2_{chip} accounted for by eQTLs mapped in either LCLs or adipose and skeletal muscle tissue. eQTLs mapped in insulin-responsive peripheral tissues significantly explained large proportions of chip heritability that exceeded the heritability accounted for

by eQTLs mapped in LCLs (Figures 2.3-2.4). Moreover, we observed that eQTLs mapped in two or more tissues (including LCLs) disproportionately accounted for the most chip heritability in the WTCCC relatively to the number of SNPs included in the subset (Figure 2.3) - indicating a pronounced contribution from "cross-tissue" eQTLs. Furthermore, we corroborated the robustness of these relationships through a series of permutation analyses and analyses adjusting for genetic relatedness, MAF distribution, differences in local LD at "tagged" SNPs and other factors that may inflate h^2_{chip} estimates.

eQTLs mapped in adipose tissue significantly explained heritability in the Hispanic cohort where more than half of the subjects were obese 2.4. We performed BMI-stratified analyses and found that "cross-tissue" eQTLs yielded the highest h^2_{chip} estimates in the non-obese cohort - similar to the results from the WTCCC study. However, adipose-specific eQTLs accounted for the most heritability in the obese cohort and provided support for distinct genetic architectures of T2D. Indeed, this observation is consistent with results from a large-scale, BMI-stratified GWAS of T2D that discovered T2D risk alleles that were significant in either a lean cohort or an obese cohort, but not both [120].

Taken together, these results extend the observations of Nicolae et al. (2010) and Below et al. (2011) to the whole genome whereby metabolic-tissue eQTLs are not only enriched among GWAS-significant variants but also "concentrate" the overall genetic component of T2D risk. Moreover, these results are consistent with the findings of Gusev et al. (2014) whereby the authors partitioned chip heritability for 11 complex traits according to regulatory annotations from the ENCODE project and found that heritability was predominantly concentrated in DHS regions [60]. Degner et al. (2012) had previously shown that 55% of eQTLs in Yoruba LCLs (West African descent) were also associated with chromatin accessibility at DHS [27]. Therefore, this work further demonstrates the importance of inherited variants that regulate gene expression on T2D and provides avenues for further research.

Although, eQTLs mapped in adipose and skeletal muscle tissues accounted for disproportionate shares of chip heritability - providing genetic support for the importance of tissues

involved in insulin signaling - this does not preclude the influence of eQTLs that regulate gene expression in pancreatic β-cells. Indeed, variants in the cross-tissue eQTL subsets may have regulatory effects in β-cells as well as other T2D-relevant tissues (i.e. liver, brain, kidneys, GI tract, α-cells, etc.). As maps of the regulatory landscape for primary human tissues improve with greater sample sizes, researchers can use information gleaned from these efforts to further delineate parsimonious SNP sets that capture much of the heritability of T2D. This would inform fine-mapping studies and could be used to elucidate transcripts involved in genetic mechanisms that promote T2D. Lastly, the contribution of population-specific eQTLs is yet to be determined. Although, eQTLs mapped in Europeans and African Americans disproportionately explained phenotypic variance in the Hispanic cohort, this is only a lower bound as eQTL-mapping studies have largely been directed to populations with less complex genetic admixture [147].

We also leveraged eQTL information to confront the challenge of mapping disease genes in light of the fact that the majority of trait-associated SNPs are in non-coding regions of the genome [67]. We applied the PrediXcan approach to estimate the *genetic component* of gene expression and perform an *in silico* study that directly addressed the mechanism of transcription [9]. We leveraged 42 prediction models trained on genotype and expression data from the DGN study and GTEx Project (the most comprehensive resource of such data from primary human tissues) to predict gene expression from GWAS summary data from the DIAGRAM trans-ethnic meta-analysis (i.e. the largest such study to date). Through this extensive set of analyses, we provided support for 109 tissue-level associations and implicated 52 putative T2D genes from regulatory genetic information rather than proximity-based assumptions. Furthermore, 63 of these associations and 25 genes replicated in an independent cohort representing a medical community.

The vast majority of gene associations mapped to known T2D loci. However, many of these associations corresponded to genes that had not been previously implicated by genetic studies of T2D and therefore represent novel T2D genes. Gene set enrichment analyses

revealed that sets of significantly-associated genes were overrepresented for processes that involve the endocrine pancreas and insulin signaling in peripheral tissues. Moreover, genes mapping to regions outside of loci harboring T2D-associated SNPs were enriched for processes involved in kidney function and circadian rhythm - indicating that leveraging eQTL information can extend the breadth of results gleaned from GWAS.

Importantly, these results inform the regulatory consequences of T2D-associated variants and may help resolve the relevant disease genes at loci harboring significant single variant associations. We found evidence at several T2D loci that not only supported the putative disease gene but also indicates the direction of expression that confers disease risk. For example, decreased expression of *PPARG* and increased expression of *WFS1* across multiple tissue models significantly associated with T2D (Figure 3.1). Moreover, several loci contained multiple significant gene associations in addition to the putative T2D gene. The most salient example was the *TCF7L2* locus where predicted expression of multiple genes exceeded the most stringent significance threshold (Figure 3.2). Moreover, *GPAM* - a novel T2D gene previously implicated as an LDL cholesterol gene from GWAS - exhibited the strongest gene-level association at the *TCF7L2* locus.

Presumably, instances where the predicted expression of multiple genes (including novel candidate genes) associated with T2D can be attributed to risk alleles having multiple regulatory effects at a locus. We are currently working to explore the correlation structure between predicted expression traits and the LD relationships between sets of predictor SNPs used to estimate the expression of genes that reach genome-wide significance at T2D loci. Moreover, we will determine the basis for divergent directional associations (e.g. *KCNJ11*) and resolve whether there are independent signals (i.e. predictor SNPs with low LD) between models or if the same SNPs have divergent effects on expression in different tissues. However, the results from this body of work demonstrate that more insight can be obtained by incorporating information gleaned from eQTL mapping studies into genetic studies of T2D.

References

[1] Vignini A., Raffaelli F., Cester A., Iannilli A., Cherubini V., Mazzanti L., and Nanetti L. Environmental and genetical aspects of the link between pregnancy, birth size, and type 2 diabetes, 2012.

[2] Frank W Albert and Leonid Kruglyak. The role of regulatory variation in complex traits and disease. *Nat Rev Genet*, 16(4):197–212, apr 2015.

[3] K G M M Alberti, P Zimmet, and J Shaw. International Diabetes Federation: a consensus on Type 2 diabetes prevention. *Diabetic medicine : a journal of the British Diabetic Association*, 24(5):451–63, 2007.

[4] P. Almgren, M. Lehtovirta, B. Isomaa, L. Sarelin, M. R. Taskinen, V. Lyssenko, T. Tuomi, and L. Groop. Heritability and familiality of type 2 diabetes and related quantitative traits in the Botnia Study, 2011.

[5] D Altshuler, J N Hirschhorn, M Klannemark, C M Lindgren, M C Vohl, J Nemesh, C R Lane, S F Schaffner, S Bolk, C Brewer, T Tuomi, D Gaudet, T J Hudson, M Daly, L Groop, and E S Lander. The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat Genet*, 26(1):76–80, 2000.

[6] Qasim Ayub, Loukas Moutsianas, Yuan Chen, Kalliope Panoutsopoulou, Vincenza Colonna, Luca Pagani, Inga Prokopenko, Graham R S Ritchie, Chris Tyler-Smith, Mark I. McCarthy, Eleftheria Zeggini, and Yali Xue. Revisiting the thrifty gene hypothesis via 65 loci associated with susceptibility to type 2 diabetes. *American Journal of Human Genetics*, 94(2):176–185, 2014.

[7] L J Baier, P a Permana, X Yang, R E Pratley, R L Hanson, G Q Shen, D Mott, W C Knowler, N J Cox, Y Horikawa, N Oda, G I Bell, and C Bogardus. A calpain-10 gene polymorphism is associated with reduced muscle mRNA levels and insulin resistance. *The Journal of clinical investigation*, 106(7):R69–R73, 2000.

[8] Michael Bamshad and Stephen P Wooding. Signatures of natural selection in the human genome. *Nat Rev Genet*, 4(2):99–111, feb 2003.

[9] Alvaro Barbeira, Kaanan P Shah, Jason M Torres, Heather E Wheeler, Eric S Torstenson, Todd Edwards, Tzintzuni Garcia, Graeme I Bell, Dan Nicolae, Nancy J Cox, and Hae Kyung Im. MetaXcan: Summary Statistics Based Gene-Level Association Method Infers Accurate PrediXcan Results. *bioRxiv*, mar 2016.

[10] David J P Barker. Adult consequences of fetal growth restriction. *Clinical Obstetrics and Gynecology*, 49(2):270–283, 2006.

[11] Douglas Bates, Martin Mächler, Benjamin M. Bolker, and Steven C. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

[12] Alexis Battle, Sara Mostafavi, Xiaowei Zhu, James B Potash, Myrna M Weissman, Courtney McCormick, Christian D Haudenschild, Kenneth B Beckman, Jianxin Shi, Rui Mei, Alexander E Urban, Stephen B Montgomery, Douglas F Levinson, and Daphne Koller. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Research*, 24(1):14–24, jan 2014.

[13] Harold Bays, Lawrence Mandarino, and Ralph A DeFronzo. Role of the Adipocyte, Free Fatty Acids, and Ectopic Fat in Pathogenesis of Type 2 Diabetes Mellitus: Peroxisomal Proliferator-Activated Receptor Agonists Provide a Rational Therapeutic Approach. *The Journal of Clinical Endocrinology & Metabolism*, 89(2):463–478, feb 2004.

[14] J.E. Below, E.R. Gamazon, J.V. Morrison, A. Konkashbaev, A. Pluzhnikov, P.M. McKeigue, E.J. Parra, S.C. Elbein, D.M. Hallman, D.L. Nicolae, G.I. Bell, M. Cruz, N.J. Cox, and C.L. Hanis. Genome-wide association and meta-analysis in populations from Starr County, Texas, and Mexico City identify type 2 diabetes susceptibility loci and enrichment for expression quantitative trait loci in top signals. *Diabetologia*, 54(8):2047–2055, 2011.

[15] Liana K Billings and Jose C Florez. The genetics of type 2 diabetes: what have we learned from GWAS? *Annals of the New York Academy of Sciences*, 1212:59–77, 2010.

[16] G Boden. Role of fatty acids in the pathogenesis of insulin resistance and NIDDM. *Diabetes*, 46:3–10, 1997.

[17] Josiane L. Broussard, David A. Ehrmann, Eve Van Cauter, Esra Tasali, and Matthew J. Brady. Impaired insulin signaling in human adipocytes after experimental sleep restriction: A randomized, crossover study. *Annals of Internal Medicine*, 157(8):549–557, 2012.

[18] Jason D Buenrostro, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods*, 10(12):1213–8, 2013.

[19] A Carpentier, S D Mittelman, R N Bergman, A Giacca, and G F Lewis. Prolonged elevation of plasma free fatty acids impairs pancreatic beta-cell function in obese nondiabetic humans but not in individuals with type 2 diabetes. *Diabetes*, 49:399–408, 2000.

[20] Yoon Shin Cho, Chien-Hsiun Chen, Cheng Hu, Jirong Long, Rick Twee Hee Ong, Xueling Sim, Fumihiko Takeuchi, Ying Wu, Min Jin Go, Toshimasa Yamauchi, Yi-Cheng Chang, Soo Heon Kwak, Ronald C W Ma, Ken Yamamoto, Linda S Adair, Tin Aung, Qiuyin Cai, Li-Ching Chang, Yuan-Tsong Chen, Yutang Gao, Frank B Hu, Hyung-Lae Kim, Sangsoo Kim, Young Jin Kim, Jeannette Jen-Mai Lee, Nanette R Lee, Yun Li, Jian Jun Liu, Wei Lu, Jiro Nakamura, Eitaro Nakashima, Daniel Peng-Keat Ng, Wan Ting Tay, Fuu-Jen Tsai, Tien Yin Wong, Mitsuhiro Yokota, Wei Zheng,

Rong Zhang, Congrong Wang, Wing Yee So, Keizo Ohnaka, Hiroshi Ikegami, Kazuo Hara, Young Min Cho, Nam H Cho, Tien-Jyun Chang, Yuqian Bao, Åsa K Hedman, Andrew P Morris, Mark I McCarthy, Ryoichi Takayanagi, Kyong Soo Park, Weiping Jia, Lee-Ming Chuang, Juliana C N Chan, Shiro Maeda, Takashi Kadowaki, Jong-Young Lee, Jer-Yuarn Wu, Yik Ying Teo, E Shyong Tai, Xiao Ou Shu, Karen L Mohlke, Norihiro Kato, Bok-Ghee Han, and Mark Seielstad. Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. *Nature Genetics*, 44(1):67–72, 2011.

[21] Lea K Davis, Dongmei Yu, Clare L Keenan, Eric R Gamazon, Anuar I Konkashbaev, Eske M Derks, Benjamin M Neale, Jian Yang, S Hong Lee, Patrick Evans, Cathy L Barr, Laura Bellodi, Fortu Benarroch, Gabriel Bedoya Berrio, Oscar J Bienvenu, Michael H Bloch, Rianne M Blom, Ruth D Bruun, Cathy L Budman, Beatriz Camarena, Desmond Campbell, Carolina Cappi, Julio C Cardona Silgado, Danielle C Cath, Maria C Cavallini, Denise a Chavira, Sylvain Chouinard, David V Conti, Edwin H Cook, Vladimir Coric, Bernadette a Cullen, Dieter Deforce, Richard Delorme, Yves Dion, Christopher K Edlund, Karin Egberts, Peter Falkai, Thomas V Fernandez, Patience J Gallagher, Helena Garrido, Daniel Geller, Simon L Girard, Hans J Grabe, Marco a Grados, Benjamin D Greenberg, Varda Gross-Tsur, Stephen Haddad, Gary a Heiman, Sian M J Hemmings, Ana G Hounie, Cornelia Illmann, Joseph Jankovic, Michael a Jenike, James L Kennedy, Robert a King, Barbara Kremeyer, Roger Kurlan, Nuria Lanzagorta, Marion Leboyer, James F Leckman, Leonhard Lennertz, Chunyu Liu, Christine Lochner, Thomas L Lowe, Fabio Macciardi, James T McCracken, Lauren M McGrath, Sandra C Mesa Restrepo, Rainald Moessner, Jubel Morgan, Heike Muller, Dennis L Murphy, Allan L Naarden, William Cornejo Ochoa, Roel a Ophoff, Lisa Osiecki, Andrew J Pakstis, Michele T Pato, Carlos N Pato, John Piacentini, Christopher Pittenger, Yehuda Pollak, Scott L Rauch, Tobias J Renner, Victor I Reus, Margaret a Richter, Mark a Riddle, Mary M Robertson, Roxana Romero, Maria C Rosàrio, David Rosenberg, Guy a Rouleau, Stephan Ruhrmann, Andres Ruiz-Linares, Aline S Sampaio, Jack Samuels, Paul Sandor, Brooke Sheppard, Harvey S Singer, Jan H Smit, Dan J Stein, E Strengman, Jay a Tischfield, Ana V Valencia Duarte, Homero Vallada, Filip Van Nieuwerburgh, Jeremy Veenstra-Vanderweele, Susanne Walitza, Ying Wang, Jens R Wendland, Herman G M Westenberg, Yin Yao Shugart, Euripedes C Miguel, William McMahon, Michael Wagner, Humberto Nicolini, Danielle Posthuma, Gregory L Hanna, Peter Heutink, Damiaan Denys, Paul D Arnold, Ben a Oostra, Gerald Nestadt, Nelson B Freimer, David L Pauls, Naomi R Wray, S Evelyn Stewart, Carol a Mathews, James a Knowles, Nancy J Cox, and Jeremiah M Scharf. Partitioning the heritability of tourette syndrome and obsessive compulsive disorder reveals differences in genetic architecture. *PLoS genetics*, 9:e1003864, 2013.

[22] N. Maneka G De Silva, Rachel M. Freathy, Tom M. Palmer, Louise A. Donnelly, Jian'an Luan, Tom Gaunt, Claudia Langenberg, Michael N. Weedon, Beverley Shields, Beatrice A. Knight, Kirsten J. Ward, Manjinder S. Sandhu, Roger M. Harbord, Mark I. McCarthy, George Davey Smith, Shah Ebrahim, Andrew T. Hattersley, Nicholas Wareham, Debbie A. Lawlor, Andrew D. Morris, Colin N A Palmer, and Timothy M.

Frayling. Mendelian randomization studies do not support a role for raised circulating triglyceride levels influencing type 2 diabetes, glucose levels, or insulin resistance. *Diabetes*, 60(3):1008–1018, 2011.

[23] R. A. DEFRONZO. The triumvirate: β-cell, muscle, liver. A collusion responsible for NIDDM. *Diabetes*, 37(6):667–687, 1988.

[24] Ralph A DeFronzo. Pathogenesis of type 2 diabetes mellitus. *The Medical clinics of North America*, 88:787–835, ix, 2004.

[25] Ralph A. DeFronzo. From the Triumvirate to the Ominous Octet: A New Paradigm for the Treatment of Type 2 Diabetes Mellitus. *Diabetes Care*, 58(4):773–795, 2009.

[26] Ralph A. DeFronzo, Eleuterio Ferrannini, and Donald C. Simonson. Fasting hyperglycemia in non-insulin-dependent diabetes mellitus: Contributions of excessive hepatic glucose production and impaired tissue glucose uptake. *Metabolism*, 38(4):387–395, 1989.

[27] Jacob F. Degner, Athma A. Pai, Roger Pique-Regi, Jean-Baptiste Veyrieras, Daniel J. Gaffney, Joseph K. Pickrell, Sherryl De Leon, Katelyn Michelini, Noah Lewellen, Gregory E. Crawford, Matthew Stephens, Yoav Gilad, and Jonathan K. Pritchard. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*, 482(7385):390–394, 2012.

[28] Olivier Delaneau, Jonathan Marchini, and Jean-François Zagury. A linear complexity phasing method for thousands of genomes. *Nature methods*, 9(2):179–81, 2012.

[29] Christopher M. Depner, Ellen R. Stothard, and Kenneth P. Wright. Metabolic consequences of sleep and circadian disorders. *Current Diabetes Reports*, 14(7), 2014.

[30] M P Diamond, K Thornton, M Connolly-Diamond, R S Sherwin, and R A DeFronzo. Reciprocal variations in insulin-stimulated glucose uptake and pancreatic insulin secretion in women with normal glucose tolerance. *Journal of the Society for Gynecologic Investigation*, 2(5):708–15, 1995.

[31] Samuel P Dickson, Kai Wang, Ian Krantz, Hakon Hakonarson, and David B Goldstein. Rare variants create synthetic genome-wide associations. *PLoS biology*, 8:e1000294, 2010.

[32] J H Dominguez, K Camp, L Maianu, H Feister, and W T Garvey. Molecular adaptations of GLUT1 and GLUT2 in renal proximal tubules of diabetic rats. *The American journal of physiology*, 266(2 Pt 2):F283–90, 1994.

[33] Alfred Doyle, Michael P McGarry, Nancy A Lee, and James J Lee. The construction of transgenic and gene knockout/knockin mouse models of human disease. *Transgenic Research*, 21(2):327–349, 2011.

[34] Daniel J. Drucker. The biology of incretin hormones, 2006.

[35] Josée Dupuis, Claudia Langenberg, Inga Prokopenko, Richa Saxena, Nicole Soranzo, Anne U Jackson, Eleanor Wheeler, Nicole L Glazer, Nabila Bouatia-Naji, Anna L Gloyn, Cecilia M Lindgren, Reedik Mägi, Andrew P Morris, Joshua Randall, Toby Johnson, Paul Elliott, Denis Rybin, Gudmar Thorleifsson, Valgerdur Steinthorsdottir, Peter Henneman, Harald Grallert, Abbas Dehghan, Jouke Jan Hottenga, Christopher S Franklin, Pau Navarro, Kijoung Song, Anuj Goel, John R B Perry, Josephine M Egan, Taina Lajunen, Niels Grarup, Thomas Sparsø, Alex Doney, Benjamin F Voight, Heather M Stringham, Man Li, Stavroula Kanoni, Peter Shrader, Christine Cavalcanti-Proença, Meena Kumari, Lu Qi, Nicholas J Timpson, Christian Gieger, Carina Zabena, Ghislain Rocheleau, Erik Ingelsson, Ping An, Jeffrey O'Connell, Jian'an Luan, Amanda Elliott, Steven A McCarroll, Felicity Payne, Rosa Maria Roccasecca, François Pattou, Praveen Sethupathy, Kristin Ardlie, Yavuz Ariyurek, Beverley Balkau, Philip Barter, John P Beilby, Yoav Ben-Shlomo, Rafn Benediktsson, Amanda J Bennett, Sven Bergmann, Murielle Bochud, Eric Boerwinkle, Amélie Bonnefond, Lori L Bonnycastle, Knut Borch-Johnsen, Yvonne Böttcher, Eric Brunner, Suzannah J Bumpstead, Guillaume Charpentier, Yii-Der Ida Chen, Peter Chines, Robert Clarke, Lachlan J M Coin, Matthew N Cooper, Marilyn Cornelis, Gabe Crawford, Laura Crisponi, Ian N M Day, Eco J C de Geus, Jerome Delplanque, Christian Dina, Michael R Erdos, Annette C Fedson, Antje Fischer-Rosinsky, Nita G Forouhi, Caroline S Fox, Rune Frants, Maria Grazia Franzosi, Pilar Galan, Mark O Goodarzi, Jürgen Graessler, Christopher J Groves, Scott Grundy, Rhian Gwilliam, Ulf Gyllensten, Samy Hadjadj, Göran Hallmans, Naomi Hammond, Xijing Han, Anna-Liisa Hartikainen, Neelam Hassanali, Caroline Hayward, Simon C Heath, Serge Hercberg, Christian Herder, Andrew A Hicks, David R Hillman, Aroon D Hingorani, Albert Hofman, Jennie Hui, Joe Hung, Bo Isomaa, Paul R V Johnson, Torben Jørgensen, Antti Jula, Marika Kaakinen, Jaakko Kaprio, Y Antero Kesaniemi, Mika Kivimaki, Beatrice Knight, Seppo Koskinen, Peter Kovacs, Kirsten Ohm Kyvik, G Mark Lathrop, Debbie A Lawlor, Olivier Le Bacquer, Cécile Lecoeur, Yun Li, Valeriya Lyssenko, Robert Mahley, Massimo Mangino, Alisa K Manning, María Teresa Martínez-Larrad, Jarred B McAteer, Laura J McCulloch, Ruth McPherson, Christa Meisinger, David Melzer, David Meyre, Braxton D Mitchell, Mario A Morken, Sutapa Mukherjee, Silvia Naitza, Narisu Narisu, Matthew J Neville, Ben A Oostra, Marco Orrù, Ruth Pakyz, Colin N A Palmer, Giuseppe Paolisso, Cristian Pattaro, Daniel Pearson, John F Peden, Nancy L Pedersen, Markus Perola, Andreas F H Pfeiffer, Irene Pichler, Ozren Polasek, Danielle Posthuma, Simon C Potter, Anneli Pouta, Michael A Province, Bruce M Psaty, Wolfgang Rathmann, Nigel W Rayner, Kenneth Rice, Samuli Ripatti, Fernando Rivadeneira, Michael Roden, Olov Rolandsson, Annelli Sandbaek, Manjinder Sandhu, Serena Sanna, Avan Aihie Sayer, Paul Scheet, Laura J Scott, Udo Seedorf, Stephen J Sharp, Beverley Shields, Gunnar Sigurethsson, Eric J G Sijbrands, Angela Silveira, Laila Simpson, Andrew Singleton, Nicholas L Smith, Ulla Sovio, Amy Swift, Holly Syddall, Ann-Christine Syvänen, Toshiko Tanaka, Barbara Thorand, Jean Tichet, Anke Tönjes, Tiinamaija Tuomi, André G Uitterlinden, Ko Willems van Dijk, Mandy van Hoek, Dhiraj Varma, Sophie Visvikis-Siest, Veronique Vitart, Nicole Vogelzangs, Gérard Waeber, Peter J Wagner, Andrew Walley, G Bragi Walters, Kim L Ward, Hugh Watkins, Michael N Weedon, Sarah H Wild, Gonneke Willemsen, Jaqueline C M Witteman, John W G Yarnell,

Eleftheria Zeggini, Diana Zelenika, Björn Zethelius, Guangju Zhai, Jing Hua Zhao, M Carola Zillikens, DIAGRAM Consortium, GIANT Consortium, Global BPgen Consortium, Ingrid B Borecki, Ruth J F Loos, Pierre Meneton, Patrik K E Magnusson, David M Nathan, Gordon H Williams, Andrew T Hattersley, Kaisa Silander, Veikko Salomaa, George Davey Smith, Stefan R Bornstein, Peter Schwarz, Joachim Spranger, Fredrik Karpe, Alan R Shuldiner, Cyrus Cooper, George V Dedoussis, Manuel Serrano-Ríos, Andrew D Morris, Lars Lind, Lyle J Palmer, Frank B Hu, Paul W Franks, Shah Ebrahim, Michael Marmot, W H Linda Kao, James S Pankow, Michael J Sampson, Johanna Kuusisto, Markku Laakso, Torben Hansen, Oluf Pedersen, Peter Paul Pramstaller, H Erich Wichmann, Thomas Illig, Igor Rudan, Alan F Wright, Michael Stumvoll, Harry Campbell, James F Wilson, Anders Hamsten on behalf of Procardis Consortium, MAGIC investigators, Richard N Bergman, Thomas A Buchanan, Francis S Collins, Karen L Mohlke, Jaakko Tuomilehto, Timo T Valle, David Altshuler, Jerome I Rotter, David S Siscovick, Brenda W J H Penninx, Dorret I Boomsma, Panos Deloukas, Timothy D Spector, Timothy M Frayling, Luigi Ferrucci, Augustine Kong, Unnur Thorsteinsdottir, Kari Stefansson, Cornelia M van Duijn, Yurii S Aulchenko, Antonio Cao, Angelo Scuteri, David Schlessinger, Manuela Uda, Aimo Ruokonen, Marjo-Riitta Jarvelin, Dawn M Waterworth, Peter Vollenweider, Leena Peltonen, Vincent Mooser, Goncalo R Abecasis, Nicholas J Wareham, Robert Sladek, Philippe Froguel, Richard M Watanabe, James B Meigs, Leif Groop, Michael Boehnke, Mark I McCarthy, Jose C Florez, and Inês Barroso. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nature genetics*, 42(2):105–16, 2010.

[36] Robert H. Eckel, Christopher M. Depner, Leigh Perreault, Rachel R. Markwald, Mark R. Smith, Andrew W. McHill, Janine Higgins, Edward L. Melanson, and Kenneth P. Wright. Morning Circadian Misalignment during Short Sleep Duration Impacts Insulin Sensitivity, 2015.

[37] Evan E Eichler, Jonathan Flint, Greg Gibson, Augustine Kong, Suzanne M Leal, Jason H Moore, and Joseph H Nadeau. Missing heritability and strategies for finding the underlying causes of complex disease. *Nature reviews. Genetics*, 11:446–450, 2010.

[38] Steven C Elbein, Eric R Gamazon, Swapan K Das, Neda Rasouli, Philip A Kern, and Nancy J Cox. Genetic risk factors for type 2 diabetes: a trans-regulatory genetic architecture? *American journal of human genetics*, 91:466–77, 2012.

[39] Steven C. Elbein, Michael D. Hoffman, Kui Teng, Mark F. Leppert, and Sandra J. Hasstedt. A genome-wide search for type 2 diabetes susceptibility genes in Utah Caucasians. *Diabetes*, 48(5):1175–1182, 1999.

[40] Anna L A.L. et al. Gloyn, Michael N Weedon, Katharine R Owen, Martina J Turner, Bridget A Knight, Graham Hitman, Mark Walker, Jonathan C Levy, Mike Sampson, Stephanie Halford, Mark I McCarthy, Andrew T Hattersley, and Timothy M Frayling. Large-scale association studies of variants in genes encoding the pancreatic beta-cell KATP channel subunits Kir6.2 (KCNJ11) and SUR1 (ABCC8) confirm that

the KCNJ11 E23K variant is associated with type 2 diabetes. *Diabetes*, 52(2):568–572, feb 2003.

[41] I. Sadaf Farooqi. FTO and Obesity: The missing link. *Cell Metabolism*, 13:7–8, 2011.

[42] Ingo Feinerer, Kurt Hornik, and David Meyer. Text Mining Infrastructure in R. *Journal Of Statistical Software*, 25(5):1–54, 2008.

[43] E Ferrannini, D C Simonson, L D Katz, G Reichard, S Bevilacqua, E J Barrett, M Olsson, and R a DeFronzo. The disposal of an oral glucose load in patients with non-insulin-dependent diabetes. *Metabolism: clinical and experimental*, 37(1):79–85, 1988.

[44] Christodoulos S Flordellis, Ioannis Ilias, and Athanasios G Papavassiliou. New therapeutic options for the metabolic syndrome: what's next? *Trends in Endocrinology & Metabolism*, 16(6):254–260, aug 2005.

[45] Timothy M Frayling, Andrew T Hattersley, Anne McCarthy, Jeff Holly, Simon M S Mitchell, Anna L Gloyn, Katharine Owen, David Davies, George Davey Smith, and Yoav Ben-Shlomo. A putative functional polymorphism in the IGF-I gene: association studies with type 2 diabetes, adult height, glucose tolerance, and fetal growth in U.K. populations. *Diabetes*, 51:2313–2316, 2002.

[46] Timothy M Frayling, Nicholas J Timpson, Michael N Weedon, Eleftheria Zeggini, Rachel M Freathy, Cecilia M Lindgren, John R B Perry, Katherine S Elliott, Hana Lango, Nigel W Rayner, Beverley Shields, Lorna W Harries, Jeffrey C Barrett, Sian Ellard, Christopher J Groves, Bridget Knight, Ann-Marie Patch, Andrew R Ness, Shah Ebrahim, Debbie A Lawlor, Susan M Ring, Yoav Ben-Shlomo, Marjo-Riitta Jarvelin, Ulla Sovio, Amanda J Bennett, David Melzer, Luigi Ferrucci, Ruth J F Loos, Inês Barroso, Nicholas J Wareham, Fredrik Karpe, Katharine R Owen, Lon R Cardon, Mark Walker, Graham A Hitman, Colin N A Palmer, Alex S F Doney, Andrew D Morris, George Davey Smith, Andrew T Hattersley, and Mark I McCarthy. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science (New York, N.Y.)*, 316(5826):889–94, 2007.

[47] Rachel M Freathy, Dennis O Mook-Kanamori, Ulla Sovio, Inga Prokopenko, Nicholas J Timpson, Diane J Berry, Nicole M Warrington, Elisabeth Widen, Jouke Jan Hottenga, Marika Kaakinen, Leslie a Lange, Jonathan P Bradfield, Marjan Kerkhof, Julie a Marsh, Reedik Mägi, Chih-Mei Chen, Helen N Lyon, Mirna Kirin, Linda S Adair, Yurii S Aulchenko, Amanda J Bennett, Judith B Borja, Nabila Bouatia-Naji, Pimphen Charoen, Lachlan J M Coin, Diana L Cousminer, Eco J C de Geus, Panos Deloukas, Paul Elliott, David M Evans, Philippe Froguel, Beate Glaser, Christopher J Groves, Anna-Liisa Hartikainen, Neelam Hassanali, Joel N Hirschhorn, Albert Hofman, Jeff M P Holly, Elina Hyppönen, Stavroula Kanoni, Bridget a Knight, Jaana Laitinen, Cecilia M Lindgren, Wendy L McArdle, Paul F O'Reilly, Craig E Pennell, Dirkje S Postma, Anneli Pouta, Adaikalavan Ramasamy, Nigel W Rayner, Susan M Ring, Fernando Rivadeneira, Beverley M Shields, David P Strachan, Ida Surakka, Anja

Taanila, Carla Tiesler, Andre G Uitterlinden, Cornelia M van Duijn, Alet H Wijga, Gonneke Willemsen, Haitao Zhang, Jianhua Zhao, James F Wilson, Eric a P Steegers, Andrew T Hattersley, Johan G Eriksson, Leena Peltonen, Karen L Mohlke, Struan F a Grant, Hakon Hakonarson, Gerard H Koppelman, George V Dedoussis, Joachim Heinrich, Matthew W Gillman, Lyle J Palmer, Timothy M Frayling, Dorret I Boomsma, George Davey Smith, Chris Power, Vincent W V Jaddoe, Marjo-Riitta Jarvelin, and Mark I McCarthy. Variants in ADCY5 and near CCNL1 are associated with fetal growth and birth weight. *Nature genetics*, 42(5):430–435, 2010.

[48] Eric R Gamazon, Heather E Wheeler, Kaanan P Shah, Sahar V Mozaffari, Keston Aquino-Michaels, Robert J Carroll, Anne E Eyler, Joshua C Denny, Dan L Nicolae, Nancy J Cox, and Hae Kyung Im. A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics*, 47(9):1091–1098, 2015.

[49] Eric R Gamazon, Wei Zhang, Anuar Konkashbaev, Shiwei Duan, Emily O Kistner, Dan L Nicolae, M Eileen Dolan, and Nancy J Cox. SCAN: SNP and copy number annotation. *Bioinformatics (Oxford, England)*, 26:259–262, 2010.

[50] Daphne Sl Gardner and E Shyong Tai. Clinical features and treatment of maturity onset diabetes of the young (MODY). *Diabetes, metabolic syndrome and obesity : targets and therapy*, 5:101–8, 2012.

[51] Greg Gibson. Hints of hidden heritability in GWAS., 2010.

[52] Greg Gibson. Rare and common variants: twenty arguments, 2012.

[53] A L Gloyn, Y Hashim, S J Ashcroft, R Ashfield, S Wiltshire, and R C Turner. Association studies of variants in promoter and coding regions of beta-cell ATP-sensitive K-channel genes SUR1 and Kir6.2 with Type 2 diabetes mellitus (UKPDS 53). Technical report, 2001.

[54] Anna L Gloyn, Ewan R Pearson, Jennifer F Antcliff, Peter Proks, G Jan Bruining, Annabelle S Slingerland, Neville Howard, Shubha Srinivasan, José M C L Silva, Janne Molnes, Emma L Edghill, Timothy M Frayling, I Karen Temple, Deborah Mackay, Julian P H Shield, Zdenek Sumnik, Adrian van Rhijn, Jerry K H Wales, Penelope Clark, Shaun Gorman, Javier Aisenberg, Sian Ellard, Pål R Njølstad, Frances M Ashcroft, and Andrew T Hattersley. Activating Mutations in the Gene Encoding the ATP-Sensitive Potassium-Channel Subunit Kir6.2 and Permanent Neonatal Diabetes. *New England Journal of Medicine*, 350(18):1838–1849, apr 2004.

[55] Struan F a Grant, Gudmar Thorleifsson, Inga Reynisdottir, Rafn Benediktsson, Andrei Manolescu, Jesus Sainz, Agnar Helgason, Hreinn Stefansson, Valur Emilsson, Anna Helgadottir, Unnur Styrkarsdottir, Kristinn P Magnusson, G Bragi Walters, Ebba Palsdottir, Thorbjorg Jonsdottir, Thorunn Gudmundsdottir, Arnaldur Gylfason, Jona Saemundsdottir, Robert L Wilensky, Muredach P Reilly, Daniel J Rader, Yu Bagger,

Claus Christiansen, Vilmundur Gudnason, Gunnar Sigurdsson, Unnur Thorsteinsdottir, Jeffrey R Gulcher, Augustine Kong, and Kari Stefansson. Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nature genetics*, 38(3):320–323, 2006.

[56] L C Groop, R C Bonadonna, S DelPrato, K Ratheiser, K Zyck, E Ferrannini, and R a DeFronzo. Glucose and free fatty acid metabolism in non-insulin-dependent diabetes mellitus. Evidence for multiple sites of insulin resistance. *The Journal of clinical investigation*, 84(1):205–213, 1989.

[57] Leif Groop and Flemming Pociot. Genetics of diabetes - Are we missing the genes or the disease? *Molecular and cellular endocrinology*, 382:726–39, 2014.

[58] Weihua Guan, Anna Pluzhnikov, Nancy J. Cox, and Michael Boehnke. Meta-analysis of 23 type 2 diabetes linkage studies from the international type 2 diabetes linkage analysis consortium. *Human Heredity*, 66(1):35–49, 2008.

[59] Alexander Gusev, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda W J Penninx, Rick Jansen, Eco J C de Geus, Dorret I Boomsma, Fred A Wright, Patrick F Sullivan, Elina Nikkola, Marcus Alvarez, Mete Civelek, Aldons J Lusis, Terho Lehtimaki, Emma Raitoharju, Mika Kahonen, Ilkka Seppala, Olli Raitakari, Johanna Kuusisto, Markku Laakso, Alkes L Price, Paivi Pajukanta, and Bogdan Pasaniuc. Integrative approaches for large-scale transcriptome-wide association studies. *bioRxiv*, aug 2015.

[60] Alexander Gusev, S. Hong Lee, Gosia Trynka, Hilary Finucane, Bjarni J. Vilhj??lmsson, Han Xu, Chongzhi Zang, Stephan Ripke, Brendan Bulik-Sullivan, Eli Stahl, Anna K. K??hler, Christina M. Hultman, Shaun M. Purcell, Steven A. McCarroll, Mark Daly, Bogdan Pasaniuc, Patrick F. Sullivan, Benjamin M. Neale, Naomi R. Wray, Soumya Raychaudhuri, and Alkes L. Price. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *American Journal of Human Genetics*, 95(5):535–552, 2014.

[61] C L Hanis, E Boerwinkle, R Chakraborty, D L Ellsworth, P Concannon, B Stirling, V A Morrison, B Wapelhorst, R S Spielman, K J Gogolin-Ewens, J M Shephard, S R Williams, N Risch, D Hinds, N Iwasaki, M Ogata, Y Omori, C Petzold, H Rietzsch, H.-E. Schroder, J Schulze, N J Cox, S Menzel, V V Boriraj, X Chen, L R Lim, T Lindner, L E Mereu, Y.-Q. Wang, K Xiang, K Yamagata, Y Yang, and G I Bell. A genome-wide search for human non-insulin-dependent (type 2) diabetes genes reveals a major susceptibility locus on chromosome 2. *Nat Genet*, 13(2):161–166, jun 1996.

[62] R L Hanson, M G Ehm, D J Pettitt, M Prochazka, D B Thompson, D Timberlake, T Foroud, S Kobes, L Baier, D K Burns, L Almasy, J Blangero, W T Garvey, P H Bennett, and W C Knowler. An autosomal genomic scan for loci linked to type II diabetes mellitus and body-mass index in Pima Indians. *American journal of human genetics*, 63(4):1130–1138, 1998.

[63] Jennifer Harrow, Adam Frankish, Jose M. Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L. Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, If Barnes, Alexandra Bignell, Veronika Boychenko, Toby Hunt, Mike Kay, Gaurab Mukherjee, Jeena Rajan, Gloria Despacio-Reyes, Gary Saunders, Charles Steward, Rachel Harte, Michael Lin, C??dric Howald, Andrea Tanzer, Thomas Derrien, Jacqueline Chrast, Nathalie Walters, Suganthi Balasubramanian, Baikang Pei, Michael Tress, Jose Manuel Rodriguez, Iakes Ezkurdia, Jeltje Van Baren, Michael Brent, David Haussler, Manolis Kellis, Alfonso Valencia, Alexandre Reymond, Mark Gerstein, Roderic Guig??, and Tim J. Hubbard. GENCODE: The reference human genome annotation for the ENCODE project. *Genome Research*, 22(9):1760–1774, 2012.

[64] A. T. Hattersley, R. C. Turner, P. Patel, S. O'Rahilly, A. T. Hattersley, P. Patel, J. S. Wainscoat, M. A. Permutt, Y. Tanazawa, K. C. Chin, and P. Watkins. Linkage of type 2 diabetes to the glucokinase gene. *The Lancet*, 339(8805):1307–1310, 1992.

[65] Kari Hemminki, Xinjun Li, Kristina Sundquist, and Jan Sundquist. Familial risks for type 2 diabetes in Sweden. *Diabetes Care*, 33(2):293–297, 2010.

[66] Dena G. Hernandez, Mike A. Nalls, Matthew Moore, Sean Chong, Allissa Dillman, Daniah Trabzuni, J. Raphael Gibbs, Mina Ryten, Sampath Arepalli, Michael E. Weale, Alan B. Zonderman, Juan Troncoso, Richard O'Brien, Robert Walker, Colin Smith, Stefania Bandinelli, Bryan J. Traynor, John Hardy, Andrew B. Singleton, and Mark R. Cookson. Integration of GWAS SNPs and tissue specific expression profiling reveal discrete eQTLs for human traits in blood and brain. *Neurobiology of Disease*, 47(1):20–28, 2012.

[67] Lucia A Hindorff, Praveen Sethupathy, Heather A Junkins, Erin M Ramos, Jayashri P Mehta, Francis S Collins, and Teri A Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367, jun 2009.

[68] Thomas J Hoffmann, Mark N Kvale, Stephanie E Hesselson, Yiping Zhan, Christine Aquino, Yang Cao, Simon Cawley, Elaine Chung, Sheryl Connell, Jasmin Eshragh, Marcia Ewing, Jeremy Gollub, Mary Henderson, Earl Hubbell, Carlos Iribarren, Jay Kaufman, Richard Z Lao, Yontao Lu, Dana Ludwig, Gurpreet K Mathauda, William McGuire, Gangwu Mei, Sunita Miles, Matthew M Purdy, Charles Quesenberry, Dilrini Ranatunga, Sarah Rowell, Marianne Sadler, Michael H Shapero, Ling Shen, Tanushree R Shenoy, David Smethurst, Stephen K Van den Eeden, Larry Walter, Eunice Wan, Reid Wearley, Teresa Webster, Christopher C Wen, Li Weng, Rachel A Whitmer, Alan Williams, Simon C Wong, Chia Zau, Andrea Finn, Catherine Schaefer, Pui-Yan Kwok, and Neil Risch. Next generation genome-wide association tool: Design and coverage of a high-throughput European-optimized SNP array. *Genomics*, 98(2):79–89, aug 2011.

[69] Thomas J Hoffmann, Yiping Zhan, Mark N Kvale, Stephanie E Hesselson, Jeremy Gollub, Carlos Iribarren, Yontao Lu, Gangwu Mei, Matthew M Purdy, Charles Quesenberry, Sarah Rowell, Michael H Shapero, David Smethurst, Carol P Somkin, Stephen K

Van den Eeden, Larry Walter, Teresa Webster, Rachel A Whitmer, Andrea Finn, Catherine Schaefer, Pui-Yan Kwok, and Neil Risch. Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm. *Genomics*, 98(6):422–430, dec 2011.

[70] Y Horikawa, N Oda, N J Cox, X Li, M Orho-Melander, M Hara, Y Hinokio, T H Lindner, H Mashima, P E Schwarz, L del Bosque-Plata, Y Horikawa, Y Oda, I Yoshiuchi, S Colilla, K S Polonsky, S Wei, P Concannon, N Iwasaki, J Schulze, L J Baier, C Bogardus, L Groop, E Boerwinkle, C L Hanis, and G I Bell. Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nature genetics*, 26(2):163–175, 2000.

[71] Momoko Horikoshi, Hanieh Yaghootkar, Dennis O Mook-Kanamori, Ulla Sovio, H Rob Taal, Branwen J Hennig, Jonathan P Bradfield, Beate St Pourcain, David M Evans, Pimphen Charoen, Marika Kaakinen, Diana L Cousminer, Terho Lehtimäki, Eskil Kreiner-Møller, Nicole M Warrington, Mariona Bustamante, Bjarke Feenstra, Diane J Berry, Elisabeth Thiering, Thiemo Pfab, Sheila J Barton, Beverley M Shields, Marjan Kerkhof, Elisabeth M van Leeuwen, Anthony J Fulford, Zoltán Kutalik, Jing Hua Zhao, Marcel den Hoed, Anubha Mahajan, Virpi Lindi, Liang-Kee Goh, Jouke-Jan Hottenga, Ying Wu, Olli T Raitakari, Marie N Harder, Aline Meirhaeghe, Ioanna Ntalla, Rany M Salem, Karen a Jameson, Kaixin Zhou, Dorota M Monies, Vasiliki Lagou, Mirna Kirin, Jani Heikkinen, Linda S Adair, Fowzan S Alkuraya, Ali Al-Odaib, Philippe Amouyel, Ehm Astrid Andersson, Amanda J Bennett, Alexandra I F Blakemore, Jessica L Buxton, Jean Dallongeville, Shikta Das, Eco J C de Geus, Xavier Estivill, Claudia Flexeder, Philippe Froguel, Frank Geller, Keith M Godfrey, Frédéric Gottrand, Christopher J Groves, Torben Hansen, Joel N Hirschhorn, Albert Hofman, Mads V Hollegaard, David M Hougaard, Elina Hyppönen, Hazel M Inskip, Aaron Isaacs, Torben Jørgensen, Christina Kanaka-Gantenbein, John P Kemp, Wieland Kiess, Tuomas O Kilpeläinen, Norman Klopp, Bridget a Knight, Christopher W Kuzawa, George McMahon, John P Newnham, Harri Niinikoski, Ben a Oostra, Louise Pedersen, Dirkje S Postma, Susan M Ring, Fernando Rivadeneira, Neil R Robertson, Sylvain Sebert, Olli Simell, Torsten Slowinski, Carla M T Tiesler, Anke Tönjes, Allan Vaag, Jorma S Viikari, Jacqueline M Vink, Nadja Hawwa Vissing, Nicholas J Wareham, Gonneke Willemsen, Daniel R Witte, Haitao Zhang, Jianhua Zhao, James F Wilson, Michael Stumvoll, Andrew M Prentice, Brian F Meyer, Ewan R Pearson, Colin a G Boreham, Cyrus Cooper, Matthew W Gillman, George V Dedoussis, Luis a Moreno, Oluf Pedersen, Maiju Saarinen, Karen L Mohlke, Dorret I Boomsma, Seang-Mei Saw, Timo a Lakka, Antje Körner, Ruth J F Loos, Ken K Ong, Peter Vollenweider, Cornelia M van Duijn, Gerard H Koppelman, Andrew T Hattersley, John W Holloway, Berthold Hocher, Joachim Heinrich, Chris Power, Mads Melbye, Mònica Guxens, Craig E Pennell, Klaus Bønnelykke, Hans Bisgaard, Johan G Eriksson, Elisabeth Widén, Hakon Hakonarson, André G Uitterlinden, Anneli Pouta, Debbie a Lawlor, George Davey Smith, Timothy M Frayling, Mark I McCarthy, Struan F a Grant, Vincent W V Jaddoe, Marjo-Riitta Jarvelin, Nicholas J Timpson, Inga Prokopenko, and Rachel M Freathy. New

loci associated with birth weight identify genetic links between intrauterine growth and adult height and metabolism. *Nature genetics*, 45(1):76–82, 2012.

[72] B N Howie, P Donnelly, and J Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.*, 5:e1000529, 2009.

[73] Ke Huang, Anup K Nair, Yunhua Li Muller, Paolo Piaggi, Li Bian, Melissa Del Rosario, William C Knowler, Sayuko Kobes, Robert L Hanson, Clifton Bogardus, and Leslie J Baier. Whole exome sequencing identifies variation in CYB5A and RNF10 associated with adiposity and type 2 diabetes. *Obesity (Silver Spring, Md.)*, 2013.

[74] Jim R Hughes, Nigel Roberts, Simon McGowan, Deborah Hay, Eleni Giannoulatou, Magnus Lynch, Marco De Gobbi, Stephen Taylor, Richard Gibbons, and Douglas R Higgs. Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nature genetics*, 46(2):205–12, 2014.

[75] Minako Imamura and Shiro Maeda. Genetics of type 2 diabetes: the GWAS era and future perspectives [Review], 2011.

[76] H Inoue, Y Tanizawa, J Wasson, P Behn, K Kalidas, E Bernal-Mizrachi, M Mueckler, H Marshall, H Donis-Keller, P Crock, D Rogers, M Mikuni, H Kumashiro, K Higashi, G Sobue, Y Oka, and M A Permutt. A gene encoding a transmembrane protein is mutated in patients with diabetes mellitus and optic atrophy (Wolfram syndrome). *Nature genetics*, 20(2):143–8, 1998.

[77] The International and Hapmap Consortium. The International HapMap Project. *Nature*, 426(6968):789–796, 2003.

[78] Beben Benyamin Jian Yang Michael E Goddard, Peter M Visscher, et al. Common SNPs explain a large portion of the heritability for human height. *Nature Genetics*, 42:565–569, 2010.

[79] Dharambir K. Sanghera and Piers R. Blackett. Type 2 Diabetes Genetics: Beyond GWAS, 2012.

[80] Steven E Kahn, Rebecca L Hull, and Kristina M Utzschneider. Mechanisms linking obesity to insulin resistance and type 2 diabetes. *Nature*, 444(7121):840–846, 2006.

[81] J Kaprio, J Tuomilehto, M Koskenvuo, K Romanov, A Reunanen, J Eriksson, J Stengård, and Y A Kesäniemi. Concordance for type 1 (insulin-dependent) and type 2 (non-insulin-dependent) diabetes mellitus in a population-based cohort of twins in Finland. *Diabetologia*, 35:1060–1067, 1992.

[82] Jeffrey M Kidd, Gregory M Cooper, William F Donahue, Hillary S Hayden, Nick Sampas, Tina Graves, Nancy Hansen, Brian Teague, Can Alkan, Francesca Antonacci, Eric Haugen, Troy Zerr, N Alice Yamada, Peter Tsang, Tera L Newman, Eray Tüzün, Ze Cheng, Heather M Ebling, Nadeem Tusneem, Robert David, Will Gillett, Karen a

Phelps, Molly Weaver, David Saranga, Adrianne Brand, Wei Tao, Erik Gustafson, Kevin McKernan, Lin Chen, Maika Malig, Joshua D Smith, Joshua M Korn, Steven a Mccarroll, David a Altshuler, Daniel a Peiffer, Michael Dorschner, John Stamatoyannopoulos, David Schwartz, Deborah a Nickerson, James C Mullikin, Richard K Wilson, Laurakay Bruhn, Maynard V Olson, Rajinder Kaul, Douglas R Smith, and Evan E Eichler. Mapping and sequencing of structural variation from eight human genomes. *Nature*, 453(7191):56–64, 2008.

[83] Augustine Kong, Valgerdur Steinthorsdottir, Gisli Masson, Gudmar Thorleifsson, Patrick Sulem, Soren Besenbacher, Aslaug Jonasdottir, Asgeir Sigurdsson, Kari Th Kristinsson, Adalbjorg Jonasdottir, Michael L Frigge, Arnaldur Gylfason, Pall I Olason, Sigurjon a Gudjonsson, Sverrir Sverrisson, Simon N Stacey, Bardur Sigurgeirsson, Kristrun R Benediktsdottir, Helgi Sigurdsson, Thorvaldur Jonsson, Rafn Benediktsson, Jon H Olafsson, Oskar Th Johannsson, Astradur B Hreidarsson, Gunnar Sigurdsson, Anne C Ferguson-Smith, Daniel F Gudbjartsson, Unnur Thorsteinsdottir, and Kari Stefansson. Parental origin of sequence variants associated with complex diseases. *Nature*, 462(7275):868–874, 2009.

[84] Jaspal S Kooner, Danish Saleheen, Xueling Sim, Joban Sehmi, Weihua Zhang, Philippe Frossard, Latonya F Been, Kee-Seng Chia, Antigone S Dimas, Neelam Hassanali, Tazeen Jafar, Jeremy B M Jowett, Xinzhong Li, Venkatesan Radha, Simon D Rees, Fumihiko Takeuchi, Robin Young, Tin Aung, Abdul Basit, Manickam Chidambaram, Debashish Das, Elin Grundberg, Åsa K Hedman, Zafar I Hydrie, Muhammed Islam, Chiea-Chuen Khor, Sudhir Kowlessur, Malene M Kristensen, Samuel Liju, Wei-Yen Lim, David R Matthews, Jianjun Liu, Andrew P Morris, Alexandra C Nica, Janani M Pinidiyapathirage, Inga Prokopenko, Asif Rasheed, Maria Samuel, Nabi Shah, a Samad Shera, Kerrin S Small, Chen Suo, Ananda R Wickremasinghe, Tien Yin Wong, Mingyu Yang, Fan Zhang, Goncalo R Abecasis, Anthony H Barnett, Mark Caulfield, Panos Deloukas, Timothy M Frayling, Philippe Froguel, Norihiro Kato, Prasad Katulanda, M Ann Kelly, Junbin Liang, Viswanathan Mohan, Dharambir K Sanghera, James Scott, Mark Seielstad, Paul Z Zimmet, Paul Elliott, Yik Ying Teo, Mark I McCarthy, John Danesh, E Shyong Tai, and John C Chambers. Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nature Genetics*, 43(10):984–989, 2011.

[85] K O Kyvik, a Green, and H Beck-Nielsen. Concordance rates of insulin dependent diabetes mellitus: a population based study of young Danish twins. *BMJ (Clinical research ed.)*, 311:913–917, 1995.

[86] Hana Lango Allen, Karol Estrada, Guillaume Lettre, Sonja I Berndt, Michael N Weedon, Fernando Rivadeneira, Cristen J Willer, Anne U Jackson, Sailaja Vedantam, Soumya Raychaudhuri, Teresa Ferreira, Andrew R Wood, Robert J Weyant, Ayellet V Segre, Elizabeth K Speliotes, Eleanor Wheeler, Nicole Soranzo, Ju-Hyun Park, Jian Yang, Daniel Gudbjartsson, Nancy L Heard-Costa, Joshua C Randall, Lu Qi, Albert Vernon Smith, Reedik Magi, Tomi Pastinen, Liming Liang, Iris M Heid, Jian/'an Luan, Gudmar Thorleifsson, Thomas W Winkler, Michael E Goddard, Ken Sin Lo, Cameron

Palmer, Tsegaselassie Workalemahu, Yurii S Aulchenko, Asa Johansson, M Carola Zillikens, Mary F Feitosa, Tonu Esko, Toby Johnson, Shamika Ketkar, Peter Kraft, Massimo Mangino, Inga Prokopenko, Devin Absher, Eva Albrecht, Florian Ernst, Nicole L Glazer, Caroline Hayward, Jouke-Jan Hottenga, Kevin B Jacobs, Joshua W Knowles, Zoltan Kutalik, Keri L Monda, Ozren Polasek, Michael Preuss, Nigel W Rayner, Neil R Robertson, Valgerdur Steinthorsdottir, Jonathan P Tyrer, Benjamin F Voight, Fredrik Wiklund, Jianfeng Xu, Jing Hua Zhao, Dale R Nyholt, Niina Pellikka, Markus Perola, John R B Perry, Ida Surakka, Mari-Liis Tammesoo, Elizabeth L Altmaier, Najaf Amin, Thor Aspelund, Tushar Bhangale, Gabrielle Boucher, Daniel I Chasman, Constance Chen, Lachlan Coin, Matthew N Cooper, Anna L Dixon, Quince Gibson, Elin Grundberg, Ke Hao, M Juhani Junttila, Lee M Kaplan, Johannes Kettunen, Inke R Konig, Tony Kwan, Robert W Lawrence, Douglas F Levinson, Mattias Lorentzon, Barbara McKnight, Andrew P Morris, Martina Muller, Julius Suh Ngwa, Shaun Purcell, Suzanne Rafelt, Rany M Salem, Erika Salvi, Serena Sanna, Jianxin Shi, Ulla Sovio, John R Thompson, Michael C Turchin, Liesbeth Vandenput, Dominique J Verlaan, Veronique Vitart, Charles C White, Andreas Ziegler, Peter Almgren, Anthony J Balmforth, Harry Campbell, Lorena Citterio, Alessandro De Grandi, Anna Dominiczak, Jubao Duan, Paul Elliott, Roberto Elosua, Johan G Eriksson, Nelson B Freimer, Eco J C Geus, Nicola Glorioso, Shen Haiqing, Anna-Liisa Hartikainen, Aki S Havulinna, Andrew A Hicks, Jennie Hui, Wilmar Igl, Thomas Illig, Antti Jula, Eero Kajantie, Tuomas O Kilpelainen, Markku Koiranen, Ivana Kolcic, Seppo Koskinen, Peter Kovacs, Jaana Laitinen, Jianjun Liu, Marja-Liisa Lokki, Ana Marusic, Andrea Maschio, Thomas Meitinger, Antonella Mulas, Guillaume Pare, Alex N Parker, John F Peden, Astrid Petersmann, Irene Pichler, Kirsi H Pietilainen, Anneli Pouta, Martin Ridderstrale, Jerome I Rotter, Jennifer G Sambrook, Alan R Sanders, Carsten Oliver Schmidt, Juha Sinisalo, Jan H Smit, Heather M Stringham, G Bragi Walters, Elisabeth Widen, Sarah H Wild, Gonneke Willemsen, Laura Zagato, Lina Zgaga, Paavo Zitting, Helene Alavere, Martin Farrall, Wendy L McArdle, Mari Nelis, Marjolein J Peters, Samuli Ripatti, Joyce B J van Meurs, Katja K Aben, Kristin G Ardlie, Jacques S Beckmann, John P Beilby, Richard N Bergman, Sven Bergmann, Francis S Collins, Daniele Cusi, Martin den Heijer, Gudny Eiriksdottir, Pablo V Gejman, Alistair S Hall, Anders Hamsten, Heikki V Huikuri, Carlos Iribarren, Mika Kahonen, Jaakko Kaprio, Sekar Kathiresan, Lambertus Kiemeney, Thomas Kocher, Lenore J Launer, Terho Lehtimaki, Olle Melander, Tom H Mosley Jr, Arthur W Musk, Markku S Nieminen, Christopher J O/'Donnell, Claes Ohlsson, Ben Oostra, Lyle J Palmer, Olli Raitakari, Paul M Ridker, John D Rioux, Aila Rissanen, Carlo Rivolta, Heribert Schunkert, Alan R Shuldiner, David S Siscovick, Michael Stumvoll, Anke Tonjes, Jaakko Tuomilehto, Gert-Jan van Ommen, Jorma Viikari, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, Michael A Province, Manfred Kayser, Alice M Arnold, Larry D Atwood, Eric Boerwinkle, Stephen J Chanock, Panos Deloukas, Christian Gieger, Henrik Gronberg, Per Hall, Andrew T Hattersley, Christian Hengstenberg, Wolfgang Hoffman, G Mark Lathrop, Veikko Salomaa, Stefan Schreiber, Manuela Uda, Dawn Waterworth, Alan F Wright, Themistocles L Assimes, Ines Barroso, Albert Hofman, Karen L Mohlke, Dorret I Boomsma, Mark J Caulfield, L Adrienne Cupples, Jeanette Erdmann, Caroline S Fox, Vilmundur Gudnason, Ulf Gyllensten, Tamara B Harris, Richard B

Hayes, Marjo-Riitta Jarvelin, Vincent Mooser, Patricia B Munroe, Willem H Ouwehand, Brenda W Penninx, Peter P Pramstaller, Thomas Quertermous, Igor Rudan, Nilesh J Samani, Timothy D Spector, Henry Volzke, Hugh Watkins, James F Wilson, Leif C Groop, Talin Haritunians, Frank B Hu, Robert C Kaplan, Andres Metspalu, Kari E North, David Schlessinger, Nicholas J Wareham, David J Hunter, Jeffrey R O/'Connell, David P Strachan, H.-Erich Wichmann, Ingrid B Borecki, Cornelia M van Duijn, Eric E Schadt, Unnur Thorsteinsdottir, Leena Peltonen, Andre G Uitterlinden, Peter M Visscher, Nilanjan Chatterjee, Ruth J F Loos, Michael Boehnke, Mark I McCarthy, Erik Ingelsson, Cecilia M Lindgren, Goncalo R Abecasis, Kari Stefansson, Timothy M Frayling, and Joel N Hirschhorn. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317):832–838, oct 2010.

[87] S Hong Lee, Teresa R Decandia, Stephan Ripke, Jian Yang, Patrick F Sullivan, Michael E Goddard, Matthew C Keller, Peter M Visscher, and Naomi R Wray. Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nature Genetics*, 44:247–250, 2012.

[88] Sang Hong Lee, Naomi R Wray, Michael E Goddard, and Peter M Visscher. Estimating Missing Heritability for Disease from Genome-wide Association Studies. *The American Journal of Human Genetics*, 88:294–305, 2011.

[89] Ray K M Leung and Paul A Whittaker. RNA interference: From gene silencing to gene-specific therapeutics. *Pharmacology & Therapeutics*, 107(2):222–239, aug 2005.

[90] S. Lillioja and A. Wilton. Agreement among type 2 diabetes linkage studies but a poor correlation with results from genome-wide association studies. *Diabetologia*, 52(6):1061–1074, 2009.

[91] Kirk E Lohmueller, Thomas Sparsø, Qibin Li, Ehm Andersson, Thorfinn Korneliussen, Anders Albrechtsen, Karina Banasik, Niels Grarup, Ingileif Hallgrimsdottir, Kristoffer Kiil, Tuomas O Kilpeläinen, Nikolaj T Krarup, Tune H Pers, Gaston Sanchez, Youna Hu, Michael Degiorgio, Torben Jørgensen, Annelli Sandbæk, Torsten Lauritzen, Søren Brunak, Karsten Kristiansen, Yingrui Li, Torben Hansen, Jun Wang, Rasmus Nielsen, and Oluf Pedersen. Whole-exome sequencing of 2,000 danish individuals and the role of rare coding variants in type 2 diabetes. *American journal of human genetics*, 93:1072–86, 2013.

[92] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, Barbara Foster, Mike Moser, Ellen Karasik, Bryan Gillard, Kimberley Ramsey, Susan Sullivan, Jason Bridge, Harold Magazine, John Syron, Johnelle Fleming, Laura Siminoff, Heather Traino, Maghboeba Mosavel, Laura Barker, Scott Jewell, Dan Rohrer, Dan Maxim, Dana Filkins, Philip Harbach, Eddie Cortadillo, Bree Berghuis, Lisa Turner, Eric Hudson, Kristin Feenstra, Leslie Sobin, James Robb, Phillip Branton, Greg Korzeniewski, Charles Shive, David Tabor, Liqun Qi, Kevin Groch, Sreenath Nampally, Steve Buia, Angela Zimmerman, Anna Smith, Robin Burges, Karna Robinson, Kim Valentino,

Deborah Bradbury, Mark Cosentino, Norma Diaz-Mayoral, Mary Kennedy, Theresa Engel, Penelope Williams, Kenyon Erickson, Kristin Ardlie, Wendy Winckler, Gad Getz, David DeLuca, Daniel MacArthur, Manolis Kellis, Alexander Thomson, Taylor Young, Ellen Gelfand, Molly Donovan, Yan Meng, George Grant, Deborah Mash, Yvonne Marcus, Margaret Basile, Jun Liu, Jun Zhu, Zhidong Tu, Nancy J Cox, Dan L Nicolae, Eric R Gamazon, Hae Kyung Im, Anuar Konkashbaev, Jonathan Pritchard, Matthew Stevens, Timothee Flutre, Xiaoquan Wen, Emmanouil T Dermitzakis, Tuuli Lappalainen, Roderic Guigo, Jean Monlong, Michael Sammeth, Daphne Koller, Alexis Battle, Sara Mostafavi, Mark McCarthy, Manual Rivas, Julian Maller, Ivan Rusyn, Andrew Nobel, Fred Wright, Andrey Shabalin, Mike Feolo, Nataliya Sharopova, Anne Sturcke, Justin Paschal, James M Anderson, Elizabeth L Wilder, Leslie K Derr, Eric D Green, Jeffery P Struewing, Gary Temple, Simona Volpi, Joy T Boyer, Elizabeth J Thomson, Mark S Guyer, Cathy Ng, Assya Abdallah, Deborah Colantuoni, Thomas R Insel, Susan E Koester, A Roger Little, Patrick K Bender, Thomas Lehner, Yin Yao, Carolyn C Compton, Jimmie B Vaught, Sherilyn Sawyer, Nicole C Lockhart, Joanne Demchok, Helen F Moore, and The Gtex Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*, 45(6):580–585, jun 2013.

[93] Ruth J F Loos, Cecilia M Lindgren, Shengxu Li, Eleanor Wheeler, Jing Hua Zhao, Inga Prokopenko, Michael Inouye, Rachel M Freathy, Antony P Attwood, Jacques S Beckmann, Sonja I Berndt, Kevin B Jacobs, Stephen J Chanock, Richard B Hayes, Sven Bergmann, Amanda J Bennett, Sheila A Bingham, Murielle Bochud, Morris Brown, Stéphane Cauchi, John M Connell, Cyrus Cooper, George Davey Smith, Ian Day, Christian Dina, Subhajyoti De, Emmanouil T Dermitzakis, Alex S F Doney, Katherine S Elliott, Paul Elliott, David M Evans, I Sadaf Farooqi, Philippe Froguel, Jilur Ghori, Christopher J Groves, Rhian Gwilliam, David Hadley, Alistair S Hall, Andrew T Hattersley, Johannes Hebebrand, Iris M Heid, Claudia Lamina, Christian Gieger, Thomas Illig, Thomas Meitinger, H-Erich Wichmann, Blanca Herrera, Anke Hinney, Sarah E Hunt, Marjo-Riitta Jarvelin, Toby Johnson, Jennifer D M Jolley, Fredrik Karpe, Andrew Keniry, Kay-Tee Khaw, Robert N Luben, Massimo Mangino, Jonathan Marchini, Wendy L McArdle, Ralph McGinnis, David Meyre, Patricia B Munroe, Andrew D Morris, Andrew R Ness, Matthew J Neville, Alexandra C Nica, Ken K Ong, Stephen O'Rahilly, Katharine R Owen, Colin N A Palmer, Konstantinos Papadakis, Simon Potter, Anneli Pouta, Lu Qi, Joshua C Randall, Nigel W Rayner, Susan M Ring, Manjinder S Sandhu, André Scherag, Matthew A Sims, Kijoung Song, Nicole Soranzo, Elizabeth K Speliotes, Holly E Syddall, Sarah A Teichmann, Nicholas J Timpson, Jonathan H Tobias, Manuela Uda, Carla I Ganz Vogel, Chris Wallace, Dawn M Waterworth, Michael N Weedon, Cristen J Willer, Wraight, Xin Yuan, Eleftheria Zeggini, Joel N Hirschhorn, David P Strachan, Willem H Ouwehand, Mark J Caulfield, Nilesh J Samani, Timothy M Frayling, Peter Vollenweider, Gerard Waeber, Vincent Mooser, Panos Deloukas, Mark I McCarthy, Nicholas J Wareham, Inês Barroso, Peter Kraft, Susan E Hankinson, David J Hunter, Frank B Hu, Helen N Lyon, Benjamin F Voight, Martin Ridderstrale, Leif Groop, Paul Scheet, Serena Sanna, Goncalo R Abecasis, Giuseppe Albai, Ramaiah Nagaraja, David Schlessinger, Anne U Jackson, Jaakko Tuomilehto, Francis S Collins, Michael Boehnke, and Karen L

Mohlke. Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nature genetics*, 40(6):768–75, 2008.

[94] Valeriya Lyssenko, Peter Almgren, Dragi Anevski, Roland Perfekt, Kaj Lahti, Michael Nissén, Bo Isomaa, Björn Forsen, Nils Homström, Carola Saloranta, Marja-Riitta Taskinen, Leif Groop, and Tiinamaija Tuomi. Predictors of and longitudinal changes in insulin sensitivity and secretion preceding onset of type 2 diabetes. *Diabetes*, 54:166–174, 2005.

[95] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, Judy H Cho, Alan E Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N Rotimi, Montgomery Slatkin, David Valle, Alice S Whittemore, Michael Boehnke, Andrew G Clark, Evan E Eichler, Greg Gibson, Jonathan L Haines, Trudy F C Mackay, Steven A McCarroll, and Peter M Visscher. Finding the missing heritability of complex diseases. *Nature*, 461:747–53, 2009.

[96] Jonathan Marchini, Bryan Howie, Simon Myers, Gil McVean, and Peter Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics*, 39(7):906–13, 2007.

[97] Alicia R Martin, Helio A Costa, Tuuli Lappalainen, Brenna M Henn, Jeffrey M Kidd, Muh-Ching Yee, Fabian Grubert, Howard M Cann, Michael Snyder, Stephen B Montgomery, and Carlos D Bustamante. Transcriptome sequencing from diverse human populations reveals differentiated regulatory architecture. *PLoS genetics*, 10(8):e1004549, 2014.

[98] Masafumi Matsuda, Ralph A. DeFronzo, Leonard Glass, Agostino Consoli, Mauro Giordano, Peter Bressler, and Stefano DelPrato. Glucagon dose-response curve for hepatic glucose production and glucose disposal in type 2 diabetic patients and normal individuals. *Metabolism*, 51(9):1111–1119, 2002.

[99] Matthew T Maurano, Richard Humbert, Eric Rynes, Robert E Thurman, Eric Haugen, Hao Wang, Alex P Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, Anthony Shafer, Fidencio Neri, Kristen Lee, Tanya Kutyavin, Sandra Stehling-Sun, Audra K Johnson, Theresa K Canfield, Erika Giste, Morgan Diegel, Daniel Bates, R Scott Hansen, Shane Neph, Peter J Sabo, Shelly Heimfeld, Antony Raubitschek, Steven Ziegler, Chris Cotsapas, Nona Sotoodehnia, Ian Glass, Shamil R Sunyaev, Rajinder Kaul, and John A Stamatoyannopoulos. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science*, 337(6099):1190–1195, 2012.

[100] Mark I McCarthy, Gonçalo R Abecasis, Lon R Cardon, David B Goldstein, Julian Little, John P a Ioannidis, and Joel N Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews. Genetics*, 9(5):356–369, 2008.

[101] J B Meigs, L a Cupples, and P W Wilson. Parental transmission of type 2 diabetes: the Framingham Offspring Study. *Diabetes*, 49(12):2201–7, 2000.

[102] Boyd E. Metzger, Lynn P. Lowe, Alan R. Dyer, Elisabeth R. Trimble, Brian Sheridan, Moshe Hod, Rony Chen, Yariv Yogev, Donald R. Coustan, Patrick M. Catalano, Warwick Giles, Julia Lowe, David R. Hadden, Bengt Persson, and Jeremy J N Oats. Hyperglycemia and adverse pregnancy outcome (HAPO) study: Associations with neonatal anthropometrics. *Diabetes*, 58(2):453–459, 2009.

[103] Sanghoon Moon, Kwang Su Jung, Young Jin Kim, Mi Yeong Hwang, Kyungsook Han, Jong Young Lee, Kiejung Park, and Bong Jo Kim. KGVDB: A population-based genomic map of CNVs tagged by SNPs in Koreans. *Bioinformatics*, 29(11):1481–1483, 2013.

[104] Andrew D P Morris, Benjamin F Voight, Tanya M Teslovich, Teresa Ferreira, A V Segre, Valgerdur Steinthorsdottir, Rona J Strawbridge, Hassan Khan, Harald Grallert, Anubha Mahajan, Inga Prokopenko, Hyun Min Kang, Christian Dina, Tonu Esko, Ross M Fraser, Stavroula Kanoni, Ashish Kumar, Vasiliki Lagou, Claudia Langenberg, Jian'an Luan, Cecilia M Lindgren, M Muller-Nurasyid, Sonali Pechlivanis, N William Rayner, Laura J Scott, Steven Wiltshire, Loic Yengo, Leena Kinnunen, Elizabeth J Rossin, Soumya Raychaudhuri, Andrew D Johnson, Antigone S Dimas, Ruth J F Loos, Sailaja Vedantam, Han Chen, Jose C Florez, Caroline Fox, Ching-Ti T Liu, Denis Rybin, David J Couper, Wen Hong L Kao, Man Li, Marilyn C Cornelis, Peter Kraft, Qi Sun, Rob M van Dam, Heather M Stringham, Peter S Chines, Krista Fischer, Pierre Fontanillas, Oddgeir L Holmen, Sarah E Hunt, Anne U Jackson, Augustine Kong, Robert Lawrence, Julia Meyer, John R B Perry, Carl G P Platou, Simon Potter, Emil Rehnberg, Neil Robertson, Suthesh Sivapalaratnam, A Stancakova, Kathleen Stirrups, Gudmar Thorleifsson, Emmi Tikkanen, Andrew R Wood, Peter Almgren, Mustafa Atalay, Rafn Benediktsson, Lori L Bonnycastle, Noël Burtt, Jason Carey, Guillaume Charpentier, Andrew T Crenshaw, Alex S F Doney, Mozhgan Dorkhan, Sarah Edkins, Valur Emilsson, Elodie Eury, Tom Forsen, Karl Gertow, Bruna Gigante, George B Grant, Christopher J Groves, Candace Guiducci, Christian Herder, Astradur B Hreidarsson, Jennie Hui, Alan James, Anna Jonsson, Wolfgang Rathmann, Norman Klopp, Jasmina Kravic, K Krjutskov, Cordelia Langford, Karin Leander, Eero Lindholm, Stéphane Lobbens, S Mannisto, Ghazala Mirza, T W Muhleisen, Bill Musk, Melissa Parkin, Loukianos Rallidis, Jouko Saramies, Bengt Sennblad, Sonia Shah, G Sigurethsson, Angela Silveira, Gerald Steinbach, Barbara Thorand, Joseph Trakalo, Fabrizio Veglia, Roman Wennauer, Wendy Winckler, Delilah Zabaneh, Harry Campbell, Cornelia van Duijn, Andre G Uitterlinden, Albert Hofman, Eric Sijbrands, Goncalo R Abecasis, Katharine R Owen, Eleftheria Zeggini, Mieke D Trip, Nita G Forouhi, A C Syvanen, Johan G Eriksson, Leena Peltonen, M M Nothen, Beverley Balkau, Colin N A Palmer, Valeriya Lyssenko, Tiinamaija Tuomi, Bo Isomaa, David J Hunter, Lu Qi, Consortium Wellcome Trust Case Control, Glucose Meta-Analyses of, Investigators Insulin-related traits Consortium, ANthropometric Traits Consortium Genetic Investigation of, Consortium Asian Genetic Epidemiology Network-Type 2 Diabetes, Consortium South Asian Type 2 Diabetes, Alan R Shuldiner, Michael Roden, Ines Barroso, Tom Wilsgaard, John Beilby, Kees Hovingh, Jackie F Price, James F Wilson, Rainer Rauramaa, Timo A Lakka, Lars Lind, George Dedoussis, I Njolstad, Nancy L Pedersen,

Kay-Tee T Khaw, Nicholas J Wareham, Sirkka M Keinanen-Kiukaanniemi, Timo E Saaristo, E Korpi-Hyovalti, Juha Saltevo, Markku Laakso, Johanna Kuusisto, Andres Metspalu, Francis S Collins, Karen L Mohlke, Richard N Bergman, Jaakko Tuomilehto, Bernhard O Boehm, Christian Gieger, Kristian Hveem, Stephane Cauchi, Philippe Froguel, Damiano Baldassarre, Elena Tremoli, Steve E Humphries, Danish Saleheen, John Danesh, Erik Ingelsson, Samuli Ripatti, Veikko Salomaa, Raimund Erbel, K H Jockel, Susanne Moebus, Annette Peters, Thomas Illig, Ulf de Faire, Anders Hamsten, Andrew D P Morris, Peter J Donnelly, Timothy M Frayling, Andrew T Hattersley, Eric Boerwinkle, Olle Melander, Sekar Kathiresan, Peter M Nilsson, Panos Deloukas, Unnur Thorsteinsdottir, Leif C Groop, Kari Stefansson, Frank Hu, James S Pankow, Josée Dupuis, James B Meigs, David Altshuler, Michael Boehnke, Mark I McCarthy, D IAbetes Genetics Replication, Consortium Meta-analysis, Ayellet V Segrè, Valgerdur Steinthorsdottir, Rona J Strawbridge, Hassan Khan, Harald Grallert, Anubha Mahajan, Inga Prokopenko, Hyun Min Kang, Christian Dina, Tonu Esko, Ross M Fraser, Stavroula Kanoni, Ashish Kumar, Vasiliki Lagou, Claudia Langenberg, Jian'an Luan, Cecilia M Lindgren, Martina Müller-Nurasyid, Sonali Pechlivanis, N William Rayner, Laura J Scott, Steven Wiltshire, Loic Yengo, Leena Kinnunen, Elizabeth J Rossin, Soumya Raychaudhuri, Andrew D Johnson, Antigone S Dimas, Ruth J F Loos, Sailaja Vedantam, Han Chen, Jose C Florez, Caroline Fox, Ching-Ti T Liu, Denis Rybin, David J Couper, Wen Hong L Kao, Man Li, Marilyn C Cornelis, Peter Kraft, Qi Sun, Rob M van Dam, Heather M Stringham, Peter S Chines, Krista Fischer, Pierre Fontanillas, Oddgeir L Holmen, Sarah E Hunt, Anne U Jackson, Augustine Kong, Robert Lawrence, Julia Meyer, John R B Perry, Carl G P Platou, Simon Potter, Emil Rehnberg, Neil Robertson, Suthesh Sivapalaratnam, Alena Stančáková, Kathleen Stirrups, Gudmar Thorleifsson, Emmi Tikkanen, Andrew R Wood, Peter Almgren, Mustafa Atalay, Rafn Benediktsson, Lori L Bonnycastle, Noël Burtt, Jason Carey, Guillaume Charpentier, Andrew T Crenshaw, Alex S F Doney, Mozhgan Dorkhan, Sarah Edkins, Valur Emilsson, Elodie Eury, Tom Forsen, Karl Gertow, Bruna Gigante, George B Grant, Christopher J Groves, Candace Guiducci, Christian Herder, Astradur B Hreidarsson, Jennie Hui, Alan James, Anna Jonsson, Wolfgang Rathmann, Norman Klopp, Jasmina Kravic, Kaarel Krjutškov, Cordelia Langford, Karin Leander, Eero Lindholm, Stéphane Lobbens, Satu Männistö, Ghazala Mirza, Thomas W Mühleisen, Bill Musk, Melissa Parkin, Loukianos Rallidis, Jouko Saramies, Bengt Sennblad, Sonia Shah, Gunnar Sigursson, Angela Silveira, Gerald Steinbach, Barbara Thorand, Joseph Trakalo, Fabrizio Veglia, Roman Wennauer, Wendy Winckler, Delilah Zabaneh, Harry Campbell, Cornelia van Duijn, Andre G Uitterlinden, Albert Hofman, Eric Sijbrands, Goncalo R Abecasis, Katharine R Owen, Eleftheria Zeggini, Mieke D Trip, Nita G Forouhi, Ann-Christine Syvänen, Johan G Eriksson, Leena Peltonen, Markus M Nöthen, Beverley Balkau, Colin N A Palmer, Valeriya Lyssenko, Tiinamaija Tuomi, Bo Isomaa, David J Hunter, Lu Qi, Alan R Shuldiner, Michael Roden, Ines Barroso, Tom Wilsgaard, John Beilby, Kees Hovingh, Jackie F Price, James F Wilson, Rainer Rauramaa, Timo A Lakka, Lars Lind, George Dedoussis, Inger Njølstad, Nancy L Pedersen, Kay-Tee T Khaw, Nicholas J Wareham, Sirkka M Keinanen-Kiukaanniemi, Timo E Saaristo, Eeva Korpi-Hyövälti, Juha Saltevo, Markku Laakso, Johanna Kuusisto, Andres Metspalu, Francis S Collins, Karen L Mohlke,

Richard N Bergman, Jaakko Tuomilehto, Bernhard O Boehm, Christian Gieger, Kristian Hveem, Stephane Cauchi, Philippe Froguel, Damiano Baldassarre, Elena Tremoli, Steve E Humphries, Danish Saleheen, John Danesh, Erik Ingelsson, Samuli Ripatti, Veikko Salomaa, Raimund Erbel, Karl-Heinz Jöckel, Susanne Moebus, Annette Peters, Thomas Illig, Ulf de Faire, Anders Hamsten, Andrew D P Morris, Peter J Donnelly, Timothy M Frayling, Andrew T Hattersley, Eric Boerwinkle, Olle Melander, Sekar Kathiresan, Peter M Nilsson, Panos Deloukas, Unnur Thorsteinsdottir, Leif C Groop, Kari Stefansson, Frank Hu, James S Pankow, Josée Dupuis, James B Meigs, David Altshuler, Michael Boehnke, and Mark I McCarthy. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics*, 44(9):981–990, 2012.

[105] Kiran Musunuru, Alanna Strong, Maria Frank-Kamenetsky, Noemi E Lee, Tim Ahfeldt, Katherine V Sachs, Xiaoyu Li, Hui Li, Nicolas Kuperwasser, Vera M Ruda, James P Pirruccello, Brian Muchmore, Ludmila Prokunina-Olsson, Jennifer L Hall, Eric E Schadt, Carlos R Morales, Sissel Lund-Katz, Michael C Phillips, Jamie Wong, William Cantley, Timothy Racie, Kenechi G Ejebe, Marju Orho-Melander, Olle Melander, Victor Koteliansky, Kevin Fitzgerald, Ronald M Krauss, Chad A Cowan, Sekar Kathiresan, and Daniel J Rader. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature*, 466(7307):714–719, aug 2010.

[106] J V Neel. Diabetes Mellitus A Geneticists Nightmare BT - The Genetics of Diabetes Mellitus. pages 1–11. Springer Berlin Heidelberg, Berlin, Heidelberg, 1976.

[107] J. V. Neel. Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? 1962. *Bulletin of the World Health Organization*, 77(8), 1999.

[108] B. Newman, J. V. Selby, M. C. King, C. Slemenda, R. Fabsitz, and G. D. Friedman. Concordance for type 2 (non-insulin-dependent) diabetes mellitus in male twins. *Diabetologia*, 30(10):763–768, 1987.

[109] Philip Newsholme, Deirdre Keane, Hannah J Welters, and Noel G Morgan. Life and death decisions of the pancreatic beta-cell: the role of fatty acids. *Clinical science (London, England : 1979)*, 112(1):27–42, 2007.

[110] Alexandra C. Nica, Stephen B. Montgomery, Antigone S. Dimas, Barbara E. Stranger, Claude Beazley, Inês In??s Barroso, and Emmanouil T. Dermitzakis. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS genetics*, 6(4):e1000895, 2010.

[111] Dan L. Nicolae, Eric Gamazon, Wei Zhang, Shiwei Duan, M. Eileen Dolan, Nancy J. Cox, M Eileen Dolan, and Nancy J. Cox. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS genetics*, 6(4):e1000888, 2010.

[112] Silvana Obici, Zhaohui Feng, George Karkanias, Denis G Baskin, and Luciano Rossetti. Decreasing hypothalamic insulin receptors causes hyperphagia and insulin resistance in rats. *Nature neuroscience*, 5(6):566–572, 2002.

[113] Marju Orho-Melander, Mia Klannemark, Malin K Svensson, Martin Ridderstråle, Cecilia M Lindgren, and Leif Groop. Variants in the Calpain-10 Gene Predispose to Insulin Resistance and Elevated Free Fatty Acid Levels. *Diabetes*, 51(8):2658–2664, aug 2002.

[114] Jurg Ott, Jing Wang, and Suzanne M Leal. Genetic linkage analysis in the age of whole-genome sequencing. *Nature Reviews Genetics*, 16(5):275–284, 2015.

[115] Ju-Hyun Park, Sholom Wacholder, Mitchell H Gail, Ulrike Peters, Kevin B Jacobs, Stephen J Chanock, and Nilanjan Chatterjee. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature genetics*, 42:570–575, 2010.

[116] Stephen C J Parker, Michael L Stitzel, D Leland Taylor, Jose Miguel Orozco, Michael R Erdos, Jennifer A Akiyama, Kelly Lammerts van Bueren, Peter S Chines, Narisu Narisu, Brian L Black, Axel Visel, Len A Pennacchio, and Francis S Collins. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proceedings of the National Academy of Sciences of the United States of America*, 110(44):17921–17926, 2013.

[117] E J Parra, J E Below, S Krithika, A Valladares, J L Barta, N J Cox, C L Hanis, N Wacher, J Garcia-Mena, P Hu, M D Shriver, J Kumate, P M McKeigue, J Escobedo, and M Cruz. Genome-wide association study of type 2 diabetes in a sample from Mexico City and a meta-analysis of a Mexican-American sample from Starr County, Texas. *Diabetologia*, 54:2038–2046, 2011.

[118] Lorenzo Pasquali, Kyle J Gaulton, Santiago A Rodríguez-Seguí, Loris Mularoni, Irene Miguel-Escalada, Ildem Akerman, Juan J Tena, Ignasi Morán, Carlos Gómez-Marín, Martijn van de Bunt, Joan Ponsa-Cobas, Natalia Castro, Takao Nammo, Inês Cebola, Javier García-Hurtado, Miguel Angel Maestro, François Pattou, Lorenzo Piemonti, Thierry Berney, Anna L Gloyn, Philippe Ravassard, José Luis Gómez-Skarmeta, Ferenc Müller, Mark I McCarthy, and Jorge Ferrer. Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nature genetics*, 46(2):136–43, 2014.

[119] Radhika Patnala, Judith Clements, and Jyotsna Batra. Candidate gene association studies: a comprehensive guide to useful in silico tools. *BMC genetics*, 14:39, 2013.

[120] John R B Perry, Benjamin F. Voight, Loïc Yengo, Najaf Amin, Josée Dupuis, Martha Ganser, Harald Grallert, Pau Navarro, Man Li, Lu Qi, Valgerdur Steinthorsdottir, Robert A. Scott, Peter Almgren, Dan E. Arking, Yurii Aulchenko, Beverley Balkau, Rafn Benediktsson, Richard N. Bergman, Eric Boerwinkle, Lori Bonnycastle, Noël P. Burtt, Harry Campbell, Guillaume Charpentier, Francis S. Collins, Christian Gieger, Todd Green, Samy Hadjadj, Andrew T. Hattersley, Christian Herder, Albert Hofman, Andrew D. Johnson, Anna Kottgen, Peter Kraft, Yann Labrune, Claudia Langenberg, Alisa K. Manning, Karen L. Mohlke, Andrew P. Morris, Ben Oostra, James Pankow, Ann Kristin Petersen, Peter P. Pramstaller, Inga Prokopenko, Wolfgang Rathmann,

William Rayner, Michael Roden, Igor Rudan, Denis Rybin, Laura J. Scott, Gunnar Sigurdsson, Rob Sladek, Gudmar Thorleifsson, Unnur Thorsteinsdottir, Jaakko Tuomilehto, Andre G. Uitterlinden, Sidonie Vivequin, Michael N. Weedon, Alan F. Wright, Frank B. Hu, Thomas Illig, Linda Kao, James B. Meigs, James F. Wilson, Kari Stefansson, Cornelia van Duijn, David Altschuler, Andrew D. Morris, Michael Boehnke, Mark I. McCarthy, Philippe Froguel, Colin N A Palmer, Nicholas J. Wareham, Leif Groop, Timothy M. Frayling, and Stéphane Cauchi. Stratifying type 2 diabetes cases by BMI identifies genetic risk variants in LAMA1 and enrichment for risk variants in lean compared to obese cases. *PLoS Genetics*, 8, 2012.

[121] P Poulsen, K O Kyvik, A Vaag, and H Beck-Nielsen. Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance–a population-based twin study. *Diabetologia*, 42:139–145, 1999.

[122] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38:904–909, 2006.

[123] Shaun M Purcell, Naomi R Wray, Jennifer L Stone, Peter M Visscher, Michael C O'Donovan, Patrick F Sullivan, and Pamela Sklar. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460:748–752, 2009.

[124] R R Development Core Team. *R: A Language and Environment for Statistical Computing*, volume 1. 2011.

[125] Diabetes Genetics Replication, Meta-analysis Diagram Consortium, and Asian Genetic Epidemiology. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature genetics*, 46(3):234–44, 2014.

[126] Inga Reynisdottir, Gudmar Thorleifsson, Rafn Benediktsson, Gunnar Sigurdsson, Valur Emilsson, Anna Sigurlin Einarsdottir, Eyrun Edda Hjorleifsdottir, Gudbjorg Th Orlygsdottir, Gudrun Thora Bjornsdottir, Jona Saemundsdottir, Skarphedinn Halldorsson, Soffia Hrafnkelsdottir, Steinunn Bjorg Sigurjonsdottir, Svana Steinsdottir, Mitchell Martin, Jarema P Kochan, Brian K Rhees, Struan F A Grant, Michael L Frigge, Augustine Kong, Vilmundur Gudnason, Kari Stefansson, and Jeffrey R Gulcher. Localization of a susceptibility gene for type 2 diabetes to chromosome 5q34-q35.2. *American journal of human genetics*, 73:323–335, 2003.

[127] Martin Ridderstråle and Leif Groop. Genetic dissection of type 2 diabetes. *Molecular and cellular endocrinology*, 297:10–17, 2009.

[128] M Roden, T B Price, G Perseghin, K F Petersen, D L Rothman, G W Cline, and G I Shulman. Mechanism of free fatty acid-induced insulin resistance in humans. *The Journal of clinical investigation*, 97:2859–2865, 1996.

[129] Gojka Roglic and Nigel Unwin. Mortality attributable to diabetes: Estimates for the year 2010, 2010.

[130] Jørn V Sagen, Helge Ræder, Eba Hathout, Naim Shehadeh, Kolbeinn Gudmundsson, Halvor Bævre, Dianne Abuelo, Chanika Phornphutkul, Janne Molnes, Graeme I Bell, Anna L Gloyn, Andrew T Hattersley, Anders Molven, Oddmund Søvik, and Pål R Njølstad. Permanent Neonatal Diabetes due to Mutations in KCNJ11 Encoding Kir6.2: Patient Characteristics and Initial Response to Sulfonylurea Therapy . *Diabetes*, 53(10):2713–2718, oct 2004.

[131] L. B. Salans, G. A. Bray, S. W. Cushman, E. Danforth, J. A. Glennon, E. S. Horton, and E. A. Sims. Glucose metabolism and the response to insulin by human adipose tissue in spontaneous and experimental obesity. Effects of dietary composition and adipose cell size. *Journal of Clinical Investigation*, 53(3):848–856, 1974.

[132] Daniel Savic, Honggang Ye, Ivy Aneas, Soo Young Park, Graeme I. Bell, and Marcelo A. Nobrega. Alterations in TCF7L2 expression define its role as a key regulator of glucose metabolism. *Genome Research*, 21(9):1417–1425, 2011.

[133] Laura J Scott, Karen L Mohlke, Lori L Bonnycastle, Cristen J Willer, Yun Li, William L Duren, Michael R Erdos, Heather M Stringham, Peter S Chines, Anne U Jackson, Ludmila Prokunina-Olsson, Chia-Jen Ding, Amy J Swift, Narisu Narisu, Tianle Hu, Randall Pruim, Rui Xiao, Xiao-Yi Li, Karen N Conneely, Nancy L Riebow, Andrew G Sprau, Maurine Tong, Peggy P White, Kurt N Hetrick, Michael W Barnhart, Craig W Bark, Janet L Goldstein, Lee Watkins, Fang Xiang, Jouko Saramies, Thomas A Buchanan, Richard M Watanabe, Timo T Valle, Leena Kinnunen, Gonçalo R Abecasis, Elizabeth W Pugh, Kimberly F Doheny, Richard N Bergman, Jaakko Tuomilehto, Francis S Collins, and Michael Boehnke. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science (New York, N.Y.)*, 316:1341–1345, 2007.

[134] Robert A Scott, Vasiliki Lagou, Ryan P Welch, Eleanor Wheeler, May E Montasser, Jian'an Luan, Reedik Mägi, Rona J Strawbridge, Emil Rehnberg, Stefan Gustafsson, Stavroula Kanoni, Laura J Rasmussen-Torvik, Loïc Yengo, Cecile Lecoeur, Dmitry Shungin, Serena Sanna, Carlo Sidore, Paul C D Johnson, J Wouter Jukema, Toby Johnson, Anubha Mahajan, Niek Verweij, Gudmar Thorleifsson, Jouke-Jan Hottenga, Sonia Shah, Albert V Smith, Bengt Sennblad, Christian Gieger, Perttu Salo, Markus Perola, Nicholas J Timpson, David M Evans, Beate St Pourcain, Ying Wu, Jeanette S Andrews, Jennie Hui, Lawrence F Bielak, Wei Zhao, Momoko Horikoshi, Pau Navarro, Aaron Isaacs, Jeffrey R O'Connell, Kathleen Stirrups, Veronique Vitart, Caroline Hayward, Tõnu Esko, Evelin Mihailov, Ross M Fraser, Tove Fall, Benjamin F Voight, Soumya Raychaudhuri, Han Chen, Cecilia M Lindgren, Andrew D P Morris, Nigel W Rayner, Neil Robertson, Denis Rybin, Ching-Ti Liu, Jacques S Beckmann, Sara M Willems, Peter S Chines, Anne U Jackson, Hyun Min Kang, Heather M Stringham, Kijoung Song, Toshiko Tanaka, John F Peden, Anuj Goel, Andrew A Hicks, Ping An, Martina Müller-Nurasyid, Anders Franco-Cereceda, Lasse Folkersen, Letizia Marullo, Hanneke Jansen, Albertine J Oldehinkel, Marcel Bruinenberg, James S Pankow, Kari E North, Nita G Forouhi, Ruth J F Loos, Sarah Edkins, Tibor V Varga, Göran Hallmans, Heikki Oksa, Mulas Antonella, Ramaiah Nagaraja, Stella Trompet, Ian Ford, Stephan

J L Bakker, Augustine Kong, Meena Kumari, Bruna Gigante, Christian Herder, Patricia B Munroe, Mark Caulfield, Jula Antti, Massimo Mangino, Kerrin Small, Iva Miljkovic, Yongmei Liu, Mustafa Atalay, Wieland Kiess, Alan L James, Fernando Rivadeneira, Andre G Uitterlinden, Colin N A Palmer, Alex S F Doney, Gonneke Willemsen, Johannes H Smit, Susan Campbell, Ozren Polasek, Lori L Bonnycastle, Serge Hercberg, Maria Dimitriou, Jennifer L Bolton, Gerard R Fowkes, Peter Kovacs, Jaana Lindström, Tatijana Zemunik, Stefania Bandinelli, Sarah H Wild, Hanneke V Basart, Wolfgang Rathmann, Harald Grallert, Winfried Maerz, Marcus E Kleber, Bernhard O Boehm, Annette Peters, Peter P Pramstaller, Michael A Province, Ingrid B Borecki, Nicholas D Hastie, Igor Rudan, Harry Campbell, Hugh Watkins, Martin Farrall, Michael Stumvoll, Luigi Ferrucci, Dawn M Waterworth, Richard N Bergman, Francis S Collins, Jaakko Tuomilehto, Richard M Watanabe, Eco J C de Geus, Brenda W Penninx, Albert Hofman, Ben A Oostra, Bruce M Psaty, Peter Vollenweider, James F Wilson, Alan F Wright, G Kees Hovingh, Andres Metspalu, Matti Uusitupa, Patrik K E Magnusson, Kirsten O Kyvik, Jaakko Kaprio, Jackie F Price, George V Dedoussis, Panos Deloukas, Pierre Meneton, Lars Lind, Michael Boehnke, Alan R Shuldiner, Cornelia M van Duijn, Andrew D P Morris, Anke Toenjes, Patricia A Peyser, John P Beilby, Antje Körner, Johanna Kuusisto, Markku Laakso, Stefan R Bornstein, Peter E H Schwarz, Timo A Lakka, Rainer Rauramaa, Linda S Adair, George Davey Smith, Tim D Spector, Thomas Illig, Ulf de Faire, Anders Hamsten, Vilmundur Gudnason, Mika Kivimaki, Aroon Hingorani, Sirkka M Keinanen-Kiukaanniemi, Timo E Saaristo, Dorret I Boomsma, Kari Stefansson, Pim van der Harst, Josée Dupuis, Nancy L Pedersen, Naveed Sattar, Tamara B Harris, Francesco Cucca, Samuli Ripatti, Veikko Salomaa, Karen L Mohlke, Beverley Balkau, Philippe Froguel, Anneli Pouta, Marjo-Riitta Jarvelin, Nicholas J Wareham, Nabila Bouatia-Naji, Mark I McCarthy, Paul W Franks, James B Meigs, Tanya M Teslovich, Jose C Florez, Claudia Langenberg, Erik Ingelsson, Inga Prokopenko, and Inês Barroso. Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nature genetics*, 44(9):991–1005, 2012.

[135] Luan Shu, Aleksey V. Matveyenko, Julie Kerr-Conte, Jae Hyoung Cho, C. H S McIntosh, and Kathrin Maedler. Decreased TCF7L2 protein levels in type 2 diabetes mellitus correlate with downregulation of GIP- and GLP-1 receptors and impaired beta-cell function. *Human Molecular Genetics*, 18(13):2388–2399, 2009.

[136] Robert Sladek, Ghislain Rocheleau, Johan Rung, Christian Dina, Lishuang Shen, David Serre, Philippe Boutin, Daniel Vincent, Alexandre Belisle, Samy Hadjadj, Beverley Balkau, Barbara Heude, Guillaume Charpentier, Thomas J Hudson, Alexandre Montpetit, Alexey V Pshezhetsky, Marc Prentki, Barry I Posner, David J Balding, David Meyre, Constantin Polychronakos, and Philippe Froguel. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445:881–885, 2007.

[137] Kerrin S Small, Asa K Hedman, Elin Grundberg, Alexandra C Nica, Gudmar Thorleifsson, Augustine Kong, Unnur Thorsteindottir, So-Youn Shin, Hannah B Richards, Nicole Soranzo, Kourosh R Ahmadi, Cecilia M Lindgren, Kari Stefansson, Emmanouil T Dermitzakis, Panos Deloukas, Timothy D Spector, and Mark I McCarthy.

Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. *Nature genetics*, 43(6):561–564, 2011.

[138] Scott Smemo, Juan J Tena, Kyoung-Han Kim, Eric R Gamazon, Noboru J Sakabe, Carlos Gómez-Marín, Ivy Aneas, Flavia L Credidio, Debora R Débora R Sobreira, Nora F Wasserman, Ju Hee Lee, Vijitha Puviindran, Davis Tam, Michael Shen, Joe Eun Son, Niki Alizadeh Vakili, Hoon-Ki Sung, Silvia Naranjo, Rafael D Acemel, Miguel Manzanares, Andras Nagy, Nancy J Cox, Chi-Chung Hui, Jose Luis Gomez-Skarmeta, Marcelo A Nóbrega, Carlos Gomez-Marin, Ivy Aneas, Flavia L Credidio, Debora R Débora R Sobreira, Nora F Wasserman, Ju Hee Lee, Vijitha Puviindran, Davis Tam, Michael Shen, Joe Eun Son, Niki Alizadeh Vakili, Hoon-Ki Sung, Silvia Naranjo, Rafael D Acemel, Miguel Manzanares, Andras Nagy, Nancy J Cox, Chi-Chung Hui, Jose Luis Gomez-Skarmeta, and Marcelo A Nobrega. Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature*, advance on(7492):371–5, mar 2014.

[139] Doug Speed, Gibran Hemani, Michael R. Johnson, and David J. Balding. Improved heritability estimation from genome-wide SNPs. *American Journal of Human Genetics*, 91(6):1011–1021, 2012.

[140] Eli A Stahl, Daniel Wegmann, Gosia Trynka, Javier Gutierrez-Achury, Ron Do, Benjamin F Voight, Peter Kraft, Robert Chen, Henrik J Kallberg, Fina A S Kurreeman, Sekar Kathiresan, Cisca Wijmenga, Peter K Gregersen, Lars Alfredsson, Katherine A Siminovitch, Jane Worthington, Paul I W de Bakker, Soumya Raychaudhuri, and Robert M Plenge. Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis, 2012.

[141] Oliver Stegle, Leopold Parts, Matias Piipari, John Winn, and Richard Durbin. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature protocols*, 7(3):500–7, 2012.

[142] Kent D Taylor, Jill M Norris, and Jerome I Rotter. Genome-Wide Association: Which Do You Want First: the Good News, the Bad News, or the Good News? . *Diabetes*, 56(12):2844–2848, dec 2007.

[143] Albert Tenesa and Chris S Haley. The heritability of human disease: estimation, uses and abuses. *Nature reviews. Genetics*, 14:139–49, 2013.

[144] Tanya M Teslovich, Kiran Musunuru, Albert V Smith, Andrew C Edmondson, Ioannis M Stylianou, Masahiro Koseki, James P Pirruccello, Samuli Ripatti, Daniel I Chasman, Cristen J Willer, Christopher T Johansen, Sigrid W Fouchier, Aaron Isaacs, Gina M Peloso, Maja Barbalic, Sally L Ricketts, Joshua C Bis, Yurii S Aulchenko, Gudmar Thorleifsson, Mary F Feitosa, John Chambers, Marju Orho-Melander, Olle Melander, Toby Johnson, Xiaohui Li, Xiuqing Guo, Mingyao Li, Yoon Shin Cho, Min Jin Go, Young Jin Kim, Jong-Young Lee, Taesung Park, Kyunga Kim, Xueling Sim, Rick Twee-Hee Ong, Damien C Croteau-Chonka, Leslie A Lange, Joshua D Smith, Kijoung Song, Jing Hua Zhao, Xin Yuan, Jian'an Luan, Claudia Lamina, Andreas

Ziegler, Weihua Zhang, Robert Y L Zee, Alan F Wright, Jacqueline C M Witteman, James F Wilson, Gonneke Willemsen, H-Erich Wichmann, John B Whitfield, Dawn M Waterworth, Nicholas J Wareham, Gérard Waeber, Peter Vollenweider, Benjamin F Voight, Veronique Vitart, Andre G Uitterlinden, Manuela Uda, Jaakko Tuomilehto, John R Thompson, Toshiko Tanaka, Ida Surakka, Heather M Stringham, Tim D Spector, Nicole Soranzo, Johannes H Smit, Juha Sinisalo, Kaisa Silander, Eric J G Sijbrands, Angelo Scuteri, James Scott, David Schlessinger, Serena Sanna, Veikko Salomaa, Juha Saharinen, Chiara Sabatti, Aimo Ruokonen, Igor Rudan, Lynda M Rose, Robert Roberts, Mark Rieder, Bruce M Psaty, Peter P Pramstaller, Irene Pichler, Markus Perola, Brenda W J H Penninx, Nancy L Pedersen, Cristian Pattaro, Alex N Parker, Guillaume Pare, Ben A Oostra, Christopher J O'Donnell, Markku S Nieminen, Deborah A Nickerson, Grant W Montgomery, Thomas Meitinger, Ruth McPherson, Mark I McCarthy, Wendy McArdle, David Masson, Nicholas G Martin, Fabio Marroni, Massimo Mangino, Patrik K E Magnusson, Gavin Lucas, Robert Luben, Ruth J F Loos, Marja-Liisa Lokki, Guillaume Lettre, Claudia Langenberg, Lenore J Launer, Edward G Lakatta, Reijo Laaksonen, Kirsten O Kyvik, Florian Kronenberg, Inke R König, Kay-Tee Khaw, Jaakko Kaprio, Lee M Kaplan, Asa Johansson, Marjo-Riitta Jarvelin, A Cecile J W Janssens, Erik Ingelsson, Wilmar Igl, G Kees Hovingh, Jouke-Jan Hottenga, Albert Hofman, Andrew A Hicks, Christian Hengstenberg, Iris M Heid, Caroline Hayward, Aki S Havulinna, Nicholas D Hastie, Tamara B Harris, Talin Haritunians, Alistair S Hall, Ulf Gyllensten, Candace Guiducci, Leif C Groop, Elena Gonzalez, Christian Gieger, Nelson B Freimer, Luigi Ferrucci, Jeanette Erdmann, Paul Elliott, Kenechi G Ejebe, Angela Döring, Anna F Dominiczak, Serkalem Demissie, Panagiotis Deloukas, Eco J C de Geus, Ulf de Faire, Gabriel Crawford, Francis S Collins, Yii-der I Chen, Mark J Caulfield, Harry Campbell, Noel P Burtt, Lori L Bonnycastle, Dorret I Boomsma, S Matthijs Boekholdt, Richard N Bergman, Inês Barroso, Stefania Bandinelli, Christie M Ballantyne, Themistocles L Assimes, Thomas Quertermous, David Altshuler, Mark Seielstad, Tien Y Wong, E-Shyong Tai, Alan B Feranil, Christopher W Kuzawa, Linda S Adair, Herman A Taylor, Ingrid B Borecki, Stacey B Gabriel, James G Wilson, Hilma Holm, Unnur Thorsteinsdottir, Vilmundur Gudnason, Ronald M Krauss, Karen L Mohlke, Jose M Ordovas, Patricia B Munroe, Jaspal S Kooner, Alan R Tall, Robert A Hegele, John J P Kastelein, Eric E Schadt, Jerome I Rotter, Eric Boerwinkle, David P Strachan, Vincent Mooser, Kari Stefansson, Muredach P Reilly, Nilesh J Samani, Heribert Schunkert, L Adrienne Cupples, Manjinder S Sandhu, Paul M Ridker, Daniel J Rader, Cornelia M van Duijn, Leena Peltonen, Gonçalo R Abecasis, Michael Boehnke, and Sekar Kathiresan. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466(7307):707–13, 2010.

[145] The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[146] The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(may):25–29, 2000.

[147] The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348:648–660, 2015.

[148] Chao Tian, Peter K. Gregersen, and Michael F. Seldin. Accounting for ancestry: Population substructure and genome-wide association studies. *Human Molecular Genetics*, 17(R2), 2008.

[149] Robert Tibshirani. Regression Selection and Shrinkage via the Lasso. *Journal of the Royal Statistical Society B*, 58(1):267–288, 1994.

[150] Jason M. Torres, Eric R. Gamazon, Esteban J. Parra, Jennifer E. Below, Adan Valladares-Salgado, Niels Wacher, Miguel Cruz, Craig L. Hanis, and Nancy J. Cox. Cross-tissue and tissue-specific eQTLs: Partitioning the heritability of a complex trait. *American Journal of Human Genetics*, 95(5):521–534, 2014.

[151] Michael Trauner, Marco Arrese, and Martin Wagner. Fatty liver and lipotoxicity, 2010.

[152] Shashaank Vattikuti, Juen Guo, and Carson C. Chow. Heritability and Genetic Correlations Explained by Common SNPs for Metabolic Syndrome Traits, 2012.

[153] Nathalie Vionnet, El Habib Hani, Sophie Dupont, Sophie Gallina, Stephan Francke, Sébastien Dotte, Frédérique De Matos, Emmanuelle Durand, Frédéric Leprêtre, Cécile Lecoeur, Philippe Gallina, Lirije Zekiri, Christian Dina, and Philippe Froguel. Genomewide Search for Type 2 Diabetes–Susceptibility Genes in French Whites: Evidence for a Novel Susceptibility Locus for Early-Onset Diabetes on Chromosome 3q27-qter and Independent Replication of a Type 2–Diabetes Locus on Chromosome 1q21. *The American Journal of Human Genetics*, 67(6):1470–1480, apr 2016.

[154] Peter M. Visscher, Matthew A. Brown, Mark I. McCarthy, and Jian Yang. Five years of GWAS discovery, 2012.

[155] Peter M Visscher, Brian McEvoy, and Jian Yang. From Galton to GWAS: quantitative genetics of human height. *Genetics research*, 92:371–379, 2010.

[156] Peter M Visscher, Jian Yang, and Michael E Goddard. A commentary on 'common SNPs explain a large proportion of the heritability for human height' by Yang et al. (2010). *Twin research and human genetics : the official journal of the International Society for Twin Studies*, 13:517–524, 2010.

[157] Benjamin F BF Voight, Laura J LJ Scott, Valgerdur Steinthorsdottir, P Andrew, YS Yurii S Aulchenko, Gudmar Thorleifsson, Laura J LJ McCulloch, Teresa Ferreira, AD AP Morris, C Dina, RP Welch, E Zeggini, C Huth, YS Yurii S Aulchenko, Gudmar Thorleifsson, Laura J LJ McCulloch, Teresa Ferreira, H Grallert, N Amin, GM Wu, CJ Willer, S Raychaudhuri, SA McCarroll, C Langenberg, OM Hofmann, J Dupuis, L Qi, AV Segre, M Van Hoek, P Navarro, K Ardlie, B Balkau, R Benediktsson, AJ Bennett, R Blagieva, E Boerwinkle, LL Bonnycastle, KB Bostrom, B Bravenboer, S Bumpstead, NP Burtt, G Charpentier, PS Chines, M Cornelis, DJ Couper, G Crawford, Asf

Doney, KS Elliott, AL Elliott, Erdos, CS Fox, CS Franklin, M Ganser, C Gieger, N Grarup, T Green, S Griffin, CJ Groves, C Guiducci, S Hadjadj, N Hassanali, C Herder, B Isomaa, AU Jackson, Prv Johnson, T Jorgensen, Whl Kao, N Klopp, A Kong, P Kraft, J Kuusisto, T Lauritzen, M Li, A Lieverse, CM Lindgren, V Lyssenko, M Marre, T Meitinger, K Midthjell, MA Morken, N Narisu, P Nilsson, KR Owen, F Payne, Jrb Perry, AK Petersen, C Platou, C Proenca, I Prokopenko, W Rathmann, NW Rayner, NR Robertson, G Rocheleau, M Roden, MJ Sampson, R Saxena, BM Shields, P Shrader, G Sigurdsson, T Sparso, K Strassburger, HM Stringham, Q Sun, AJ Swift, B Thorand, J Tichet, T Tuomi, RM Van Dam, TW Van Haeften, T Van Herpt, JV Van Vliet-Ostaptchouk, GB Walters, MN Weedon, C Wijmenga, J Witteman, RN Bergman, S Cauchi, FS Collins, AL Gloyn, U Gyllensten, T Hansen, WA Hide, GA Hitman, A Hofman, DJ Hunter, K Hveem, M Laakso, KL Mohlke, AD AP Morris, Cna Palmer, PP Pramstaller, I Rudan, E Sijbrands, LD Stein, J Tuomilehto, A Uitterlinden, M Walker, NJ Wareham, RM Watanabe, GR Abecasis, BO Boehm, H Campbell, MJ Daly, AT Hattersley, FB Hu, JB Meigs, JS Pankow, O Pedersen, HE Wichmann, I Barroso, JC Florez, TM Frayling, L Groop, R Sladek, U Thorsteinsdottir, JF Wilson, T Illig, P Froguel, CM Van Duijn, K Stefansson, D Altshuler, M Boehnke, MI McCarthy, Magic Investigators, and Giant Consortium. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nature Genetics*, 42(7):579–89, 2010.

[158] K M Waters, D O Stram, M T Hassanein, L Le Marchand, L R Wilkens, G Maskarinec, K R Monroe, L N Kolonel, D Altshuler, B E Henderson, and C A Haiman. Consistent association of type 2 diabetes risk variants found in europeans in diverse racial and ethnic groups. *PLoS Genet*, 6(8), 2010.

[159] The Wellcome, Trust Case, and Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447:661–678, 2007.

[160] Jennifer Wessel, Audrey Y Chu, Sara M Willems, Shuai Wang, Hanieh Yaghootkar, Jennifer A Brody, Marco Dauriz, Marie-France Hivert, Sridharan Raghavan, Leonard Lipovich, Bertha Hidalgo, Keolu Fox, Jennifer E Huffman, Ping An, Yingchang Lu, Laura J Rasmussen-Torvik, Niels Grarup, Margaret G Ehm, Li Li, Abigail S Baldridge, Alena Stančáková, Ravinder Abrol, Céline Besse, Anne Boland, Jette Bork-Jensen, Myriam Fornage, Daniel F Freitag, Melissa E Garcia, Xiuqing Guo, Kazuo Hara, Aaron Isaacs, Johanna Jakobsdottir, Leslie A Lange, Jill C Layton, Man Li, Jing Hua Zhao, Karina Meidtner, Alanna C Morrison, Mike A Nalls, Marjolein J Peters, Maria Sabater-Lleal, Claudia Schurmann, Angela Silveira, Albert V Smith, Lorraine Southam, Marcus H Stoiber, Rona J Strawbridge, Kent D Taylor, Tibor V Varga, Kristine H Allin, Najaf Amin, Jennifer L Aponte, Tin Aung, Caterina Barbieri, Nathan A Bihlmeyer, Michael Boehnke, Cristina Bombieri, Donald W Bowden, Sean M Burns, Yuning Chen, Yii-DerI Chen, Ching-Yu Cheng, Adolfo Correa, Jacek Czajkowski, Abbas Dehghan, Georg B Ehret, Gudny Eiriksdottir, Stefan A Escher, Aliki-Eleni Farmaki, Mattias Frånberg, Giovanni Gambaro, Franco Giulianini, William A Goddard, Anuj Goel, Omri Gottesman, Megan L Grove, Stefan Gustafsson, Yang Hai,

Göran Hallmans, Jiyoung Heo, Per Hoffmann, Mohammad K Ikram, Richard A Jensen, Marit E Jørgensen, Torben Jørgensen, Maria Karaleftheri, Chiea C Khor, Andrea Kirkpatrick, Aldi T Kraja, Johanna Kuusisto, Ethan M Lange, I T Lee, Wen-Jane Lee, Aaron Leong, Jiemin Liao, Chunyu Liu, Yongmei Liu, Cecilia M Lindgren, Allan Linneberg, Giovanni Malerba, Vasiliki Mamakou, Eirini Marouli, Nisa M Maruthur, Angela Matchan, Roberta McKean-Cowdin, Olga McLeod, Ginger A Metcalf, Karen L Mohlke, Donna M Muzny, Ioanna Ntalla, Nicholette D Palmer, Dorota Pasko, Andreas Peter, Nigel W Rayner, Frida Renström, Ken Rice, Cinzia F Sala, Bengt Sennblad, Ioannis Serafetinidis, Jennifer A Smith, Nicole Soranzo, Elizabeth K Speliotes, Eli A Stahl, Kathleen Stirrups, Nikos Tentolouris, Anastasia Thanopoulou, Mina Torres, Michela Traglia, Emmanouil Tsafantakis, Sundas Javad, Lisa R Yanek, Eleni Zengini, Diane M Becker, Joshua C Bis, James B Brown, L Adrienne Cupples, Torben Hansen, Erik Ingelsson, Andrew J Karter, Carlos Lorenzo, Rasika A Mathias, Jill M Norris, Gina M Peloso, Wayne H-H Sheu, Daniela Toniolo, Dhananjay Vaidya, Rohit Varma, Lynne E Wagenknecht, Heiner Boeing, Erwin P Bottinger, George Dedoussis, Panos Deloukas, Ele Ferrannini, Oscar H Franco, Paul W Franks, Richard A Gibbs, Vilmundur Gudnason, Anders Hamsten, Tamara B Harris, Andrew T Hattersley, Caroline Hayward, Albert Hofman, Jan-Håkan Jansson, Claudia Langenberg, Lenore J Launer, Daniel Levy, Ben A Oostra, Christopher J O'Donnell, Stephen O'Rahilly, Sandosh Padmanabhan, James S Pankow, Ozren Polasek, Michael A Province, Stephen S Rich, Paul M Ridker, Igor Rudan, Matthias B Schulze, Blair H Smith, André G Uitterlinden, Mark Walker, Hugh Watkins, Tien Y Wong, Eleftheria Zeggini, Markku Laakso, Ingrid B Borecki, Daniel I Chasman, Oluf Pedersen, Bruce M Psaty, E Shyong Tai, Cornelia M van Duijn, Nicholas J Wareham, Dawn M Waterworth, Eric Boerwinkle, W H Linda Kao, Jose C Florez, Ruth J F Loos, James G Wilson, Timothy M Frayling, David S Siscovick, Josée Dupuis, Jerome I Rotter, James B Meigs, Robert A Scott, and Mark O Goodarzi. Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. *Nature communications*, 6:5897, 2015.

[161] Heather E Wheeler, Kaanan P Shah, Jonathon Brenner, Tzintzuni Garcia, Keston Aquino-Michaels, Nancy J Cox, Dan L Nicolae, and Hae Kyung Im. Survey of the Heritability and Sparsity of Gene Expression Traits Across Human Tissues. *bioRxiv*, mar 2016.

[162] David R. Whiting, Leonor Guariguata, Clara Weil, and Jonathan Shaw. IDF Diabetes Atlas: Global estimates of the prevalence of diabetes for 2011 and 2030. *Diabetes Research and Clinical Practice*, 94(3):311–321, 2011.

[163] WHO. Global status report on noncommunicable diseases 2014. *World Health*, page 176, 2014.

[164] Hadley Wickham. ggplot2. *Wiley Interdisciplinary Reviews: Computational Statistics*, 3(2):180–185, 2011.

[165] Cristen J. Willer, Yun Li, and Gon??alo R. Abecasis. METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, 26(17):2190–2191, 2010.

[166] Jian Yang, Beben Benyamin, Brian P Mcevoy, Scott Gordon, Anjali K Henders, R Dale, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, Michael E Goddard, and Peter M Visscher. Common SNPs explain a large proportion of heritability for human height. *Nature Genetics*, 42:565–569, 2011.

[167] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. GCTA: a tool for genome-wide complex trait analysis. *American journal of human genetics*, 88:76–82, 2011.

[168] Jian Yang, Sang Hong Lee, Michael E Goddard, and Peter M Visscher. Genome-wide complex trait analysis (GCTA): methods, data analyses, and interpretations. *Methods in molecular biology (Clifton, N.J.)*, 1019:215–36, 2013.

[169] Jian Yang, Teri A Manolio, Louis R Pasquale, Eric Boerwinkle, Neil Caporaso, Julie M Cunningham, Mariza De Andrade, Bjarke Feenstra, Eleanor Feingold, M Geoffrey Hayes, William G Hill, Maria Teresa Landi, Alvaro Alonso, Guillaume Lettre, Peng Lin, Hua Ling, William Lowe, Rasika A Mathias, Mads Melbye, Elizabeth Pugh, Marilyn C Cornelis, Bruce S Weir, Michael E Goddard, and Peter M Visscher. Genome partitioning of genetic variation for complex traits using common SNPs. *Nature Genetics*, 43:519–525, 2011.

[170] Matthew D Young, Matthew J Wakefield, Gordon K Smyth, and Alicia Oshlack. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome biology*, 11(2):R14, 2010.

[171] Noah Zaitlen and Peter Kraft. Heritability in the genome-wide association era, 2012.

[172] Noah Zaitlen, Peter Kraft, Nick Patterson, Bogdan Pasaniuc, Gaurav Bhatia, Samuela Pollack, and Alkes L. Price. Using Extended Genealogy to Estimate Components of Heritability for 23 Quantitative and Dichotomous Traits. *PLoS Genetics*, 9, 2013.

[173] Eleftheria Zeggini, Michael N Weedon, Cecilia M Lindgren, Timothy M Frayling, Katherine S Elliott, Hana Lango, Nicholas J Timpson, John R B Perry, Nigel W Rayner, Rachel M Freathy, Jeffrey C Barrett, Beverley Shields, Andrew P Morris, Sian Ellard, Christopher J Groves, Lorna W Harries, Jonathan L Marchini, Katharine R Owen, Beatrice Knight, Lon R Cardon, Mark Walker, Graham A Hitman, Andrew D Morris, Alex S F Doney, Mark I McCarthy, and Andrew T Hattersley. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science (New York, N.Y.)*, 316:1336–1341, 2007.

[174] Zhihong Zhu, Futao Zhang, Han Hu, Andrew Bakshi, Matthew R Robinson, Joseph E Powell, Grant W Montgomery, Michael E Goddard, Naomi R Wray, Peter M Visscher, and Jian Yang. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet*, advance on, mar 2016.