Identification of Potential Molecular Traits Underlying the Genetic Predisposition to Complex diseases Using Multi-omics and Meta-analysis

by

Li Guan

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2021

Doctoral Committee:

Professor Michael Boehnke, Co-Chair
Research Professor Laura Scott, Co-Chair
Associate Professor Stephen C. J. Parker
Associate Professor Maureen Sartor
Professor Kerby Shedden

Li Guan
guanli@umich.edu
ORCID iD:0000-0003-1205-7613

Dedication

To my family and friends.

Acknowledgments

I am grateful to Prof. Laura Scott and Prof. Michael Boehnke for providing the opportunity to work on these exciting data sets and their continual help on my research over the years. I thank all members of the Boehnke-Scott group for always willing to help and for their friendship. I am grateful for having been able to work with so many kind and brilliant people from the FUSION Tissue Biopsy group, from Prof. Karen Mohlke's group, and from Prof. Kerrin Small's group. I benefitted greatly from observing Dr. Leland Taylor, Dr. Anne Jackson, Dr. Ryan Welch, and Dr. Narisu Narisu's work. Not only did I learn analysis skills from them, but I also learned teamwork and good work ethic. I wish to express my appreciation to Dr. Corbin Quick and Prof. Xihong Lin for providing an opportunity to contribute to statistical software development. I especially thank my thesis committee members Prof. Stephen C. J. Parker, Prof. Maureen Sartor, and Prof. Kerby Shedden for their insightful ideas and continual support over the years.

I would like to thank the dedicated staff from the Department of Bioinformatics and from the Boehnke-Scott group, who made my life so much easier. I am particularly grateful to Prof. Margit Burmeister and Prof. Maureen Sartor, the co-directors of the Bioinformatics Ph.D. program, for their advice and guidance throughout my Ph.D. studies. I especially thank the department of Bioinformatics for providing an exceptional training environment to students. I am grateful for the opportunity to begin my Ph.D. studies in the PIBS program, which provided a flexible environment to find and pursue my research interests. I am extremely happy to have been a part of the Department of Bioinformatics and the broader University of Michigan community.

Finally, I would like to thank my family and friends. It is impossible for me to express how thankful I am for your unconditional love and unceasing support.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

1000G	1000 Genomes project
ACAT	Aggregated Cauchy association test
ALT	Alanine transaminase
APEX	All-in-one Package for Efficient Xqtl analysis
ATAC-seq	Assay for transposase accessible chromatin followed by sequencing
BMI	Body mass index
CAD	Coronary artery disease
Cas9	CRISPR-associated protein 9
CMD	Cardiometabolic diseases
CPM	Counts per million
CRISPR	Clustered regularly interspaced short palindromic repeats
DAP	Deterministic approximation of posteriors
DNAme	DNA methylation
EBI	European Bioinformatics Institute
eQTL	Expression quantitative trait loci
FDR	False discovery rate
FFA	Free fatty acid
FUSION	The Finland-United States Investigation of NIDDM Genetics
Geuvadis	The Geuvadis project
GTEx	The Genotype-Tissue Expression
GWAS	Genome-wide association study
HapMap	The International HapMap Project

HBB	Hemoglobin subunit beta
HOMA	Homeostatic model assessment
HRS	Haplotype Reference Consortium
IR	Insulin resistance
iPSCs	Induced pluripotent stem cells
ks.test	Kolmogorov-Smirnov test
LD	Linkage disequilibrium
IDF	International Diabetes Federation
MAF	Minor allele frequency
METSIM	METabolic Syndrome in Men
miRNA	microRNA
mQTL	Methylation QTLs
NASH	Non-alcoholic steatohepatitis
NHGRI	National Human Genome Research Institute
NIDDM	Non-insulin-dependent diabetes mellitus
NGT	Normal glucose tolerance
PC	Principal conponetns
PIP	posterior inclusion probability
POPRES	Population Reference Sample
PRS	Polygenic risk score
QC	Quality control
QTL	Quantitative trait loci
RIN	RNA integrity number
RPMMM	Reads per million mapped to microRNAs
RCP	Regional colocalization probability
SCP	SNP-level colocalization probability
SNP	Single nucleotide polymorphisms
SS	Simple steatosis
SV	Surrogate variables

SGBS	Simpson-Golabi-Behmel syndrome
T2D	Type 2 diabetes
TIN	Transcript integrity number
TPM	Transcripts per million
TMM	Trimmed Mean of M-values
TSS	Transcription start site
TwinsUK	The UK Adult Twin Registry
UKB	UK Biobank
VIF	Variance inflation factor
var-cov	Variance-covariance
WGS	Whole genome sequencing
WHR	Wasit hip ratio
WHRadjBMI	BMI adjusted waist hip ratio

Abstract

Complex diseases are multifactorial diseases caused by a complex combination of genetic, environmental and lifestyle effects. Numerous non-coding regions that increase the risk for complex diseases have been discovered by successive waves of genome-wide association studies (GWAS). However, the mechanistic understanding underlying GWAS loci has lagged behind GWAS discovery. The rapidly evolving innovations in high throughput molecular profiling technologies have greatly increased our ability to study the downstream transcriptional and epigenetic impacts of disease-associated variants. In my dissertation, I studied the mechanistic underpinning of GWAS loci for complex diseases by using high throughput molecular profiling data and by combining information from multiple studies via meta-analysis.

First, I prioritized mRNAs, microRNAs(miRNAs), and DNA methylation(DNAme) sites potentially involved in Type 2 diabetes (T2D) mechanisms, using data sets in skeletal muscle and subcutaneous adipose tissues from up to 301 individuals from the Finland-United States Investigation of Non-insulin-dependent diabetes mellitus (NIDDM) Genetics (FUSION) Tissue Biopsy Study. I identified quantitative trait loci (QTLs) for mRNAs and miRNAs expression levels and DNAme levels. A smaller proportion of miRNAs had *cis*-QTLs than mRNAs, and the lead variants for miRNA *cis*-QTLs had lower minor allele frequency(MAF) than the lead variants for mRNA *cis*-QTLs. These observations suggest that compared to mRNAs, miRNAs may be under stronger selective pressure and therefore have a lower level of *cis*-QTL regulation. By integrating the QTLs for molecular traits with T2D GWAS associations, I identified mRNAs and DNAme sites potentially underlying T2D GWAS loci. By testing for associations of molecular trait levels with 48

T2D related traits, we identified mRNAs, miRNAs, and DNAme sites associated with T2D related traits. Multiple lines of evidence suggested that *INHBB* was likely to underlie the GWAS locus rs11688682 as its eQTL was colocalized with the rs11688682 GWAS locus in both tissues, and *INHBB* was positively correlated with insulin-related physiological traits in subcutaneous adipose tissue. In addition, the luciferase assay conducted by our collaborators confirmed that the T2D risk allele rs11688682-G increased transcriptional activity in preadipocytes and adipocytes.

Second, I describe a collaborative project using data sets from TwinsUK, METSIM, GTEx and FUSION to perform RNA-seq based eQTL meta-analysis in subcutaneous adipose tissue from 2256 individuals of European ancestry. Of the 19,108 genes present in all studies, the meta-analysis revealed $\geq$ 1 eQTL for 15335 (80.3%) genes: 6440 (33.7%) genes had exactly one eQTL, 8895 genes (46.6%) had $\geq$ 2 eQTL eQTLs. I evaluated the evidence for colocalization between the meta-analysis eQTLs and the GWAS signals for seven cardiometabolic traits: T2D, Body mass index, Waist-hip ratio, BMI adjusted waist-hip ratio, Coronary artery disease, fasting glucose and fasting insulin. I identified 334 genes that had primary eQTLs colocalized with at least one GWAS signal, and 202 genes that had secondary eQTLs colocalized with at least one GWAS signal.

Throughout my dissertation work, I used molecular profiling data of multiple types of molecular traits and combined eQTL associations from multiple studies to provide clues to the molecular traits that may mediate complex disease risks. These prioritized molecular traits are promising candidates for functional follow-up of their roles in disease etiology.

Chapter 1

Introduction

Complex diseases, such as psychiatric disorders, cardiovascular diseases, autoimmune diseases, and various types of cancers[1], are caused by a combination of genetic, environmental, and lifestyle effects[2]. Complex diseases do not follow Mendelian inheritance patterns, but show familial aggregation of cases and have moderate to high evidence of heritability[3], [4]. A central goal of human complex disease studies is to identify and functionally characterize the genetic basis of the diseases and thereby to discover therapeutic targets and to develop precision medicine strategies. Linkage analysis, which tests for cosegregation of a gene marker and a disease of interest within a family, has localized the causal genes for many Mendelian diseases, such as Duchenne muscular dystrophy[5], cystic fibrosis[6]–[8] and Huntington disease[9]. Inspired by its success in unraveling the genetics for Mendelian diseases, linkage analysis was used as a main strategy in the early efforts to identify the genetic factors implicated in complex disease predisposition. However, the application of linkage analysis to complex diseases achieved limited success[10], which is in part explained by the fact that common variants of multiple genes comprise the genetic architecture of complex diseases[11]–[13] and that the majority of common variants have small to modest effects[14].

A genome-wide association study (GWAS) design was proposed to improve the power to detect common variants with small effects[15]. In GWAS, a genome-wide dense map of genetic variants, most commonly single nucleotide polymorphisms (SNPs), is used to

test for allele-frequency difference between case and control or between individuals with various levels of continuous traits[16]. GWAS soon became feasible with the advent of large-scale array-based genotype technologies. Over the past two decades due to rapid technological advances in cataloging human DNA sequence variation and their declining cost[17], successive waves of GWAS have revolutionized the search for genetic risk loci that predispose to complex diseases. As of 2020 July, the NHGRI-EBI GWAS catalog has curated single nucleotide variation associations for 4466 diseases or traits from 4054 research papers[18]. 49,451 of the 89,588 (55.2%) recorded associations meet genome-wide significance threshold(p-value < 5×10^{-8})[19]. GWAS-discovered variants explained a much larger proportion of genetic variation than variants discovered in the pre-GWAS era[20]. Despite the huge success in the discovery of risk-conferring loci, GWAS provides little mechanistic insights into how the discovered loci affect disease susceptibility, especially for the non-coding variants, which comprise the majority of loci identified by GWAS.

To study how GWAS variants confer disease risk, it is crucial to consider how genetic information propagates through biological processes to exert effects. Recent technological and computational advances have provided increasingly reliable measurements of intermediate molecular traits, from epigenetic markers to gene, protein, and metabolite abundance. One typical strategy that leverages molecular traits to decipher mechanisms at GWAS loci is to examine whether genetic regulators for phenotypic traits overlap those for molecular traits. This strategy has inspired a growing body of research looking for genetic regulators of molecular traits in human tissues. Large collaborative efforts such as GTEx, BLUEPRINT[21], and SCALLOP[22] have been established to link the genetic variants to splicing, histone modification peaks, gene expression, methylation and protein levels. The genetic variants associated with the molecular trait levels across individuals are termed molecular quantitative traits loci (QTL). In particular, QTLs for gene expression levels are termed eQTLs.

Type 2 diabetes (T2D) is a complex disease that accounted for 4.2 million deaths around the world in 2019 according to the International Diabetes Federation (IDF) consortium[23]. Globally about 1 in 11 adults has diabetes mellitus, with 90% of them belonging to T2D[24].

T2D develops when pancreatic islets fail to secrete enough insulin to compensate for the increased demand of insulin mainly driven by the insulin resistance in peripheral tissue such as skeletal muscle and subcutaneous adipose tissues[25]. High throughput molecular profiling in these T2D-relevant tissues has started to reveal molecular traits, such as genes and DNA methylation sites, that cause or respond to T2D and relevant physiological changes. In skeletal muscle tissue, Scott et al.[26] and Taylor et al.[27] have identified mRNAs and DNAme sites whose levels are associated with T2D and relevant traits as well as those that overlap T2D GWAS loci. In subcutaneous adipose tissue, Mete et al.[28] and Raulerson et al.[29] have identified QTLs of mRNAs and microRNAs(miRNAs) that colocalized with T2D. Nilsson et al. has identified mRNAs with differential expression levels and sites with differential DNA methylation levels between diabetic patients and non-diabetic controls[30]. The GTEx study has identified mRNAs with eQTLs colocalized with T2D in both skeletal muscle tissue and subcutaneous adipose tissues[31].

In chapter two, I present my work with the Finland-United States Investigation of NIDDM (FUSION) tissue biopsy study, a study that aims to understand the molecular basis of T2D. T2D is caused by genetic risk factors at many loci in combination with environmental factors[32]. To date, the largest T2D meta-analysis in individuals of European ancestry (n=898,130) has identified 403 distinct signals in 243 loci that increase susceptibility to T2D[33], and the vast majority of the signals are outside of coding region[34], [35]. To aid in elucidating the molecular mechanisms underlying the T2D GWAS loci and advancing the understanding of T2D etiology, the FUSION tissue biopsy group collected skeletal muscle[26], [27] and subcutaneous adipose tissue biopsies from up to 331 Finnish participants along with T2D-relevant physiological traits data (e.g., BMI, fasting serum insulin, and fasting glucose). The FUSION tissue biopsy study also generated genotype and molecular profiling (mRNA-sequencing, miRNA-sequencing, and DNA methylation array) data from these samples. With these rich datasets, I identified quantitative trait loci (QTL) for the mRNA and miRNA expression and DNA methylation levels and pinpointed those that potentially underlie T2D GWAS loci. We also identified genes and methylation sites associated with 48 T2D-relevant physiological traits. My contributions to this project in-

clude 1) participating in data processing and quality control, 2) conducting QTL detection and colocalization analyses, 3) conducting the physiological trait association analysis in collaboration with Anne Jackson and 4) writing the manuscript and creating the figures.

Cardiometabolic disease (CMD) is a category of complex diseases characterized by insulin resistance, impaired glucose tolerance, atherogenic dyslipidemia, hypertension, and intra-abdominal adiposity[36]. CMD includes diseases such as T2D, obesity, and cardiovascular diseases (CVD) and is the leading cause of mortality across the world[37]. Although traditionally visceral adipose has received the most attention in terms of its role in the pathophysiology of obesity and relevant metabolic disorders[38], increasing interest has been attracted to subcutaneous adipose tissue. Subcutaneous adipose tissue exists in a larger amount than visceral adipose tissue[39]. Different depots of subcutaneous adipose tissue may act in a coordinate or compensatory manner in disease development[40]–[42].

There has been a growing interest to integrate expression QTL (eQTL) with the CMD GWAS signals to prioritize potential genes involved in the mechanisms that contribute to disease susceptibility[43] in CMD-relevant tissues, such as subcutaneous adipose tissue. Many genes are regulated by more than one eQTL[29], [31]. Conditional analysis is commonly used to identify multiple eQTLs with independent effects on a given gene. The eQTLs displaying the strongest statistical evidence for associations in a locus without conditioning on any other genetic variants are considered primary eQTLs. The eQTLs that show statistical significance after adjusting for the previously selected eQTL variants are considered secondary eQTLs.

To my knowledge, seven single-study eQTL analyses have identified genome-wide eQTLs in human subcutaneous adipose tissue with sample sizes ranging from 63 to 855[28], [29], [31], [44]–[47]. These sample sizes are relatively small compared with sample sizes of whole blood studies, a more accessible tissue type[48]. To my knowledge, at least five single-cohort whole blood eQTL studies exceed the sample size of one thousand[49]–[53]. Combining data across studies through meta-analysis is a commonly used approach to

achieve greater power for eQTL studies. Meta-analysis of cerebral cortical eQTL from 1433 samples[54], cerebellar eQTL from 261 samples[54], and whole blood eQTL from 31,684 samples[55] have improved the discovery of genes with $\geq$ 1 eQTL in corresponding tissues. Using colocalization analysis, Raulerson et al.[29] and the GTEx Consortium[31] identified genes that may underlie CMD GWAS loci. Notably, Raulerson et al.[29] identified 21 genes whose secondary eQTLs, not primary eQTLs colocalized with CMD GWAS loci.

In chapter three, I describe my work in a collaborative research group to perform RNA-seq based eQTL meta-analysis from 2256 individuals of European ancestry in subcutaneous adipose, combining eQTL associations from TwinsUK, METSIM, GTEx, and FUSION studies. Of the 19,108 genes present in all studies, the meta-analysis revealed $\geq$ 1 eQTL for 15335 (80.3%) of the genes: 6440 (33.7%) genes had exactly one eQTL, 8895 genes (46.6%) had ≥ 2 eQTL eQTLs. Integrating the conditional eQTLs identified in the meta-analysis with the GWAS signals for seven cardiometabolic traits: T2D, Body mass index, Waist-hip ratio, BMI adjusted waist-hip ratio, Coronary artery disease, fasting glucose and fasting insulin, I identified 517 genes that had eQTLs colocalized with GWAS signals. My role in the meta-analysis was to perform *cis*-eQTL detection within FUSION, to compare software, to perform conditional eQTL meta-analyses and colocalization analyses.

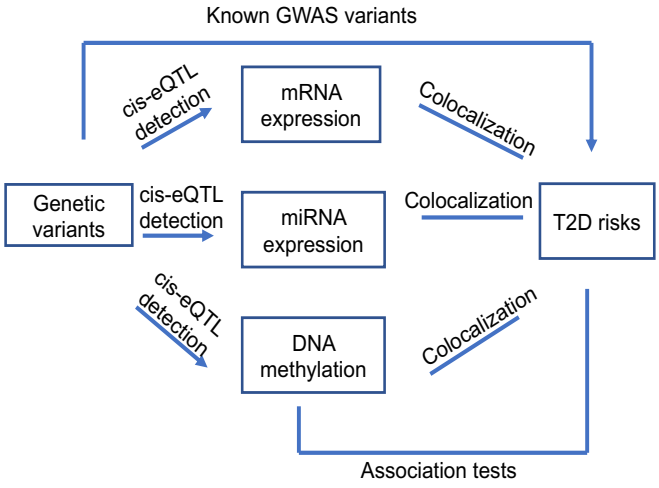
Throughout my dissertation, I used data sets of different omics data types and from multiple studies to help generate hypotheses for the biological mechanisms underlying the genetic predisposition to complex diseases. (Figure1.0.1).

In chapter four, I discuss the limitations of my work and future directions to further our knowledge of the regulatory mechanisms that link genetic risk loci to complex disease susceptibility.

Figure 1.0.1. Dissertation overview

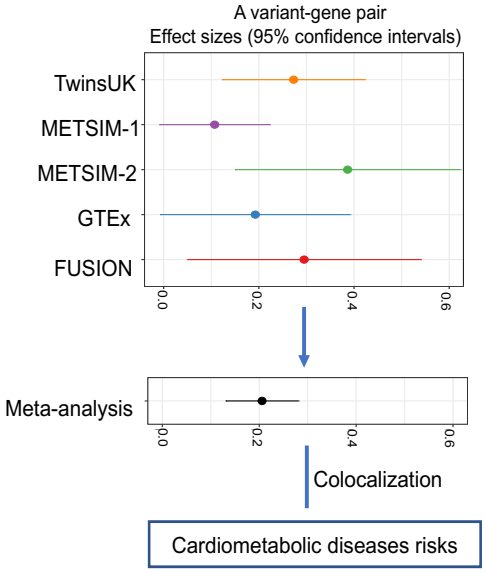
Chapter 2

What are the molecular traits through which genetic variants increase T2D risk?



What are the molecular traits that cause or respond to T2D status and T2D relevant physiological trait levels?

Chapter 3



Using eQTLs identified in a meta-analysis, what are the genes through which genetic variants increase the risk for cardiometabolic diseases?

Chapter 2

mRNA, miRNA and DNA Methylation Levels Associated with Fasting Serum Insulin and Type 2 diabetes in Human Skeletal Muscle and Subcutaneous Adipose Tissues

2.1 Introduction

Type 2 diabetes (T2D) affects 422 million adults worldwide, and its prevalence is predicted to rise even further in the next decades, presenting a tremendous clinical, economic, and social burden[56]. T2D is characterized by the inability of the beta cells to secrete enough insulin to overcome insulin resistance in peripheral tissues[57]. As with most complex diseases, an individual's risk of developing T2D is modulated by interactions between genetic and environmental factors[32] through the interplay of brain and several peripheral tissues, including pancreatic islets, skeletal muscle tissue, and subcutaneous adipose tissue[58]. Skeletal muscle is responsible for the majority (> 80%) of insulin-stimulated whole-body glucose disposal[59]. Adipose not only stores large amounts of energy in the form of triglycerides but also acts as an endocrine organ by secreting various hormones and cytokines with effects on glucose, lipid metabolism, and energy homeostasis.[60]

Despite the progress in our understanding of T2D pathophysiology[61], how genetic and environmental factors exert effects in disease-relevant tissues to increase T2D risk and how human biological systems respond to the perturbations are far from being completely understood. Omics technologies, which profile molecular traits in a high throughput man-,

ner have provided an unprecedented opportunity to study T2D pathophysiology at the molecular level, e.g., genes and DNA methylation (DNAme) sites. A growing body of research has identified genetic variants associated with molecular trait levels, termed molecular QTLs. Integrating molecular QTLs in T2D relevant tissues with risk loci identified by genome-wide association studies (GWAS) has begun to provide insights into biological mechanisms underlying the genetic predisposition to T2D. In skeletal muscle tissue, several studies have identified QTLs for gene expression levels [26], [27], [62] and DNAme levels[27]. Scott et al.[26] and Taylor et al.[27] have also identified QTLs for gene expression and DNAme levels that may overlap T2D GWAS loci. In subcutaneous adipose tissue, a few studies have investigated QTLs for genes [28], [29], [44], [45], [63], including one study[63] that has identified QTLs for microRNA expression levels. Civelek et al.[28] and Raulerson et al.[29] have also identified genes with subcutaneous tissue adipose QTLs that colocalized with cardiometabolic traits including T2D. Drong et al.[64] and Volkov et al.[65] have studied subcutaneous tissue adipose QTLs for DNAme levels, and Volkov et al.[65] has also used their identified mQTLs to identify methylation sites that may mediate genetic risk to metabolic traits. The STAGE[46] and STARNET[47] studies have identified eQTLs in skeletal muscle, subcutaneous adipose as well as five other tissues. The GTEx study[31] has surveyed eQTLs across 49 tissues and identified those that colocalized with a range of complex traits including T2D.

Most published omic studies in human skeletal muscle and subcutaneous adipose tissues focused on a particular type of molecular traits, with mRNA most often studied. To our knowledge, no population-based microRNA(miRNA) study has been performed in human skeletal muscle tissue to investigate the miRNAs involved in T2D mechanisms.

Within the Finland-United States Investigation of NIDDM Genetics (FUSION) Tissue Biopsy Study, we have collected skeletal muscle and subcutaneous adipose tissue biopsies from a cohort of 331 living donors spanning from normal glucose tolerance to newly-diagnosed diabetes. We performed mRNA- and miRNA- sequencing and DNA methylation arrays in skeletal muscle and subcutaneous adipose tissue samples and array-based genotyping in whole blood samples. Using these data, we aimed to (1) identify genes (mRNAs and

miRNAs) and DNAme sites that potentially underlie T2D GWAS loci by integrating T2D GWAS signals and molecular QTLs; (2) identify genes and methylation sites associated with 48 T2D-relevant physiological traits.

2.2 **Methods**

2.2.1 **Blood sample genotyping and genotype imputation**

Blood sample genotyping, quality control (QC), and genotype imputation are described in Taylor et al.[27]. In brief, we genotyped 331 subjects(Supplementary Table2.8.1) on the HumanOmni2.5- 4v1_H or InfiniumOmni2-5Exome-8v1-3 BeadChip arrays (Illumina, San Diego, CA, USA). We mapped the array probe sequences to the hg19 genome assembly. We examined the relatedness of samples using KING[66] and identified two pairs of first-degree relatives. We removed one sample from each pair of the first-degree relatives from genotype, mRNA-seq, miRNA-seq, and DNA methylation array data. To assess ancestry, we compared the estimated genetic principal components (PCs) to the Population Reference Sample (POPRES) European reference panel[67] and removed one non-Finnish participant (Table2.2.1). The subcutaneous adipose sample from this non-Finnish participants was unintentionally used twice in the DNA methylation array experiments. We removed the samples from the non-Finnish participant in the QC process of mRNA-seq, miRNA-seq, and DNA methylation array. A total of 328 array-genotyped samples remained for analysis. We imputed genotype dosages to the Haplotype Reference Consortium (hrc.r1.1.2016, build GRCh37/hg19) panel[68] using Minimac3[69]. We included in analyses 6.9M genetic variants with imputation quality score $R^2 > 0.3$ and MAF $\geq 2\%$ over the 328 samples for autosomes and chromosome X.

	Number of samples
One sample from each of the two first-degree relative pairs	2
Non-Finnish participants	1
Total samples passed / total samples submitted	328/331

Table 2.2.1. Number of array-genotyped samples sequentially excluded in each QC step

2.2.2 **Tissue biopsy**

We collected tissue biopsies from participants who 1) had not undergone drug treatment for diabetes, 2) were not on daily medication or on medications that increase haemorrhage risk or on medication that might confound the analyses, 3) did not have diseases that increase haemorrhage risk or might confound the analyses. Detailed participants exclusion criteria are described in Scott et al.[26] We took the tissue biopsies at clinical visits after 12 h fast and 24 h avoidance of strenuous exercise.

Muscle We surgically collected skeletal muscle biopsies from the vastus lateralis muscle from 327 of the 331 participants, following the procedures described in Scott et al.[26].

Adipose We surgically collected subcutaneous adipose tissue biopsies from the abdomen from 329 of the 331 participants concurrently with the skeletal muscle biopsy following the same general protocol. We took subcutaneous adipose biopsies from abdominal subcutaneous adipose tissue 5-10 cm lateral of the umbilicus with a surgical scalpel under local anesthetic without adrenalin.

Processing We visually dissected each frozen tissue biopsy into two pieces, one piece (30-50 mg skeletal muscle tissue, 100-150 mg subcutaneous adipose tissue) for RNA extraction (mRNA and miRNA sequencing) and the other piece (about 25 mg) for DNA extraction(DNA methylation array).

2.2.3 **RNA isolation, mRNA sequencing, and QC**

Muscle RNA integrity number (RIN) of skeletal muscle tissue biopsies ranged from 6.6 to 9.4 (median 8.4). The procedures for RNA isolation, polyA selection, sequencing, and QC were described[26], [27]. skeletal muscle samples removed from each QC step are shown in Table2.2.2. Of the 323 samples in which RNA expression levels were measured, 301 samples passed QC.

Adipose We measured RNA expression levels in 296 samples. We visually dissected 100-150 mg of each frozen biopsy sample avoiding vascularized regions. We extracted RNA as described for skeletal muscle tissue samples. Subcutaneous adipose tissue sam-

ple RIN ranged from 5.1 to 8.8 (median 7.4). We followed the same procedures of RNA extraction, mRNA-seq, processing, and quality control (QC) as described in Scott et al.[26]. We excluded six subcutaneous adipose samples, which were extreme outliers in their read coverage at the 3' end of gene bodies based on QC summary plots created by QoRTs v1.1.18[70]. To analyze the cumulative gene diversity, we first calculated the cumulative fraction of reads as a function of genes sorted by read count for each subcutaneous adipose sample. Then we compared the distribution of each sample to the distribution of median read count using the Kolmogorov-Smirnov test (ks.test function in R). We removed five subcutaneous adipose samples with p-values < 0.01. We compared the allelic RNA-seq read count distribution to known sample genotypes using verifyBamID[71] and identified two contaminated samples and one pair of sample swaps. We removed the two contaminated samples and assigned the swapped samples to the correct donors based on genotyping results. We verified the reported sex of the remaining samples using *XIST* gene expression and the mean Y chromosome gene expression.

We sought to remove outlier samples based on PCA. We performed linear regression of gene expression (Transcripts per million, TPM) as a function of age, sex, batch, and RIN. We performed PCA on the gene expression residuals[27]. We selected the first two principal components (PCs) that explained 20% of the variance in gene expression and transformed the two PCs to z-scores; No sample had a |z-score| > 5.

Description	Muscle	Adipose
One sample from each of the two first-degree relative pairs	2	2
Non-Finnish participants	1	1
Extreme 3’ bias in gene body coverage	4	6
Outliers in transcriptional diversity	7	5
Contaminated with a different sample	1	2
Outliers based on within-tissue expression PCA	0	0
One sample from each of the intentionally duplicated pairs	7	Not applicable
Total samples passed / total samples submitted	301/323	280/296

Table 2.2.2. Number of mRNA-seq samples sequentially excluded in each QC step

Samples removed from each QC step are shown in Table2.2.2. After the QC steps, 280 unique subcutaneous adipose tissue RNA-seq samples remained for analysis.

2.2.4 **DNA isolation, methylation quantification, and QC**

Muscle We isolated DNA, quantified DNA methylation levels using the Illumina Infinium HD Methylation Array with Infinium MethylationEPIC BeadChips, and performed QC as described in Taylor et al.[27].

Adipose We measured DNA methylation levels in 299 subcutaneous adipose tissue samples. We isolated DNA, quantified methylation levels, and performed QC in the same way as for muscle samples[27]. In brief, we calculated the two widely used metrics to measure methylation levels (beta-values and M-values) using the Illumina normalization method implemented in minfi v1.20.2915 with default parameters. Beta-value is the ratio of the methylated probe intensity and the overall intensity. M-value is the log2 ratio of the intensities of methylated probes versus unmethylated probes.The QC metrics used below are described in detail in Taylor et al.[27]. (1) For each probe, we calculated a detection p-value which compared the combined raw methylated and un-methylated signals to the background noise. A probe with detection p-value > 0.05 was defined as a low-quality probes[72], [73]. We excluded three samples for which > 1% of probes had detection p-values > 0.05[72], [73]. (2) We computed the median signal intensity of the methylated and un-methylated signals per sample. We excluded two samples with median methylated and/or un-methylated signals < 10[70]. (3) We excluded eight samples with evidence of multiple outlying probe signals for ≥ 1 type of control probes designed to capture different technical aspects (e.g., hybridization efficiency, staining)[74]. (4) We excluded one sample whose genotypes assayed by the EPIC array were not consistent with the expected dosages (based on the array-and-imputation based dosages) for the 47 variants designed to detect common variants on the EPIC array. (5) We excluded two samples with outlying M-value DNAme distributions identified by comparing the M-value percentiles for each sample to the median M-value distribution using the Kolmogorov-Smirnov test (p-values <0.01). (6) We excluded two subcutaneous adipose samples that did not cluster with the other subcutaneous adipose samples in the PCA of the M-values across a dataset from multiple tissues, including skeletal muscle, subcutaneous adipose, EndoC-β H1 and whole blood samples. (7) We excluded one within tissue outlier in the PCA of the M-

values. We verified the reported sex of the remaining samples using the X chromosome DNAme. Samples removed from each step are shown in Table2.2.3. After the QC steps, 276 unique adipose DNAme samples remained for analysis.

Description	Muscle	Adipose
One sample from each of the two first-degree relative pair	2	2
Non-Finnish participants	1	2 (Including one unintentional duplicate)
Failed low-quality probe filter	5	3
Outliers in the median methylated and unmethylated plot	1	2
Outliers in control probe	3	8
No clear genotype match	4	1
Outliers in methylation distribution	1	2
Outliers based on multiple-tissue PCA	3	2
Outliers based on within-tissue PCA	1	1
Total samples passed/total samples submitted	282/303	276/299

Table 2.2.3. Number of DNA methylation array samples sequentially excluded in each QC step

2.2.5 **miRNA sequencing**

We measured miRNA expression levels for 296 skeletal muscle and 270 subcutaneous adipose tissue samples. The total RNA isolated for mRNA-sequencing was also used for miRNA isolation and sequencing. miRNA libraries were prepared at the NIH Intramural Sequencing Core (NISC) from 1 µg total RNA using Illumina's TruSeq Small RNA Library Kit according to the manufacturer's guidelines, except a 10% acrylamide gel was used to better separate the library from adapters. Libraries were pooled in groups of four to eight for gel purification. Single-end 51-base sequencing was performed on Illumina HiSeq 2500 sequencers in Rapid Mode using version 2 chemistry. We mapped miRNA sequence reads using the exceRpt[75] pipeline (v4.4.0) with default parameters. We counted reads mapped to each miRNA of miRBase (version 21) [76] and quantified miRNA expression using reads per million mapped to miRNAs (RPMMM).

For each of the two pairs of quality control duplicate samples of skeletal muscle and subcutaneous adipose tissue, we retained the sample from the tissue piece used in the mRNA analyses. Because the same RNA extracts were used for both mRNA-seq and miRNA-seq, we excluded samples identified as contaminated in mRNA-seq (one skeletal muscle

tissue and two subcutaneous adipose tissue). We assessed the quality of each miRNA-seq dataset through metrics generated by exceRpt[75], including read length and library size, and did not observe outliers. Samples removed from each step are shown in Table2.2.4. After the QC steps, 290 skeletal muscle tissue miRNA-seq and 263 subcutaneous adipose tissue miRNA-seq remained for analysis.

Description	Muscle	Adipose
One sample from each of the two first-degree relative pairs	2	2
Non-Finnish participants	1	1
Contaminated with a different sample	1	2
One sample from each of the duplicated pairs	2	2
Total samples passed/total samples submitted	290/296	263/270

Table 2.2.4. Number of micro RNA-seq samples sequentially excluded in each QC step

After QC, the sample size for each data type is shown in Table2.2.5.

2.2.6 **Marginal *cis*-QTL analysis**

I scanned for *cis*-QTLs from variants that reside within 1 Mb of the gene transcription start site. To account for unknown biological and technical factors that may contribute to the measured expression level of a molecular trait, we performed factor analysis of the inverse normalized mRNA or miRNA expression levels or the inverse normalized M-values of DNAme sites via PEER v1.0[77]. I used the inferred PEER factors as covariates in QTL mapping.

I used a linear regression model with an additive genetic effect, adjusting for the first four genotype PCs (Eigenstrat p-value <0.1[27], [78]) and a specified number of PEER factors. To optimize the discovery of molecular traits with a QTL, I assessed various numbers of PEER factors. For mRNA and DNAme sites, I assessed 0 to10 with an increment of 1 PEER factor, and 10 to 80 with an increment of 5 PEER factors. For miRNA, I assessed 0 to 10 PEER factors with an increment of 1 PEER factor, and from 10 to 50 PEER factors

Number of samples	Genotyping	mRNA-seq		DNAme array		miRNA-seq	
		Muscle	Adipose	Muscle	Adipose	Muscle	Adipose
Submitted	331	323	296	303	333	296	270
Passed QC	328	301	280	282	276	290	263

Table 2.2.5. Sample sizes for each data type and tissue

with an increment of 5 PEER factors. I used as covariates the largest number of PEER factors that resulted in ≥ 1% increase in the number of molecular traits with a significant QTL, compared to the previous number of PEER factors (Supplementary Figure2.7.1). The number of PEER factors used in identifying QTLs for mRNA, miRNA and DNAme sites are shown in Table2.2.6.

For the most significant variant of a given gene, I used the approximate permutation analysis from QTLtools[79] to calculate a p-value accounting for all tested variants for that gene. I approximated a permutation based p-value distribution using a beta distribution fit with 1,000 permutations for the PEER factor analysis and 10,000 permutations for the final *cis*-eQTL detection. I applied the Storey-Tibshirani FDR[80] to the most significant variant for each gene to account for the number of genes tested with a threshold of FDR $\leq$ 1%. I used the same framework to conduct *cis*-QTL analysis for inverse normalized M-values of DNAme sites or expression levels of miRNA. QTLs for mRNA, miRNA, and DNAme sites are denoted as eQTL, miR-eQTL and mQTL, respectively.

	mRNA QTL	miRNA eQTL	DNAme sites QTL
Muscle	45	5	20
Adipose	45	8	15

Table 2.2.6. Number of PEER factors that maximized QTL discovery

2.2.7 **Filtering mRNA, miRNA and DNA methylation data to increase power to detect associations**

To maximize the power to detect molecular trait-phenotype and molecular trait-genotype associations, I assessed the power to detect expression QTLs (eQTL) and methylation QTLs (mQTL) at different thresholds of gene expression levels and of methylation variation, respectively. Separately for mRNA and miRNA, I ordered the expressed genes (mean read count > 0) from lowest to highest mean gene expression level and partitioned them into equal-size bins (each bin had the same number of genes, except for the last bin). I partitioned the 50K mRNAs into 100 bins, where each bin had about 500 mRNAs. As there are many fewer miRNAs, to avoid a low number of miRNAs in each bin, I partitioned

the 2K mRNAs into 20 bins, where each bin had about 100 miRNAs. For DNAme sites, I randomly selected 10% of the sites and ordered them from lowest to highest variance of beta-values, and partitioned the sites into 200 equal-sized bins. Next, I calculated the proportion of genes or DNAme sites with QTLs within each bin using a false discovery rate of 5% with the Benjamini-Hochberg procedure[81]. Next, I evaluated the impact of different filtering thresholds on the detected number of genes or DNAme sites with QTLs. I investigated the effect of setting the threshold for inclusion at the first bin in which the proportion of genes or DNAme sites with QTLs > 0% up to a threshold where half of the genes or DNAme sites were included. I chose the bin threshold that maximized the detected number of genes or DNAme sites with QTLs. The numbers of genes or DNAme sites included are shown in Table2.2.7.

mRNA		miRNA		DNAme	
Muscle	Adipose	Muscle	Adipose	Muscle	Adipose
31,518	34,120	836	950	699,825	700,333

Table 2.2.7. Number of genes and DNAme sites included in analyses

2.2.8 **Comparison of the power to detect QTLs at different thresholds of gene expression levels between mRNAs and miRNAs**

I assessed whether the power to detect QTLs differed between mRNAs and miRNAs at various read count levels. I ordered the mRNAs included in the analysis from lowest to highest mean read count and partitioned the mRNAs into 100 equal-sized bins. Within each bin, I calculated the proportion of mRNAs with a QTL using a false discovery rate of 5% with the Benjamini-Hochberg procedure. Next for each mRNA, I fitted a smooth spline between an indicator variable for it having an QTL and its read count using a general additive model. I applied the same analysis to miRNA, except that I partitioned the miRNAs into 20 equal-sized bins.

2.2.9 **Multiple independent QTL analysis**

I evaluated whether a given mRNA, miRNA or DNA methylation site had more than one independent QTL signal under a Bayesian fine-mapping framework using the Deterministic Approximation of Posteriors (DAP) algorithm[82]. I detected multiple independent QTLs for each mRNA, miRNA or DNAme site with a marginal *cis*-QTL, considering all variants

within the *cis*-region (1Mb). DAP computes the posterior probabilities of association models with different numbers of genetic variants and then calculates the posterior inclusion probability (PIP) for each variant. I created a 95% credible set of potential causal variants for each independent QTL of each molecular trait using the resulting PIP.

2.2.10 Colocalization of QTLs for genes and DNAme sites with T2D GWAS loci

I identified mRNAs, miRNAs, and DNAme sites that may underlie the genetic associations for T2D discovered in individuals of European ancestry[33]. I performed colocalization analysis using fastEnloc[83], [84] between the marginal association of each variant-T2D association and each independent QTL located within 1 Mb from a given molecular trait. FastEnloc first estimates the enrichment of molecular QTLs in the GWAS loci and then assesses the colocalization probability of a given molecular QTL overlapping a GWAS variant by calculating the variant-level colocalization probability (SCP). FastEnloc sums up the SCPs of correlated variants within an LD block to create regional colocalization probability (RCP), representing the probability of a genomic region having a colocalized signal.

2.2.11 Chromatin state and open chromatin profiling via assay for transposase-accessible chromatin (ATAC-seq) data sources

We used the chromatin states for a total of 31 tissues and cell types described in Varshney et al.[85]. We used the processed ATAC-seq data in skeletal muscle and subcutaneous adipose tissue generated in Scott et al.[26] and Cannon et al.[86], respectively. The muscle ATAC-seq experiment was performed using frozen human skeletal muscle (Zen-bio, Durham, NC USA). The subcutaneous adipose tissue ATAC-seq experiment was performed using a biopsy from a Finnish donor participating in the METabolic Syndrome in Men (METSIM) study.

2.2.12 Cell culture

The human preadipocyte cell strain derived from the subcutaneous adipose tissue of a patient with Simpson–Golabi–Behmel syndrome (SGBS)[87] was generously provided by Dr. Martin Wabitsch (University of Ulm) and cultured in basal medium consisting of DMEM/F12

(Corning) with 10% FBS and 33 µM biotin/17 µM panthotenate. To differentiate SGBS cells, we incubated cells in serum-free basal medium supplemented with 10 µg/ml transferrin, 20 nM insulin, 200 nM cortisol, 400 pM T3, 50 nM dexamethasone, 500 µM IBMX and 2 µM rosiglitazone. We maintained cell lines at 37° C with 5% CO_2.

2.2.13 **Transcriptional reporter assays**

To test allelic differences in transcriptional activity, we designed PCR primers (5'-TCTGGG CTCTTTCCAGTTTG and 5'–TCCTCATGGGTCAAGATGGT) with KpnI and XhoI restriction sites to amplify a 610-bp genomic region (chr2: 121347212 - 121347821) containing rs11688682, using DNA of individuals homozygous for each allele. As described previously[88], we cloned the restricted PCR amplicons into the multiple cloning site of the firefly luciferase reporter vector pGL4.23 (Promega, Fitchburg, Wisconsin/USA) in both orientations with respect to the promoter. Five independent clones were isolated and sequence-verified for each allele of each orientation. SGBS cells were seeded (40,000 cells per well) in 24-well plates and co-transfected with pGL4.23 constructs and phRL-TK Renilla luciferase reporter vector (Promega) in triplicate using Lipofectamine 3000. Twenty-eight hours after transfection, we measured the luciferase activity using the Dual-Luciferase® Reporter Assay System (Promega). We first normalized firefly luciferase activity to Renilla luciferase activity, and then normalized to the average of two empty pGL4.23 vectors. All experiments were carried out on a second independent day and yielded comparable results. We compared differences in luciferase activity between clones with G or C allele using unpaired two-sided t-tests.

2.2.14 **Estimation of variation in molecular profiling data likely driven by tissue/cell type composition heterogeneity**

We used two approaches to estimate the variation in the molecular profiling data due to tissue/cell type composition heterogeneity. 1) We estimated the proportions of constituent or contaminating tissue or cell types based on mRNA-seq data, using external reference mRNA-seq reference datasets. 2) We estimated surrogate variables (SVs) for mRNA-seq, miRNA-seq, and DNAme array separately using dSVA[89], which was designed to capture

the variability caused by unknown technical or biological factors for each physiological trait while protecting the effects of the physiological trait of interest.

For skeletal muscle tissue, we estimated the proportions for five tissue/cell types ("skin not sun exposed suprapubic", "whole blood", "adipose subcutaneous", "muscle skeletal", and "EBV transformed lymphocytes") using GTEx v7 (phs000424.v7.p2) mRNA expression profiles as references. We estimated the proportions for three muscle fiber types using the percentage of the expression levels of myosin heavy chain gene (*MYH1*, *MYH2*, and *MYH7*) as proxies for muscle fiber types, as previously described[27].

For subcutaneous adipose tissue, we computed two sets of estimates of tissue/cell type proportions. The first set had adipocyte, T cell, microvascular endothelial cell, macrophage, and blood (denoted as five-component estimates). The second set had endothelial cell, adipocyte, preadipocyte, B cell, lymphatic endothelial cell, fibroblast, M1-M5 macrophage, mast cell, neutrophil, perivascular cell, naive T cell, natural killer cells, and blood (denoted as 17-component estimates).

Five-component estimate We created a reference transcriptome by downloading raw fasta files of whole blood (GEO accession GSE67488), and raw fasta files of cell types present in subcutaneous adipose tissue (adipocytes, macrophages, CD4+ T cells, and microvascular endothelial cells) used in Glastonbury et al.[90]. We aligned the RNA-seq reads to the hg19 reference transcriptome using the same read mapping and quality control procedure as used for the FUSION mRNA-seq data[26], [27]. We estimated the tissue/cell-type proportions for each FUSION subcutaneous adipose tissue sample using the unmix function from DESeq2 v1.18.1[91].

17-component estimate For each cell type, we obtained a set of the predefined cell type marker genes from the single-nuclei cell data and the average log fold change of each marker gene (a measure of enrichment of this gene in this cell type) from Paivi Pajukanta (personal communication). Using the cell type marker genes, average log fold change of each marker gene, and FUSION bulk subcutaneous adipose tissue expression levels, we estimated a first principal component that represents the relative amount of each cell type

in FUSION subcutaneous adipose tissue samples.

2.2.15 **Molecular trait (mRNA, miRNA, DNAme) association with physiological traits**

We tested for associations of 48 physiological traits (T2D status and 47 continuous traits) with each molecular trait. We had $\leq$ 200 samples for eight physiological traits, T2D, Hemoglobin A1c, plasma insulin levels at four time-points (fasting, 30min, 60min and 120min) and two Matsuda index measurements (Table2.8.3). Compared to the other 40 physiological traits, we had less power to test for physiological-molecular trait associations for these eight physiological traits.

We tested for association of each inverse normalized quantitative trait with inverse normalized mRNA expression using a linear regression model with a base set of covariates, age, sex, RIN, TIN, batch, sample collection site, smoking status, median insert size and mean GC content. We also adjusted for either the estimated tissue/cell type proportions or physiological trait-specific surrogate variables in the model. We tested for association of T2D with inverse normalized mRNA expression using a logistic regression model, adjusting for the same set of covariates. For each physiological trait, we corrected for the number of tested genes or DNAme sites using the Benjamini-Hochberg procedure[81].

We separately applied the same analysis to inverse normalized M-values of DNAme sites and to the inverse normalized expression levels of miRNA, except for using a different base set of covariates. For DNAme site-physiological trait association analysis, the base set of covariates had age, sex, plate, sentrix position, plate position, sample collection site, smoking status. For miRNA-physiological trait association analysis, the base set of covariates had age, sex, plate, batch, RIN, sample collection site, smoking status.

2.3 **Results**

2.3.1 **Gene and DNAme QTLs**

Identifying QTLs for gene expression and DNA methylation levels may improve our understanding of the genetic control of gene expression and DNA methylation, and has the potential to unravel the molecular mechanisms that contribute to disease susceptibility.

We focused on discovering *cis*-QTLs, defined as QTLs residing within 1 Mb of the gene transcription start site for mRNA, start positions of the precursor miRNA for miRNA, and start position of the DNAme sites.

I mapped *cis*-QTLs for gene/DNAme sites in skeletal muscle and subcutaneous adipose tissues separately, controlling for genetic population structure using genotype PCs and tissue/cell-type composition effects using PEER factors. I identified 10,736 of 31,518 mRNAs (34.1%), 125 of 836 miRNAs (15.0%), and 147,899 of 699,825 DNAme sites (21.1%) with $\geq$ 1 QTL in skeletal muscle tissue at an FDR < 0.01. I identified similar fractions of genes/DNAme sites with $\geq$ 1 QTL in subcutaneous adipose tissue(Table2.3.1). In both skeletal muscle and subcutaneous adipose tissues, I detected smaller proportions of QTLs for miRNA (15.0% and 16.8% respectively) than for mRNA (34.1% and 35.4%, respectively).

Categories	mRNA		miRNA		DNAme	
	Muscle	Adipose	Muscle	Adipose	Muscle	Adipose
Genes/DNAme sites (N)	31,518	34,120	836	950	699,825	700,333
Genes/DNAme sites with ≥ 1 QTL (N)	10,736	12,068	125	159	147,889	125,122
Proportion of molecular traits with ≥ 1 QTL (%)	34.1	35.4	15.0	16.7	21.1	17.9

Table 2.3.1. Number of molecular traits with $\geq$ 1 QTL at 1% FDR threshold.

Of the tested mRNAs or miRNAs, I observed that 12.2% of the mRNAs in skeletal muscle tissue and 18.0% of the mRNAs in subcutaneous adipose tissue comprised 90% of reads mapped to mRNAs, whereas 2.8% of the miRNAs in skeletal muscle tissue and 3.0% of the miRNAs in subcutaneous adipose tissue comprised 90% of reads mapped to miRNAs (Supplementary Figure2.7.2). I asked whether the lower *cis*-QTL detection rate for miRNA than for mRNA was due to differences in the power to detect a QTL given transcript abundance. I examined the relationship between a gene's probability of having a QTL and its mean read count across samples. Within each tissue, at any fixed mean read count level (Figure2.3.1A; Figure2.3.2), a smaller proportion of miRNAs had detectable QTLs compared to mRNAs, suggesting a smaller proportion of variance in read counts (which may be biological variation or uncontrolled technical variation) in miRNAs was due to *cis*-eQTLs. In addition, I observed that mRNA QTL discovery rate was constant at mean read count level $\geq$ 100, whereas the miRNA QTL discovery rate was lower for miRNAs

with mean read count $\geq$ 100 (Figure2.3.1A; Figure2.3.2) than for miRNAs with mean read count < 100. To determine if the miRNAs with higher read counts had more constrained levels of expression than mRNA, I used the number of target mRNAs for each miRNA from TargetScan (computationally predicted targets)[92] and from TarBase (experimentally validated targets)[93]. Using the two resources, I observed that miRNAs with a larger mean read count had a higher number of target mRNAs than those with a lower mean read count (Supplementary Figure2.7.3). The broader regulatory impacts of miRNA with larger read counts suggests that they may be under a stronger selective pressure and therefore have a lower level of genetic regulation.

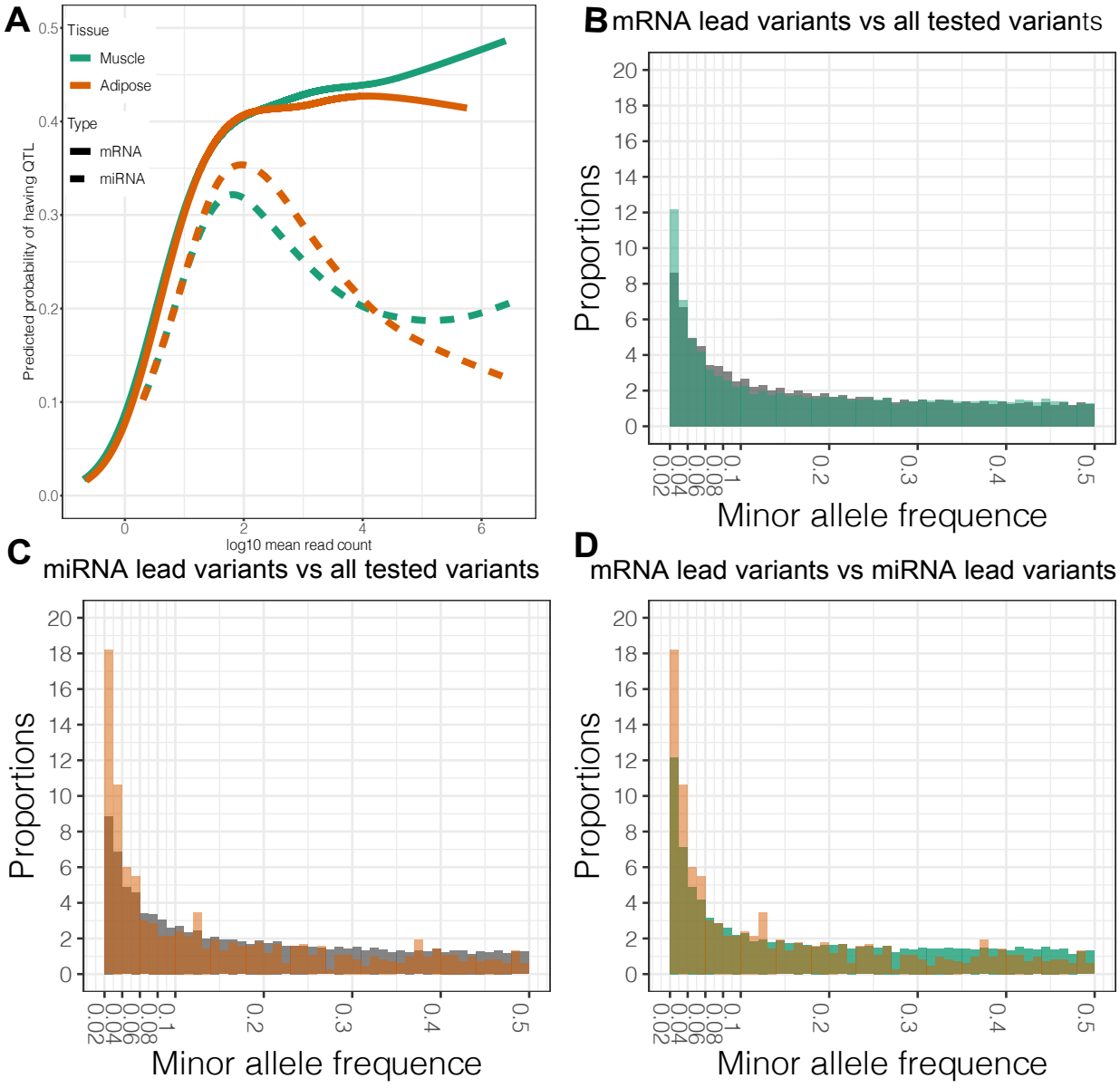


Figure 2.3.1. mRNA and miRNA *cis*-QTL discovery. (A). The probability of detecting a QTL (y-axis) as a function of the log10 mean read count for an mRNA or a miRNA (x-axis); (B). Skeletal muscle tissue mRNA QTL: distribution of minor allele frequencies (MAFs) of lead variants vs tested variants; (C). Skeletal muscle tissue miRNA QTL: distribution of MAFs of lead variants vs tested variants; (D). Skeletal muscle tissue: distribution of MAFs of lead miRNA vs lead mRNA variants. Bars show the proportions of variants within MAF bins.

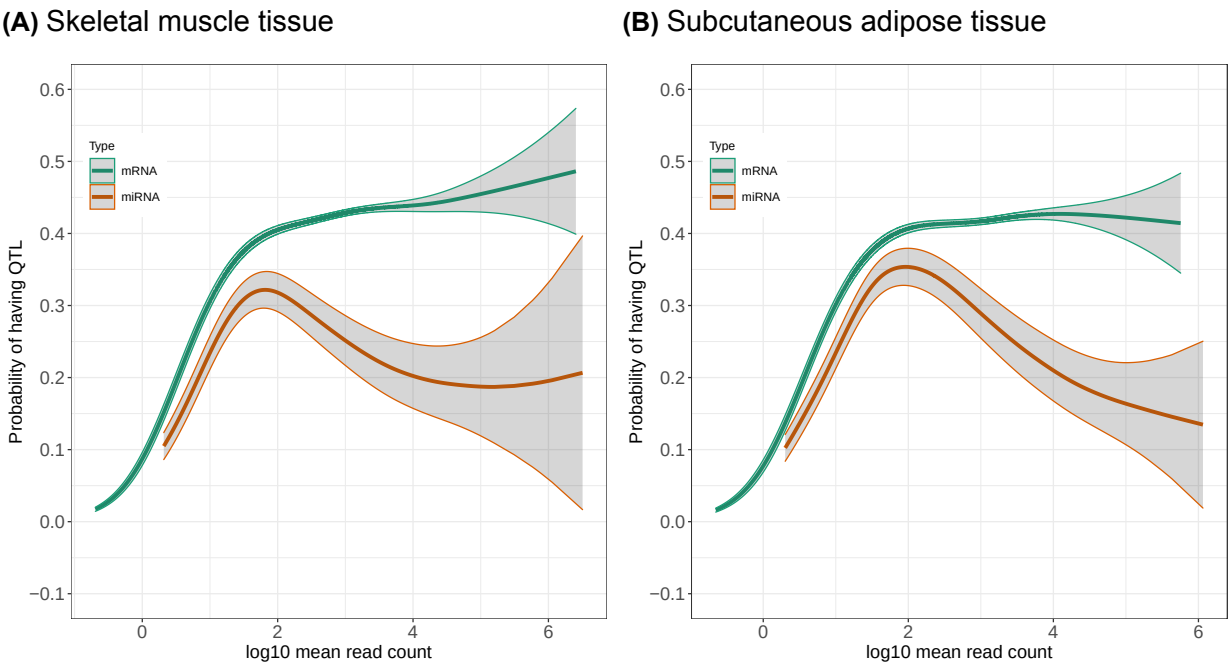


Figure 2.3.2. Scatter plots show the predicted probabilities of having QTLs (y-axis) as a function of log10 mean round counts (x-axis) of mRNAs (colored in green) and miRNA (colored in orange). Grey shaded areas represent the 95% confidence intervals around the predicted probabilities.

As fewer genetic associations were detected for miRNA than for mRNA, we hypothesized there might be stronger constraints on genetic variants affecting miRNA than mRNA. To compare the minor allele frequency (MAF) of the lead variants of mRNA and miRNA QTLs, I reran the *cis*-QTL mapping using the overlapping samples between miRNA and mRNA within each tissue (n=283 for skeletal muscle tissue and n=245 for subcutaneous adipose tissue) to avoid the bias introduced by sample size difference. In skeletal muscle tissue, for mRNA, the median MAF of tested variants and lead QTL variants were 0.159 and 0.156, respectively; for miRNA, the median MAF of tested variants and lead QTL variants were 0.157 and 0.0976. In subcutaneous adipose tissue, for mRNA, the median MAF of tested variants and lead QTL variants were 0.159 and 0.157; for miRNA, the median MAF of tested variants and lead QTL variants were 0.157 and 0.0776. The lead variants for mRNA (Figure2.3.1B) and miRNA QTLs(Figure2.3.1C) had lower MAF than the total set of tested variants (Wilcoxon rank sum test p-values for mRNA and miRNA were 5.2×10^{-8} and 2.2×10^{-16} in skeletal muscle tissue, 7.4×10^{-6} and 2.2×10^{-16}in subcutaneous

adipose tissue). In addition, lead variants for miRNA had lower MAF than lead variants for mRNA (Figure2.3.1D, Wilcoxon rank sum test p-value =2.6×10^{-5} in skeletal muscle tissue, 3.5×10^{-4} in subcutaneous adipose tissue). I did not use a p-value cutoff to select the lead variants as variants with lower MAF require a larger effect size to explain the same amount of genetic variance. This suggests that purifying selection may act on the genetic variants that influence gene expression, and may act more strongly on those that influence miRNA levels than those that influence mRNA levels. In addition, as single nucleotide mutation rate also affects allele frequency[94], I annotated mRNA and miRNA lead variants with the single nucleotide mutation rates estimated from individuals of European ancestry[95]. I did not observe a difference in the estimated mutation rate between lead variants and all tested variants or between lead variants of mRNAs and miRNAs.

Genes and DNAme sites are often regulated by more than one QTL[29], [96], [97]. Extending QTL detection to identify the multiple independent variants that affect the molecular trait levels helps in understanding the genetic architecture of molecular trait levels and capturing genes or DNAme sites that may mediate disease predisposition. Therefore, I performed multi-variant fine-mapping analysis using the Deterministic Approximation of Posteriors (DAP) algorithm to identify multiple independent association signals for genes/DNAme sites with $\geq$ 1 *cis*-QTL. Of the molecular traits that had $\geq$ 1 QTL in the marginal *cis*-eQTL analyses, I constructed 95% credible sets for molecular traits with $\geq$ 2 QTL signals for 29.1% of the mRNAs, 6.4% of the miRNA, and 16.8% of the DNAme sites in skeletal muscle tissue. Compared to skeletal muscle tissue, subcutaneous adipose tissue had similar proportions of mRNAs and DNAme sites and a larger proportion of miRNAs (13.9%) with $\geq$ 2 QTL signals(Table2.3.2, Figure2.3.3).

2.3.2 **Colocalization between T2D GWAS variants and gene/DNAme QTLs**

I used the multiple independent QTLs for the three types of molecular traits in skeletal muscle tissue and subcutaneous adipose tissue to look for potential genes and DNAme

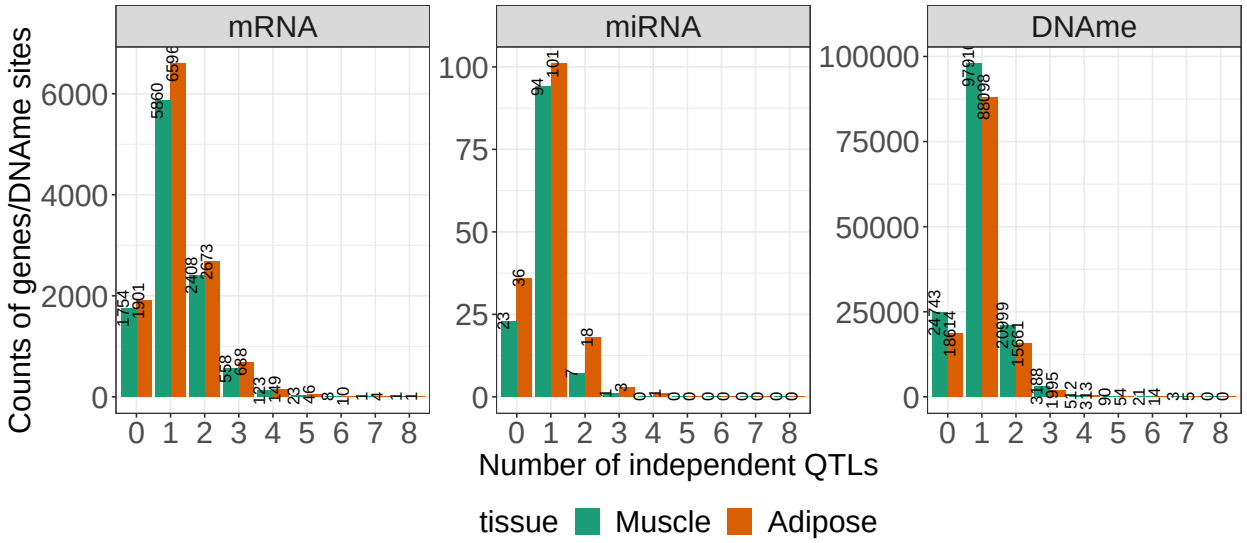


Figure 2.3.3. Multiple independent QTL discovery. Number of mRNA, miRNA, DNAme sites with 95% credible sets for N $(1 \leq N \leq 8)$ independent QTLs.

Molecular trait type	Tissue	Number of indepdent QTLs								
		0	1	2	3	4	5	6	7	8
mRNA	Muscle	1754	5860	2408	558	123	23	8	1	1
	Adipose	1901	6596	2673	688	149	46	10	4	1
miRNA	Muscle	23	94	7	1	0	0	0	0	0
	Adipose	36	101	18	3	1	0	0	0	0
DNAme	Muscle	24910	98160	21004	3189	512	90	21	3	0
	Adipose	18727	88350	15664	1995	313	54	14	5	0

Table 2.3.2. Number of mRNA, miRNA, DNAme sites with 95% credible sets for N $(1 \leq N \leq 8)$ independent QTLs.

sites implicated in the T2D GWAS signals. I used the colocalization test from fastEnloc[83], [84] to compute the probability of the sharing of causal variants between the marginal T2D GWAS signals and the QTL signals. A high colocalization probability suggests that the genetic associations for T2D may share causal variants with the QTLs for molecular traits. The sharing of causal variants between GWAS associations and QTLs can include at least three scenarios, the causal-effect scenario, the pleiotropic-effect scenario, and the different-causal-variants-in-high-LD scenario. In the causal-effect scenario, the same causal variant affects gene expression or DNA methylation level, thereby increasing T2D risk. In the pleiotropic-effect scenario, the same causal variant affects the gene expression or DNA methylation level and T2D risk independently (through different mechanisms). In the different-causal-variants-in-high-LD scenario, different causal variants in high LD affect the gene expression or DNA methylation level and T2D risk separately.

Regional colocalization probability (RCP) is a metric for the posterior probability of a genomic region having a colocalized signal. At an RCP ≥ 0.5, I identified in skeletal muscle tissue eight eQTLs colocalized with eight GWAS variants and 116 mQTLs colocalized with 74 GWAS variants; in subcutaneous adipose tissue, 14 eQTLs colocalized with 14 GWAS variants and 105 mQTLs colocalized with 69 GWAS variants (Table2.3.3; Supplementary Table2.8.2). I did not identify any miR-eQTLs that colocalized with GWAS variants. For mRNAs with eQTLs colocalized with T2D GWAS variants in skeletal muscle tissue, compared to the previous FUSION publications[26], [27], we identified an additional six mRNAs (*CEP68*, *INHBB*, *RFT1*, *FAM134C*, *PCGF3*, *AOC1*) colocalized with T2D GWAS variants. For mRNAs in subcutaneous adipose tissue, compared to the previous publications[28], [29], we identified an additional five mRNAs (*NUAK2*, *CEP68*, *HAUS6*, *PLEKHA1*, *ITGB6*) colocalized with T2D GWAS variants. For DNAme in skeletal muscle tissue, compared to Taylor et al.[27], we identified an additional 109 DNAme sites colocalized with T2D GWAS variants.

Taking the two tissues together, I identified a total of 15 unique mRNAs and 177 unique DNAme sites that had QTLs colocalized with T2D GWAS variants. Of these, there were instances where the secondary QTLs of the molecular traits colocalized with T2D GWAS variants: one secondary eQTLs for one gene (*PCGF3*) in both tissues, 13 secondary mQTLs for 13 DNAme sites in skeletal muscle tissue, and 12 secondary mQTLs for 12 DNAme sites in subcutaneous adipose tissue.

Molecular traits	Muscle		Adipose	
	GWAS variants	Molecular traits	GWAS variants	Molecular traits
mRNA	8	8	14	14
DNAme	74	116	69	105

Table 2.3.3. Number of colocalized GWAS loci-mRNA/DNAme pairs at RCP ≥ 0.5. RCP: regional colocalization probability; No GWAS variants colocalized with miRNA QTLs.

PCGF3 was the only gene whose secondary eQTL, instead of primary eQTL, colocalized with a T2D GWAS variant. The secondary eQTL of *PCGF3* was colocalized with one

of the three independent signals in a T2D GWAS locus[33] (Figure2.3.4;Supplementary Figure2.7.4). The other two GWAS signals in this region (lead variants rs1182788 and rs35654957) were not colocalized with the eQTLs of *PCGF3*. The three conditionally independent *cis*-eQTLs of *PCGF3* have different MAF (rs7672618 MAF = 0.34, rs73221128 MAF=0.04, and rs79739589 MAF= 0.09; Table2.3.4) and are in low LD R^2 with each other (max LD R^2 = 0.03), but they are in perfect D'(pairwise D'= 1).

	Lead varaint	Expression-decreasing allele	MAF	Tissue	Single-variant model		Multiple-variant model	
					Coefficient	p-value	Coefficient	p-value
1st eQTL	rs7672618	A	0.34	Muscle	-0.77	1.72E-50	-0.92	9.10E-93
				Adipose	-0.74	2.87E-47	-0.92	2.2E-93
2nd eQTL	rs73221128 (T2D GWAS variant)	T	0.04	Muscle	-0.50	4.71E-04	-0.91	1.10E-33
				Adipose	-0.54	2.59E-04	-1.02	1.10E-42
3rd eQTL	rs79739589	C	0.91	Muscle	0.83	1.89E-17	0.63	4.40E-33
				Adipose	0.69	3.18E-12	0.61	1.20E-35

Table 2.3.4. Summary statistics for the lead variants of the three independent QTL signals of *PCGF3* in the single- or multiple- variant model.Tested for the associations between *PCGF3* expression level and genotype dosages of variants adjusting for the first four genotype PCs and PEER factors. Single-variant model: only one variant was in the model; Multiple-variant model: all three variants were in the model.

Four haplotypes were formed by these three variants, `G_C_C`, `A_C_C`, `G_T_C`, `G_C_T` (alleles ordered by variants rs7672618, rs73221128, rs79739589), with haplotype frequencies of 0.54, 0.34, 0.04 and 0.09, respectively (Table2.3.5). rs73221128 T allele (T2D risk-increasing) is always on the same haplotype with the rs79739589 C allele. Compared to the haplotype `G_C_T`, the T2D risk allele carrying haplotype (T2D risk haplotype) `G_T_C` decreased the gene expression (effect size = -1.54, p-value$< 2.0 \times 10^{-16}$; effect size = -1.63, p-value$< 2.0 \times 10^{-16}$) (Table2.3.5). *PCGF3* (polycomb group ring finger 3) encodes a member of the polycomb group proteins, which are a collection of epigenetic chromatin modifiers that regulate gene expression[98]. *PCGF3* was more highly expressed in subcutaneous adipose tissue(median TPM = 23.3) than in skeletal muscle tissue(median TPM =9.7).

	Lead variants of three indepdent QTLs			Number of participants with 0/1/2 copies of haplotypes	Haplotype frequency	Muscle		Adipose	
	rs7672618	rs73221128	rs79739589			Beta	p	Beta	p
Haplotypes 1	G	C	C	72/160/96	0.54	-0.63	< 2e-16	-0.61	< 2e-16
Haplotypes 2	A	C	C	143/150/35	0.34	-1.55	< 2e-16	-1.53	< 2e-16
Haplotypes 3	G	T	C	301/27/0	0.04	-1.54	< 2e-16	-1.63	< 2e-16
Haplotypes 4	G	C	T	273/53/2	0.09	Reference			

Table 2.3.5. Four haplotypes formed by the lead variants of three independent eQTLs of *PCGF3* and their associations statistics with *PCGF3* expression level using the haplotype 4 as a reference

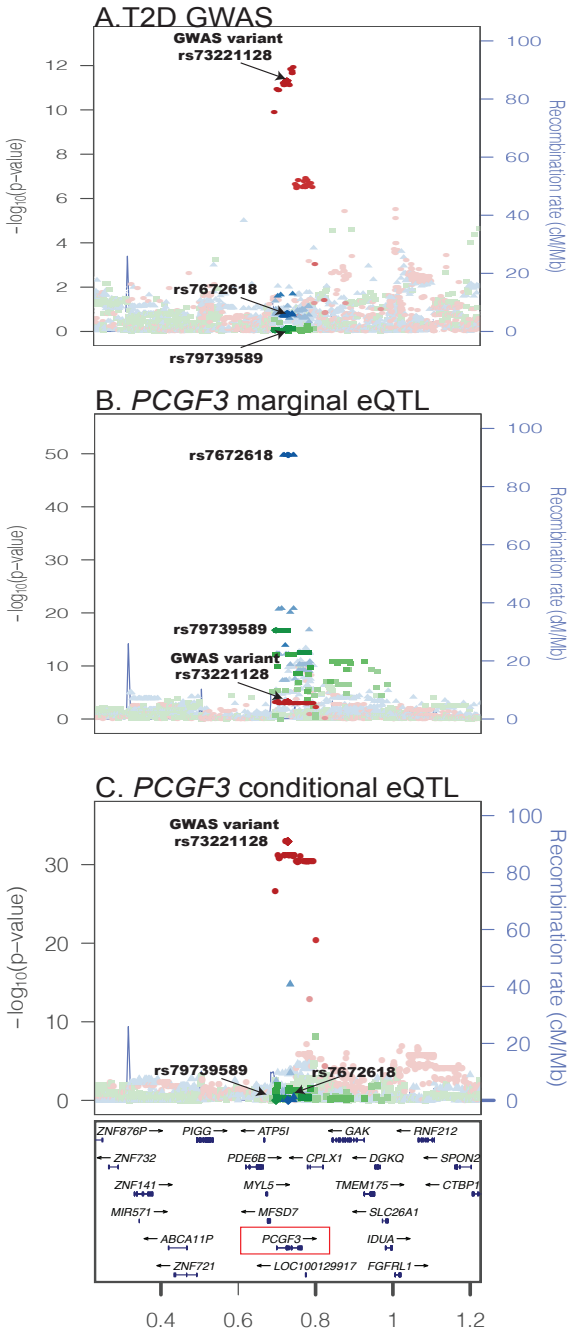


Figure 2.3.4. T2D GWAS variant rs73221128 is colocalized with the secondary eQTL for *PCGF3* in skeletal muscle tissue. Regional plots are colored by three independent eQTLs (represented by lead eQTL variants rs7672618, rs73221128, rs79739589) present in the FUSION data using FUSION LD. (A). Regional association plot for T2D meta-analysis from Mahajan et al.[33] at the rs73221128 locus (p-value = 4.5×10^{-12}); (B). Marginal eQTL association plot for *PCGF3* expression level. Marginal rs73221128-*PCGF3* association p-value = 4.71×10^{-4}. rs73221128 is in low LD R^2 (0.03) with the variant (rs7672618) that had the most significant marginal association with *PCGF3* expression (p-value = 1.72×10^{-50}); (C). After adjusting for rs7672618 and rs79739589, the T2D GWAS variant rs73221128 is more significantly associated with *PCGF3* (p-value = 1.10×10^{-33}).

In addition, I identified five GWAS variants in skeletal muscle tissue and eight in subcutaneous adipose tissue colocalized with both the eQTL for a gene and the mQTLs of nearby DNAme sites. Three of these five GWAS variants colocalized with the same mRNA (*RFT1*, *ANK1*, *INHBB*) and DNAme sites in both skeletal muscle and subcutaneous adipose tissues.

rs2581787 is the lead variant in a T2D GWAS locus with one single independent signal (rs2581787-T2D p-value= 3.0×10^{-8})[33]. rs2581787 was colocalized with the only *cis*-eQTL for *RFT1* in both tissues (Figure2.3.5;Supplementary Figure2.7.5). The T2D risk allele rs2581787-T was associated with a lower expression level of *RFT1* and a higher methylation level of cg22024966 in both tissues (Supplementary Figure2.7.6). Higher methylation levels of cg22024966 were associated with lower *RFT1* expression in both tissues (Supplementary Figure2.7.6). cg22024966 is located downstream of *RFT1*. *RFT1* encodes an enzyme involved in the translocation of the Man(5)GlcNAc(2)-PP-Dol intermediate from the cytoplasmic to the luminal side of the endoplasmic reticulum membrane[99]. rs2564940, in complete LD (R^2 = 1) with rs2581787, overlaps with ATAC-seq peaks in both tissues. rs2581787 resides in strong transcription chromatin states, and rs2564940 resides in weak/flanking TSS in both tissues (Supplementary Figure2.7.11). *RFT1* was was more highly expressed in subcutaneous adipose tissue (median TPM = 6.3) than in skeletal muscle tissue(median TPM = 1.8).

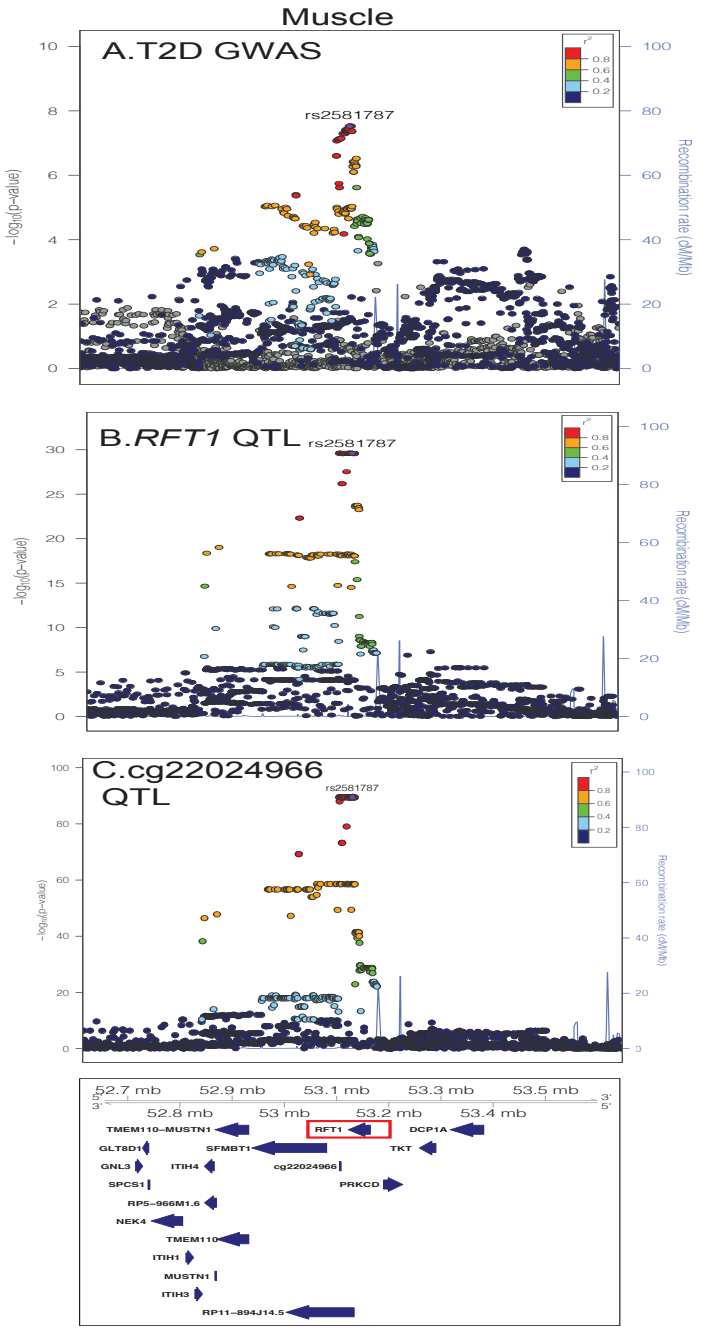
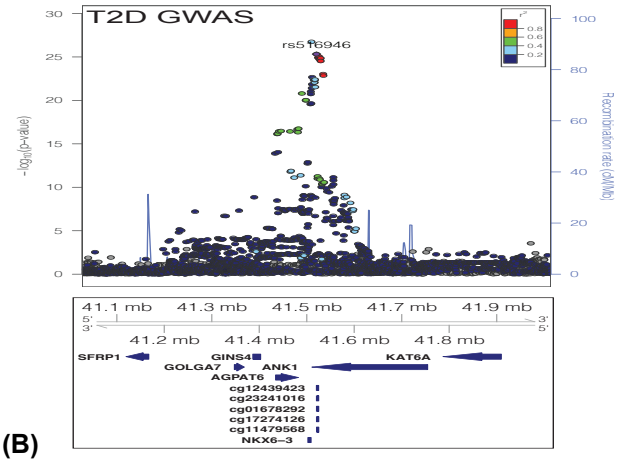


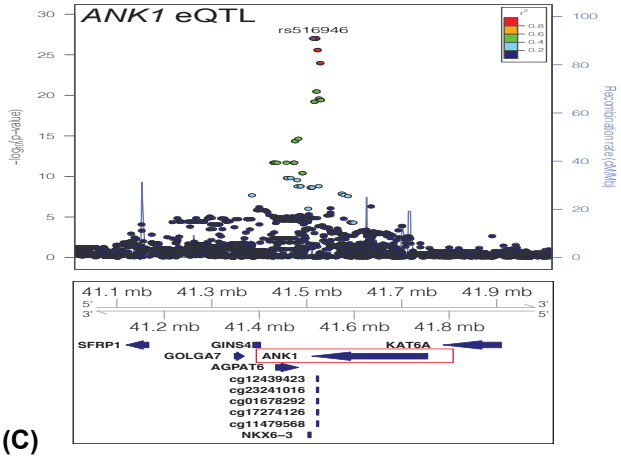
Figure 2.3.5. T2D GWAS signal is colocalized with the eQTL for *RFT1* and its nearby DNAme site cg22024966 in skeletal muscle tissue. Regional plots are colored LD R^2 with T2D GWAS variant rs2581787 using FUSION LD. (A). Regional association plot for T2D meta-analysis from Mahajan et al.[33] at the rs2581787 locus (p-value = 3.00×10^{-8}) ; (B). Marginal eQTL association plot for *RFT1* expression level. Marginal rs2581787-*RFT1* association p-value = 2.57×10^{-30}; (C). Marginal mQTL association plot for cg22024966 methylation level. Marginal rs2581787-cg22024966 association p-value = 3.14×10^{-90}.

T2D GWAS variant rs516946 (rs516946-T2D p-value=4.7×10^{-26}) was colocalized with the *cis*-eQTL signal for *ANK1* and five DNAme sites (cg11479568, cg17274126, cg23241016, cg12439423, cg01678292) in both tissues (Figure2.3.6 and Supplementary Figure2.7.7). These five DNAme sites are located within *ANK1*. *ANK1* has been identified to underlie the T2D-associated variant rs516946 in skeletal muscle tissue[26], [100] and subcutaneous adipose tissue[100]. Our results also revealed the potential connections at the DNAme level for rs516946. The T2D risk allele rs516946-C was associated with a higher expression level of *ANK1*, and lower methylation levels of cg01678292, cg12439423, cg17274126, cg11479568, cg23241016 in both tissues (Supplementary Figure2.7.8; Supplementary Figure2.7.9; Supplementary Figure2.7.10). In addition, lower methylation levels were associated with higher *ANK1* expression (Supplementary Figure2.7.8; Supplementary Figure2.7.9; Supplementary Figure2.7.10). Ankyrin 1, encoded by *ANK1*, plays a pivotal role in stabilizing the membrane structure of erythrocytes and stabilizing the sarcoplasmic reticulum around the myofibrils[101]. *ANK1* was more highly expressed in skeletal muscle tissue (median TPM = 104.8) than in subcutaneous adipose tissue (median TPM = 0.85). rs508419, in strong LD ($R^2 \geq 0.8$) with rs516946, was flanked by skeletal muscle stretch enhancers, fell in an active promoter, overlapped an ATAC-seq peak, and disrupted a TR4-binding site with in silico and in vitro evidence[26]. In subcutaneous adipose tissue, rs508419 resides in a weak promoter (Supplementary Figure2.7.11) and does not overlap with an ATAC-seq peak.

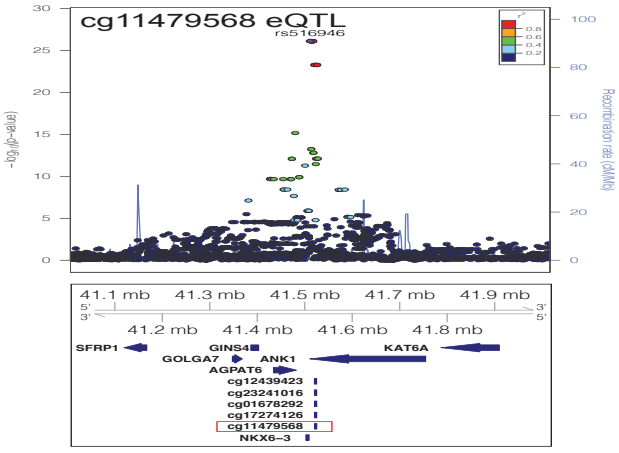
(A)


T2D GWAS
rs516946
-log10(p-value)
Recombination rate (cM/Mb)
41.1 mb
41.3 mb
41.5 mb
41.7 mb
41.9 mb
41.2 mb
41.4 mb
41.6 mb
41.8 mb
SFRP1
GINS4
KAT6A
GOLGA7
ANK1
AGPAT6
cg12439423
cg23241016
cg01678292
cg17274126
cg11479568
NKX6-3

(B)


ANK1 eQTL
rs516946
-log10(p-value)
Recombination rate (cM/Mb)
41.1 mb
41.3 mb
41.5 mb
41.7 mb
41.9 mb
41.2 mb
41.4 mb
41.6 mb
41.8 mb
SFRP1
GINS4
KAT6A
GOLGA7
ANK1
AGPAT6
cg12439423
cg23241016
cg01678292
cg17274126
cg11479568
NKX6-3

(C)


cg11479568 eQTL
rs516946
-log10(p-value)
Recombination rate (cM/Mb)
41.1 mb
41.3 mb
41.5 mb
41.7 mb
41.9 mb
41.2 mb
41.4 mb
41.6 mb
41.8 mb
SFRP1
GINS4
KAT6A
GOLGA7
ANK1
AGPAT6
cg12439423
cg23241016
cg01678292
cg17274126
cg11479568
NKX6-3

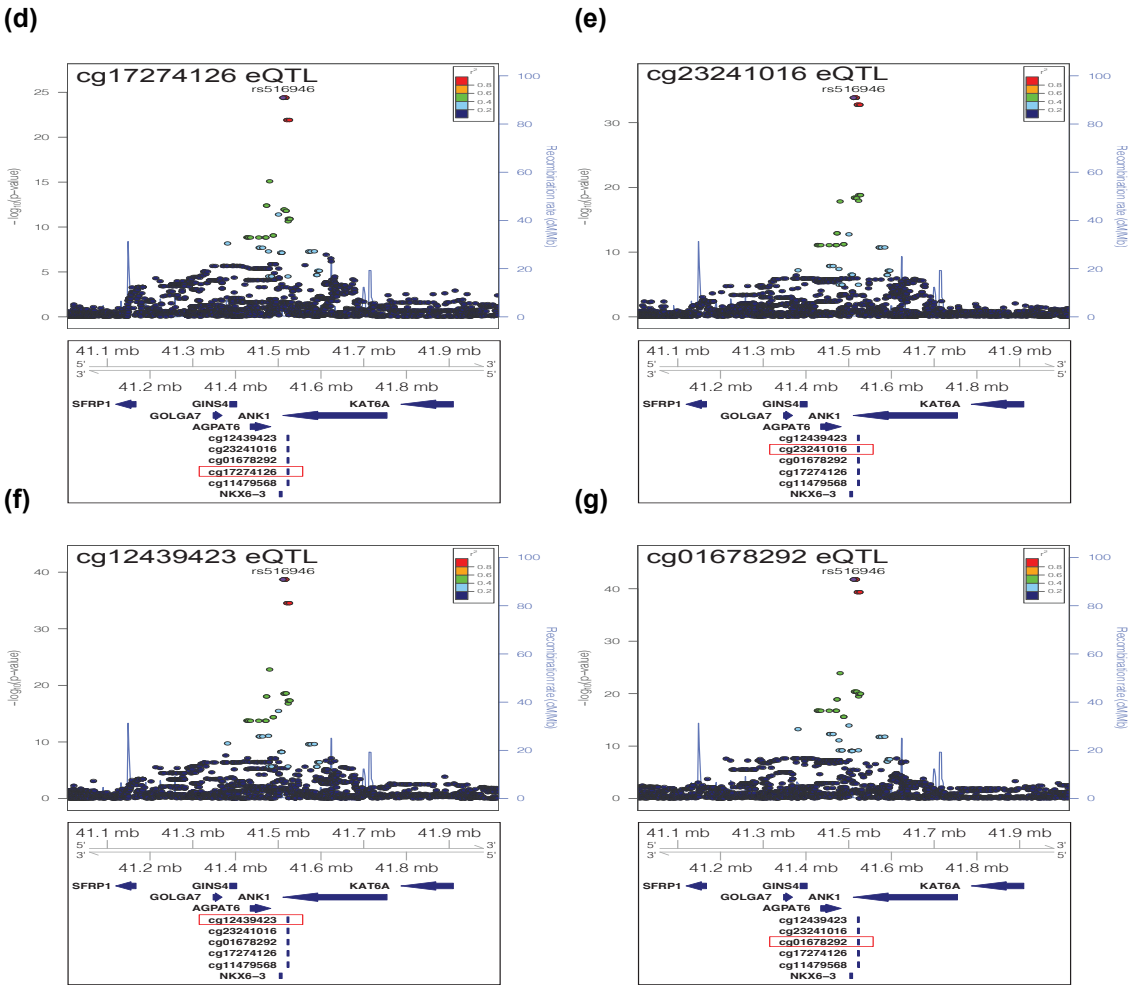


Figure 2.3.6. T2D GWAS signal is colocalized with QTLs of *ANK1* and its nearby DNAme sites in skeletal muscle tissue. Regional plots are colored by LD R^2 with T2D GWAS variant rs516946 using FUSION LD. (A). Regional association plot for T2D meta-analysis from Mahajan et al.[33] at the rs516946 locus (p-value = 4.70×10^{-26}); (B). Marginal eQTL association plot for ANK1 expression levels. rs516946-ANK1 association p-value = 9.71×10^{-28}; (C). Marginal mQTL association plot for cg11479568 methylation levels. rs516946-cg11479568 association p-value = 7.77×10^{-27}; (D). Marginal mQTL association plot for cg17274126 methylation levels. rs516946-cg17274126 association p-value = 4.04×10^{-25}; (E). Marginal mQTL association plot for cg23241016 methylation levels. rs516946-cg23241016 association p-value = 1.41×10^{-34}; (F). Marginal mQTL association plot for cg12439423 methylation levels. rs516946-cg12439423 association p-value = 1.93×10^{-39}; (G). Marginal mQTL association plot for cg01678292 methylation levels. rs516946- cg01678292 association p-value = 1.68×10^{-42}.

rs11688682 is the lead variant (rs11688682-T2D p-value=1.4×10^{-14}) at a T2D locus with three conditionally independent GWAS signals[33]. rs11688682 has been reported as a GWAS variant for triglyceride[102], HDL[102] and systolic blood pressure[103], but not fasting serum insulin or glucose[104]. In line with a previous report showing GWAS variant rs11688682 was colocalized with *INHBB*[29], I found colocalization between GWAS variant rs11688682 and an *INHBB* eQTL. *INHBB* is not the nearest gene to rs11688682, but is located 240 kb away. In addition, our results showed that rs11688682 was colocalized with the mQTLs of cg14231073 and cg15344192(Figure2.3.7 and Figure2.3.8). cg14231073 and cg15344192 are located downstream of *INHBB*. T2D risk allele rs11688682-G was associated with a higher expression level of *INHBB* and lower methylation levels of cg14231073 and cg15344192 in both tissues (Figure2.3.9;Figure2.3.10). Higher methylation levels of these DNAme sites were associated with lower *INHBB* expression (Figure2.3.9;Figure2.3.10). *INHBB* was more highly expressed in subcutaneous adipose tissue (median TPM = 59.4) than in skeletal muscle tissue (median TPM = 1.6). rs11688682 is a genotyped variant and is not in high LD ($R^2 \leq 0.43$) with any of the variants within 1 Mb. rs11688682 resides in an active enhancer and ATAC-seq peak in subcutaneous adipose tissue and a weakly repressed region in skeletal muscle tissue (Supplementary Figure2.7.11). *INHBB* encodes a subunit of activin[105], a major regulator of testicular and ovarian development[106].

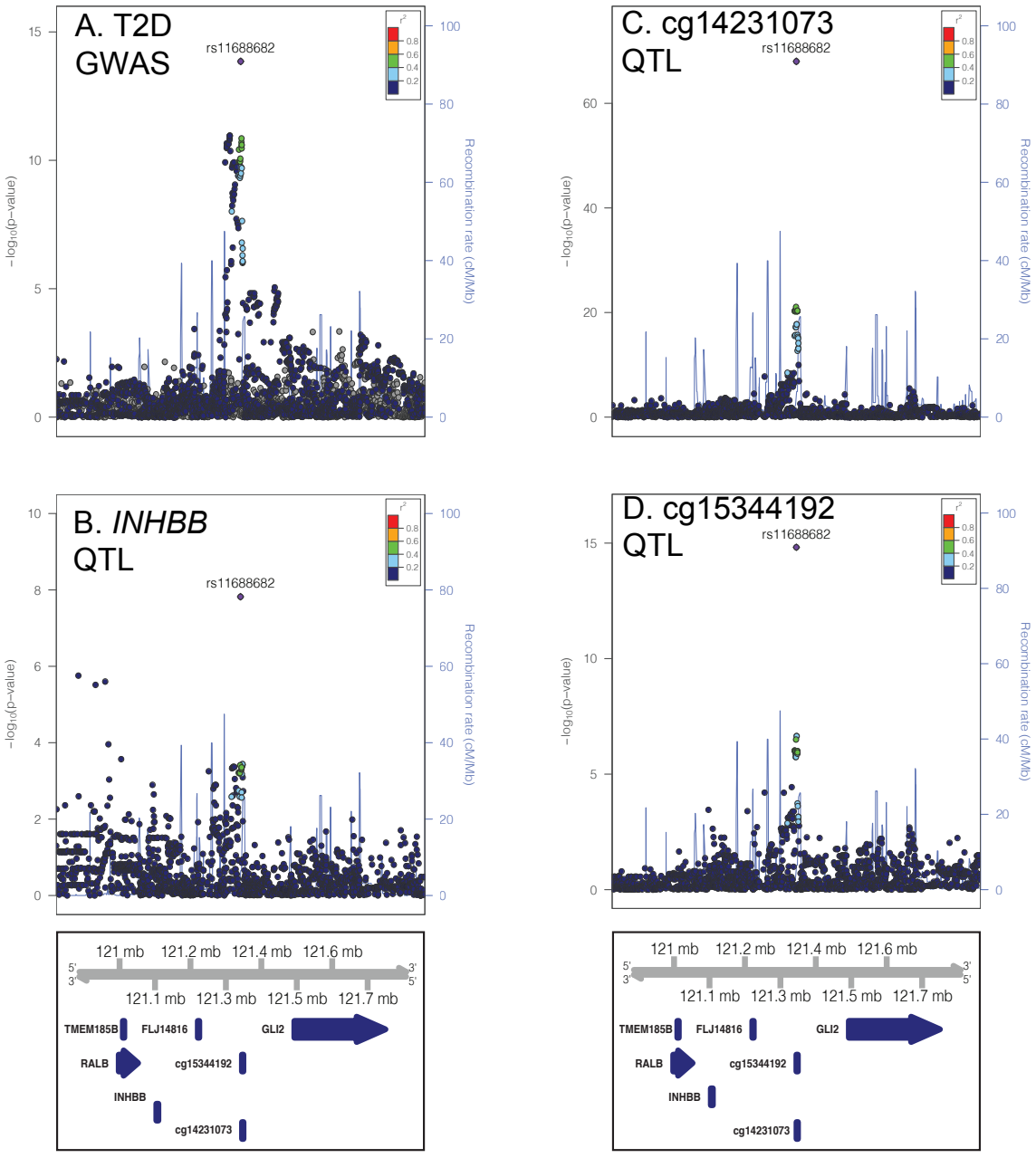


Figure 2.3.7. T2D GWAS variant rs11688682 is colocalized with the QTLs of *INHBB* and two DNAme sites cg14231073 and cg15344192 in skeletal muscle tissue. Regional plots are colored LD R^2 with T2D GWAS variant rs11688682 using FUSION LD. (A). Regional association plot for T2D meta-analysis from Mahajan et al.[33] at the rs11688682 locus (p-value = 1.40×10^{-14}); (B). Marginal eQTL association plot for *INHBB* expression level. rs11688682-*INHBB* association p-value = 1.51×10^{-8}; (C). Marginal mQTL association plot for cg14231073 methylation level. Marginal rs11688682-14231073 association p-value = 9.37×10^{-69}; (D). Marginal mQTL association plot for cg15344192 methylation level. Marginal rs11688682-cg15344192 association p-value = 1.51×10^{-15} p=1.51*10−15;

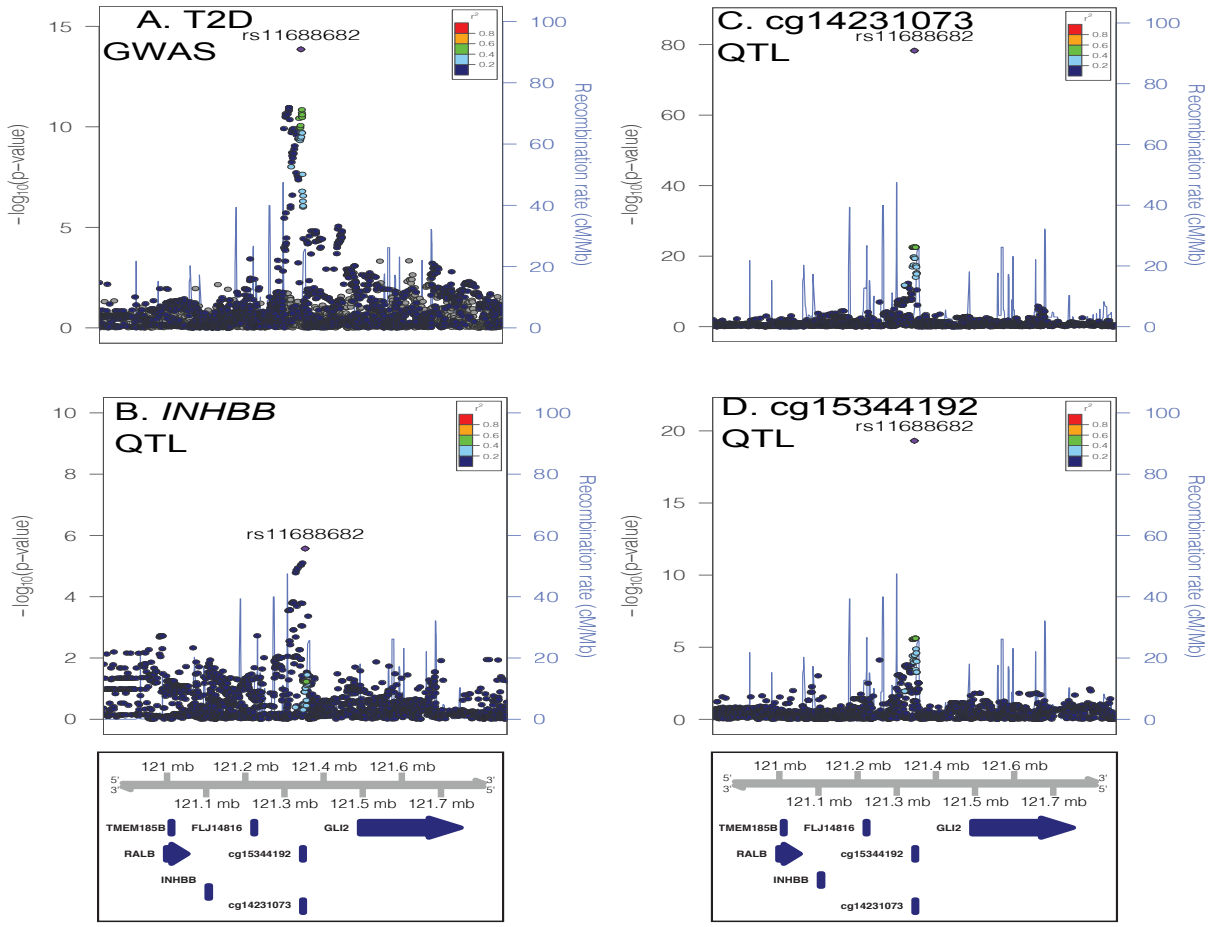


Figure 2.3.8. T2D GWAS variant rs11688682 is colocalized with the QTLs of *INHBB* and two DNAme sites cg14231073 and cg15344192 in subcutaneous adipose tissue. Regional plots are colored LD R_2 with T2D GWAS variant rs11688682 using FUSION LD. (A). Regional association plot for T2D meta-analysis from Mahajan et al.[33] at the rs11688682 locus (p-value = 1.40×10^{-14}); (B). Marginal eQTL association plot for *INHBB* expression level. rs11688682-*INHBB* association p-value = 2.68×10^{-6}; (C). Marginal mQTL association plot for cg14231073 methylation level. Marginal rs11688682-14231073 association p-value = 4.96×10^{-79}; (D). Marginal mQTL association plot for cg15344192 methylation level. Marginal rs11688682-cg15344192 association p-value = 4.83×10^{-20}.

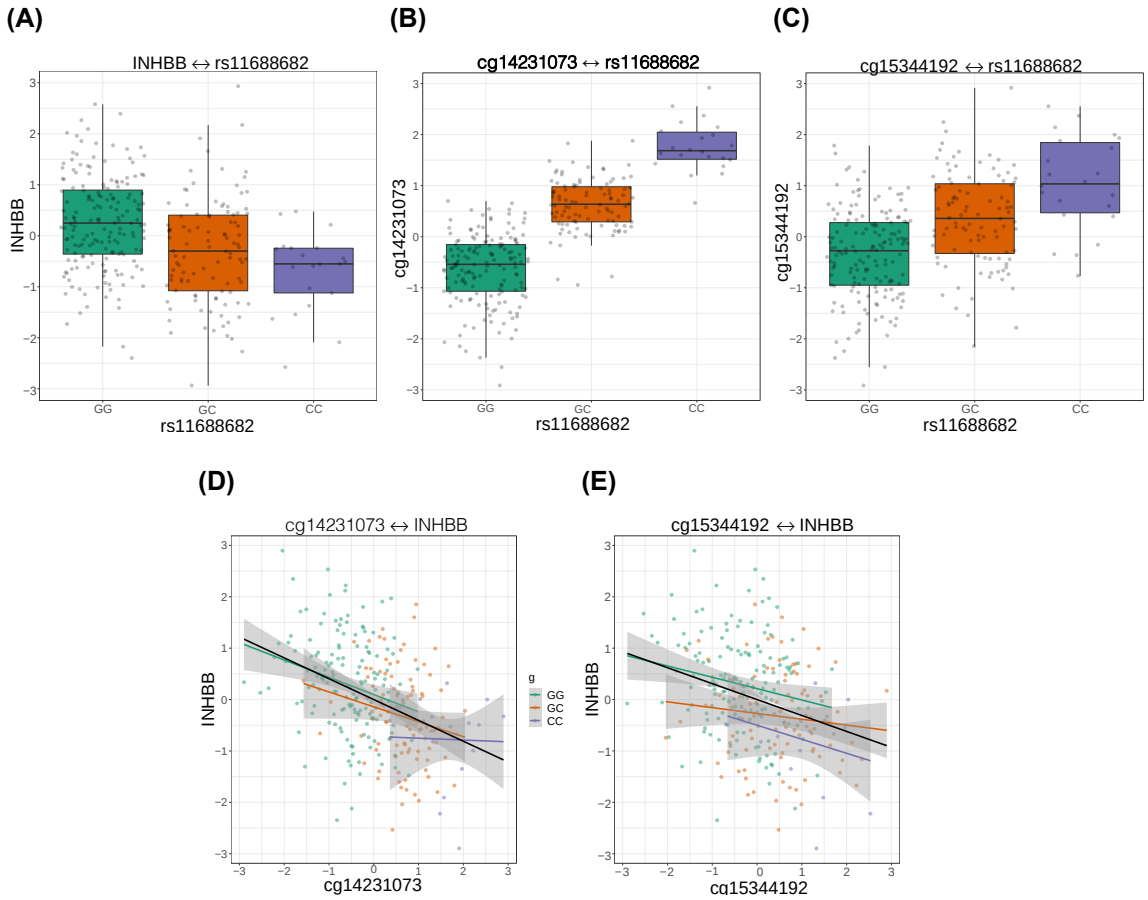


Figure 2.3.9. Effects of rs11688682 on *INHBB* and its nearby DNAme sites cg14231073 and cg15344192 in skeletal muscle tissue. (A). Box plot of residual *INHBB* expression levels by rs11688682 genotype; (B). Box plot of residual cg14231073 methylation level by rs11688682 genotype; (C). Box plot of residual cg15344192 methylation level by rs11688682 genotype; (D). Scatter plot of residual *INHBB* expression (adjusted for PEER factors used in QTL mapping; y-axis) and residual cg14231073 methylation level (adjusted for PEER factors used in QTL mapping; x-axis, colored by rs11688682 genotypes; (E). Scatter plot of residual *INHBB* expression (adjusted for PEER factors used in QTL mapping; y-axis) and residual cg15344192 methylation level (adjusted for PEER factors used in QTL mapping; x-axis, colored by rs11688682 genotypes. Linear regression lines for the relationship overall (black) and within each rs11688682 genotype (GG, green; GC, orange; CC, purple).

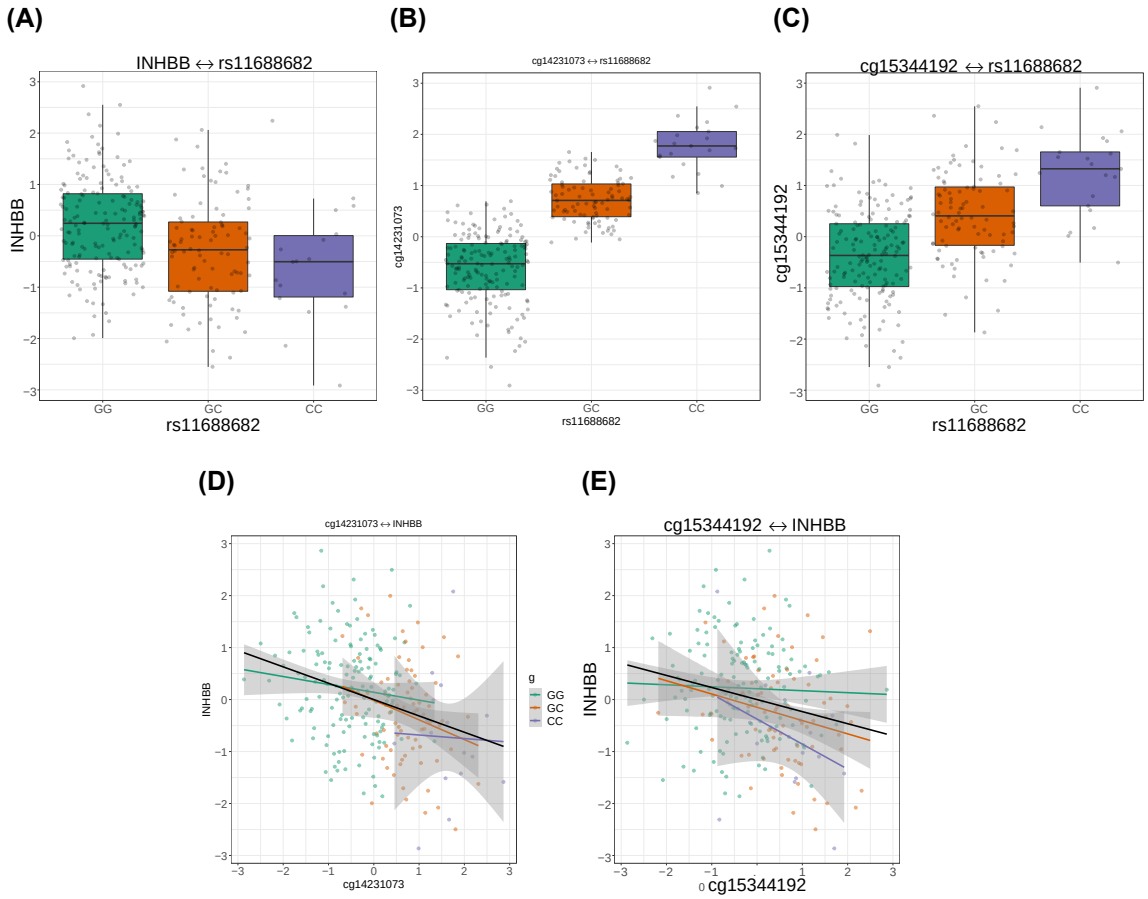


Figure 2.3.10. Effects of rs11688682 on *INHBB* and its nearby DNAme cg14231073 and cg15344192 in subcutaneous adipose tissue. (A). Box plot of residual *INHBB* expression levels by rs11688682 genotype; (B). Box plot of residual cg14231073 methylation level by rs11688682 genotype; (C). Box plot of residual cg15344192 methylation level by rs11688682 genotype; (D). Scatter plot of residual *INHBB* expression (adjusted for PEER factors used in QTL mapping; y-axis) and residual cg14231073 methylation level (adjusted for PEER factors used in QTL mapping; x-axis, colored by rs11688682 genotypes. (E). Scatter plot of residual *INHBB* expression (adjusted for PEER factors used in QTL mapping; y-axis) and residual cg15344192 methylation level (adjusted for PEER factors used in QTL mapping; x-axis, colored by rs11688682 genotypes. Linear regression lines for the relationship overall (black) and within each rs11688682 genotype (GG, green; GC orange; CC, purple).

2.3.3 **rs11688682 T2D risk allele increased transcriptional activity in luciferase assay (performed by Swarooparaniand Vadlamudi and Karen Mohlke)**

rs11688682 is located in an ATAC-seq peak in preadipocytes and in differentiated adipocytes (Hannah Perrin, unpublished data). Given the orientation of the *INHBB* expression increasing allele and the T2D GWAS risk allele, we expect the T2D GWAS risk allele would increase the expression of *INHBB*. We tested rs11688682 for allelic differences in transcriptional activity using luciferase assay in preadipocytes and adipocytes. We separately cloned DNA segments containing either the T2D risk allele (G) or the non-risk allele (C) in forward and reverse orientations to luciferase reporter constructs and conducted luciferase assay in preadipocytes and differentiated adipocytes cells. The region spanning rs11688682 showed differential allelic enhancer activity in both orientations in both preadipocytes and adipocytes. The T2D risk allele rs11688682-G had higher luciferase activity than the non-risk allele rs11688682-C (preadipotype: forward orientation p-value = 2.5×10^{-3}, reverse orientation p-value = 3.1×10^{-3}; adipocyte: forward orientation p-value = 0.07, reverse orientation p-value = 4.0×10^{-3}; Figure2.3.11A, Figure2.3.11B). The T2D risk allele showed a 1.45-fold to 1.83-fold increase in transcriptional activity relative to the non-risk allele in both orientations in preadipocytes, and 1.48-fold to 2.65-fold increase in adipocytes. These experimental results suggest that rs11688682 is located within an enhancer element and the T2D risk G allele increases transcriptional activity in preadipocytes and adipocytes.

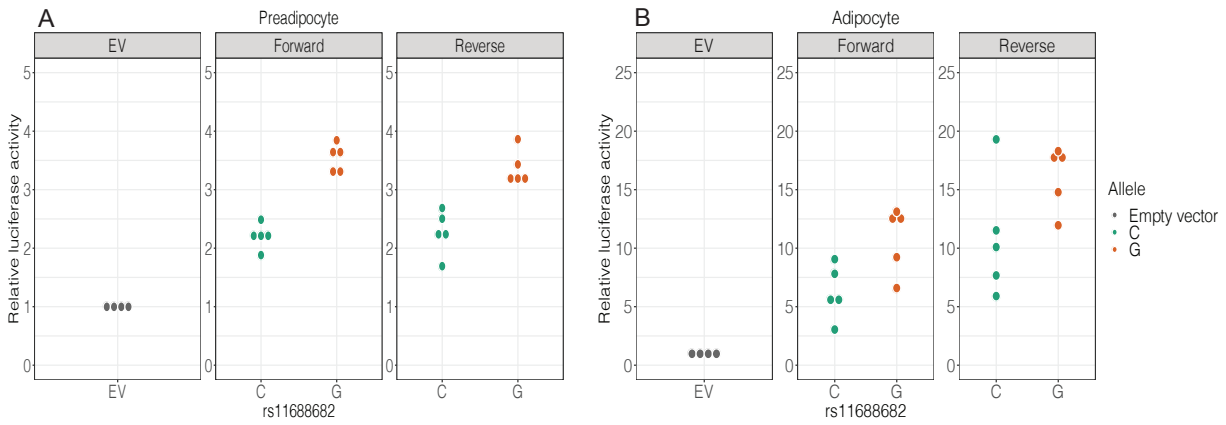


Figure 2.3.11. rs11688682 showed allelic differences in transcriptional activity using luciferase assay. (A)In preadipocytes, T2D risk rs11688682-G allele showed greater transcriptional activity than the non-risk allele. Forward orientation p-value=2.5×10^{-3}, reverse orientation p-value=3.1×10^{-3}; (B). In adipocytes, T2D risk rs11688682-G allele showed greater transcriptional activity than the non-risk allele. Forward orientation p-value=0.07, reverse orientation p-value=4.0×10^{-3}. We generated five independent clones for each allele with and measured enhancer activity in tripletes for each clone.

2.3.4 **Molecular trait (mRNA, miRNA, DNAme) association with physiological traits**

Changes in gene expression and DNA methylation levels may be causal or responsive to pathological changes. To help understand the etiology and manifestations of T2D and related traits at the molecular level, we identified mRNAs, miRNAs, and DNAme sites whose levels of expression or methylation differed between individuals with T2D and normal glucose tolerance (NGT) or by the levels of T2D-relevant physiological traits. We used an FDR of $\leq$ 1% with the Benjamini-Hochberg procedure as the statistical significance threshold for within-tissue association tests.

Bulk skeletal muscle tissue or subcutaneous adipose tissue consists of diverse cell types, including cell types in the target tissue (skeletal muscle tissue or subcutaneous adipose tissue) and cell types from non-target tissues (such as blood or skin). Cell-type composition has the potential to confound associations between molecular and physiological traits because cell-type composition can correlate with both physiological and molecular trait levels[27]. For skeletal muscle tissue, we estimated the proportions of subcutaneous adipose tissue, skeletal muscle tissue, blood, skin, and lymphocytes using GTEx RNA-seq datasets as a reference; we estimated the proportion of each of the three muscle fiber types (Type 1, Type 2A, Type 2X) using the percentages of the expression

level of the dominant myosin heavy chain gene (*MYH1*, *MYH2* or *MYH7*) of each muscle fiber(Supplementary Figure2.7.12). For subcutaneous adipose tissue, we computed two sets of estimates. For one set, we estimated the proportions of four cell types (adipocytes, macrophages, CD4+ T cells, and microvascular endothelial cells) and blood using publicly available primary or PSC/iPSC-derived RNA-seq data (five-component estimates) (Supplementary Figure2.7.13). For the other set, we used the relative amount of each of the 17 cell types identified from the subcutaneous adipose single-nuclei RNA-seq data (provided by Dr.Paivi Pajukanta) in our subcutaneous adipose tissue samples (17-component estimates). As an alternative to tissue/cell-type composition estimates, we estimated surrogate variables which were designed to represent variations from biological or other factors while protecting the effects of the physiological trait of interest.

We tested for physiological trait-molecular trait associations with and without adjusting for tissue/cell-type composition. Adjusting for tissue/cell-type composition or surrogate variables typically substantially reduced the number of mRNAs or DNAme sites associated with physiological traits in both tissues as compared to adjusting for the base set of covariates (Supplementary Figure2.7.14 and Supplementary Figure2.7.15). For miRNA, adjusting for tissue/cell-type composition or surrogate variables increased the number of associated miRNAs for a small proportion of physiological traits and decreased the number for most physiological traits (Supplementary Figure2.7.14 and Supplementary Figure2.7.15. The associations between the level of a physiological trait and the level of a molecular trait (an mRNA, a miRNA or a DNAme site) from different tissue/cell-type adjustment approaches were consistent overall in terms of direction and strength (Supplementary Figure2.7.16;Supplementary Figure2.7.17;Supplementary Figure2.7.18;Supplementary Figure2.7.19), while the number of significant associations differed. We observed that some of the estimated surrogate variables were correlated with a physiological trait of interest and adjusting for them may remove the effects of the physiological trait on molecular traits. For the following analyses we used the results adjusting for tissue/cell type composition estimates for both tissues(Figure2.3.12). Specifically for subcutaneous adipose tissue, as the 17-component estimate was a more comprehensive representation of cell types in sub-

cutaneous adipose tissue than the 5-component estimate, we used the results adjusting for the 17-component estimates.

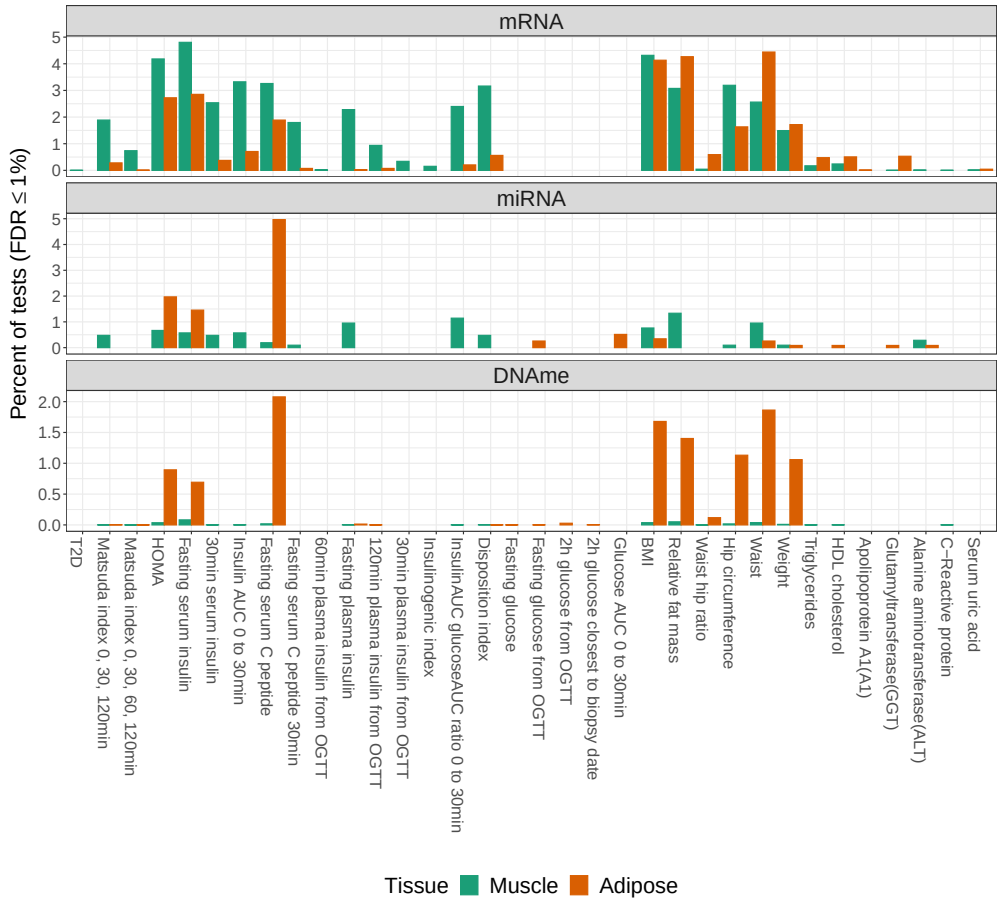


Figure 2.3.12. Percent of mRNAs/miRNAs/DNAme sites associated with the levels of physiological traits at FDR$\leq$ 1% in skeletal muscle (green) and subcutaneous adipose tissue (orange). Results shown are adjusted for tissue and fiber type estimates in skeletal muscle tissue and adjusted for the 17-component estimates in subcutaneous adipose tissue.

Within skeletal muscle tissue and subcutaneous adipose tissue2.3.12, we observed that the largest number of significant associations were found for two groups of physiological traits(Pearson correlation r=0.61, adjusted for covariates): insulin-related (e.g. fasting serum insulin) or body fat distribution-related physiological traits (e.g. BMI). Insulin-related physiological traits were associated with a slightly higher proportion of mRNAs in skeletal muscle tissue than in subcutaneous adipose tissue, whereas body fat distribution-related physiological traits were associated with a slightly higher proportion of mRNAs in subcutaneous adipose tissue than in skeletal muscle tissue. In addition, we observed that most of the physiological trait-molecular trait associations were found either in skeletal muscle tissue or subcutaneous adipose tissue, not in both (Figure2.3.13; Figure2.3.14).

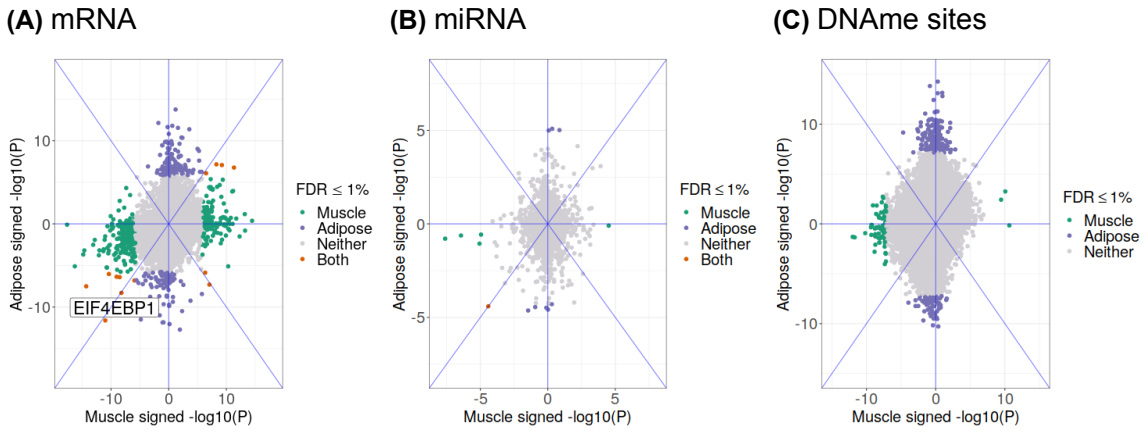


Figure 2.3.13. Fasting serum insulin associations with mRNAs/miRNAs/DNAme sites in skeletal muscle and subcutaneous adipose tissue. Scatterplots show signed -log10(p-value) of associations in skeletal muscle tissue (x-axis) and subcutaneous adipose tissue (y-axis), colored by whether an association p-value is significant in only skeletal muscle tissue, only subcutaneous adipose tissue or both, using a threshold of $\leq$ 1% FDR. The sign of an association is based on the estimated regression coefficient.

We asked whether the same molecular traits were more likely to be associated with the same physiological traits in both skeletal muscle and subcutaneous adipose tissues than expected by chance using Fisher's exact test. Using a p-value threshold of 1.04×10^{-3} (Bonferroni correction for the number of physiological traits tested, 0.05/48) for Fisher's exact test, seven physiological traits (BMI, relative fat mass, waist, fasting serum insulin, HOMA, fasting serum C peptide, fasting serum C peptide 30min) had more genes with significant trait-gene expression associations in both tissues than expected by chance

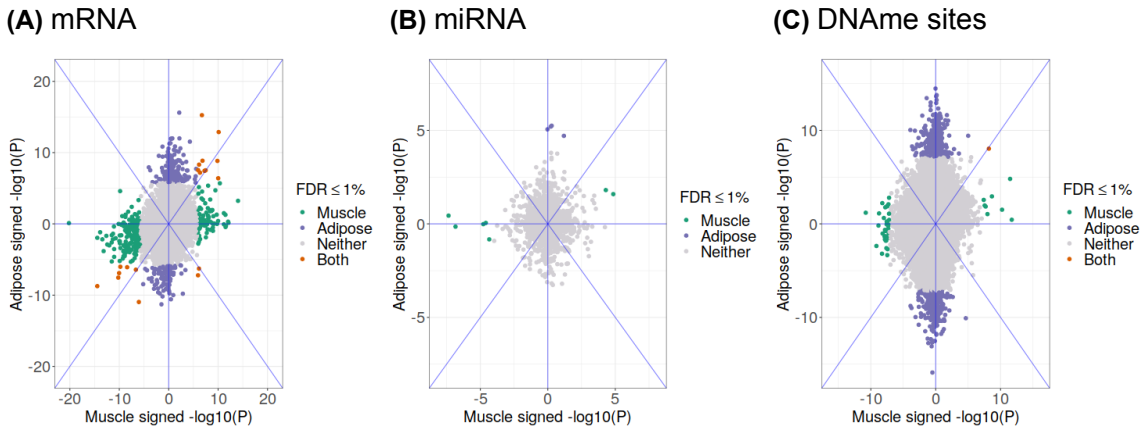


Figure 2.3.14. BMI associations with mRNAs/miRNAs/DNAme sites in skeletal muscle and subcutaneous adipose tissue. Scatterplots show signed -log10(p-value) of associations in skeletal muscle tissue (x-axis) and subcutaneous adipose tissue (y-axis) for every tested molecular traits, colored by whether an association p-value is significant in only skeletal muscle tissue, only subcutaneous adipose tissue or both, using a threshold of $\leq$ 1% FDR. The sign of an association is based on the estimated regression coefficient.

(Table2.3.6; Supplementary Table2.8.4). Of the mRNA-physiological trait associations significant in both tissues, $\geq$ 71% showed a consistent direction of effect. The seven physiological traits can be divided into two groups, one group related to body fat distribution, and the other related to insulin. Among the insulin-related physiological traits, fasting serum insulin had the largest number (n=13) of mRNAs that were significant in both tissues. Of these13 mRNAs, *EIF4EBP1* displayed highly consistent effects on physiological traits in both direction and strength (Figure2.3.15). Higher *EIF4EBP1* was associated with a beneficial physiological trait profile (lower BMI, waist, C peptide, fasting serum insulin; higher HDL and Matsuda index).

Physiological trait	Number of significant associations in muscle	Number of significant associations in adipose	Number of significant associations in both tissues	Different direction	Same direction	Fisher test p-value	Odds ratio
BMI	288	311	19	2	17	1.72E-10	7.23
Relative fat mass	211	337	17	1	16	2.51E-10	8.20
Waist	157	360	14	0	14	6.08E-09	8.48
Fasting serum insulin	350	226	13	2	11	2.44E-06	5.42
HOMA	296	213	11	2	9	8.38E-06	5.73
Fasting serum C peptide	246	152	7	2	5	2.45E-04	6.08
Fasting serum C peptide 30min	117	10	2	0	2	6.44E-04	65.89

Table 2.3.6. The seven physiological traits that had more mRNAs significant in both tissues than expected by chance, using a p-value threshold of 1.04×10^{-3} (Bonferroni correction for the number of physiological traits tested, 0.05/48) for Fisher's exact test.

As obesity and insulin resistance are interconnected physiologically and are both key risk factors for T2D, we asked whether the significant associations were driven by BMI or fasting serum insulin by adjusting for BMI or fasting serum insulin. In skeletal muscle tissue, almost none of the mRNAs associated with body fat distribution-relevant traits re-

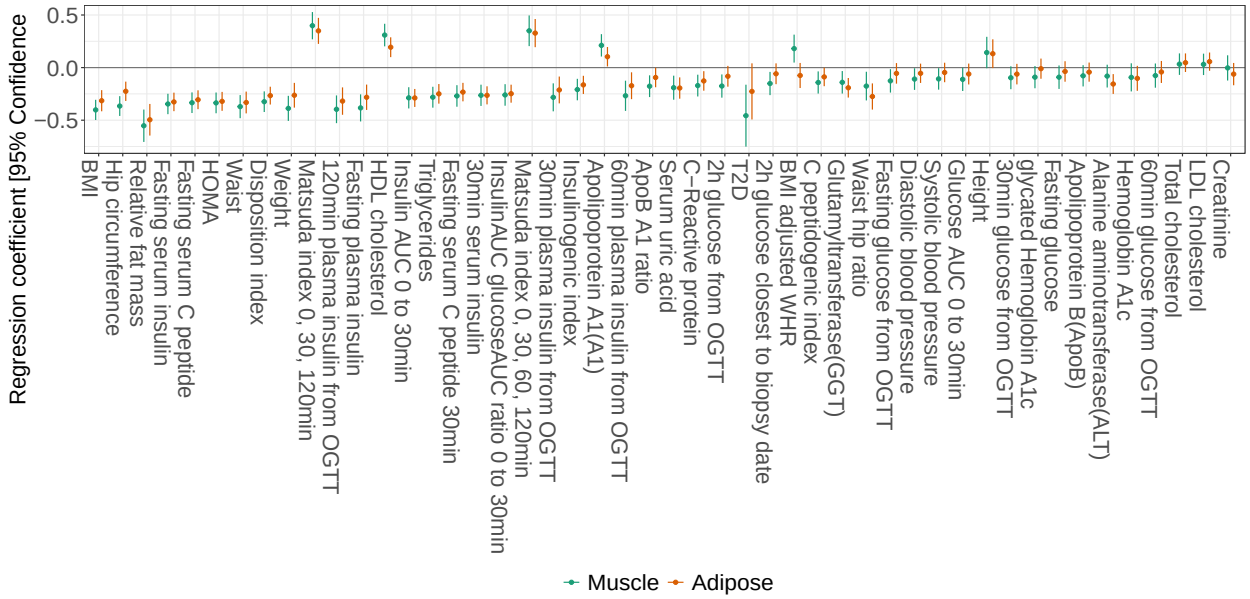


Figure 2.3.15. Coefficients and 95% confidence intervals between the *EIF4EBP1* expression level and the levels of physiological traits in skeletal muscle and subcutaneous adipose tissue.

mained significant with the additional adjustment of fasting serum insulin (92% – 100% decrease), whereas some of the mRNAs associated with insulin-relevant physiological traits remained significant with the additional adjustment of BMI (83% – 100% decrease) (Supplementary Figure2.7.20; Supplementary Figure2.7.22; Supplementary Figure2.7.23). In subcutaneous adipose tissue, there were a small number of significant physiological trait-mRNA associations remaining either additionally adjusting for fasting serum insulin (96% – 100% decrease) or BMI (79% – 99% decrease) (Supplementary Figure2.7.21; Supplementary Figure2.7.24; Supplementary Figure2.7.25). This suggests that while most of the significant physiological trait-mRNA associations were driven by the biological processes related to both fasting serum insulin and BMI, in skeletal muscle tissue there were a small number of insulin related traits-mRNA associations that could not be solely explained by BMI; in subcutaneous adipose tissue, there were a small number of insulin related traits-mRNA associations that could not be solely explained by BMI as well as a small number of BMI related traits-mRNA associations that could not be solely explained by fasting serum insulin.

Of the mRNAs whose eQTL was colocalized with T2D GWAS signals, one mRNA (*INHBB*) was significantly positively correlated with insulin related physiological traits (fasting

serum insulin, HOMA, and fasting serum c-peptide) and body fat distribution related physiological traits (waist and relative fat mass) in subcutaneous adipose tissue (Figure2.3.16). These *INHBB*-physiological traits associations were directionally consistent with the observations that the T2D risk allele rs11688682-G was associated with higher *INHBB* expression level and that diabetic individuals usually have higher levels of insulin resistance indices[107], [108]. rs11688682 has not been reported to be a GWAS signal for fasting serum insulin[104]. *INHBB* expression level was also positively but not significantly correlated with T2D versus NGT (p-value = 0.24), which might be due to the small sample size for the T2D versus NGT comparison(n=176). Since insulin resistance and obesity have shared etiology, we asked whether *INHBB*-physiological trait associations were driven by insulin resistance or obesity by adjusting for fasting serum insulin and waist (or BMI). When we adjusted for fasting serum insulin, all of the significant associations became insignificant; when we adjusted for waist (or BMI), the associations with insulin-related physiological traits were attenuated but still significant (Supplementary Table2.8.5). This indicates that significant *INHBB*-physiological traits associations in our subcutaneous adipose tissue samples may be primarily driven by insulin resistance, not obesity.

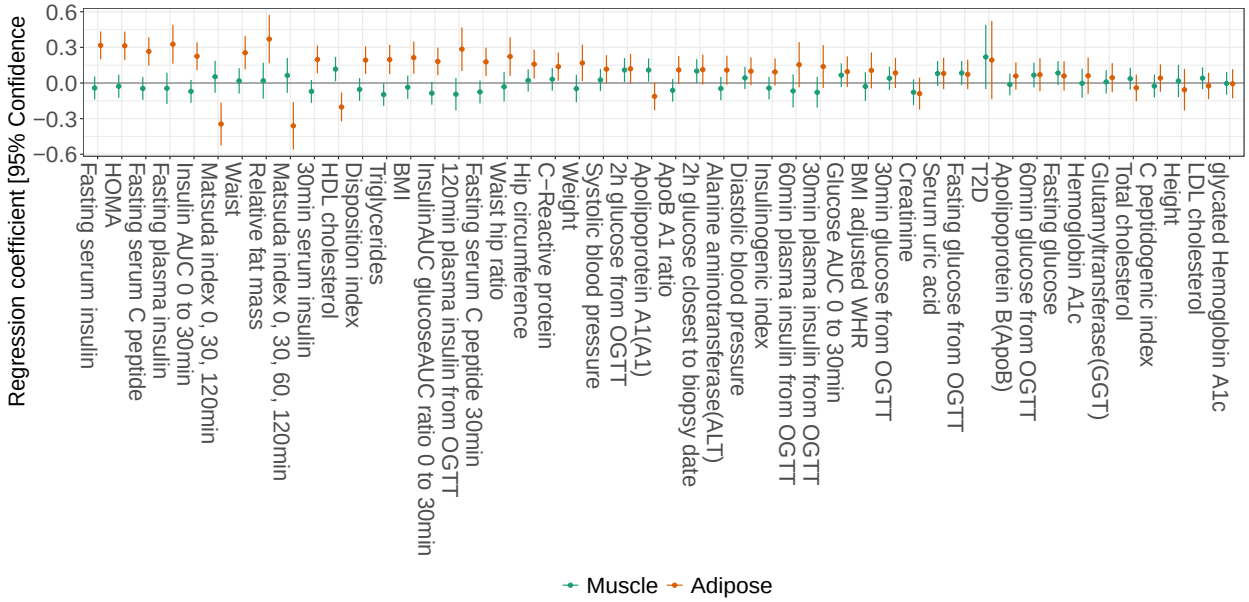


Figure 2.3.16. Coefficients and 95% confidence intervals between the *INHBB* expression level and the levels of physiological traits in skeletal muscle and subcutaneous adipose tissue.

The physiological trait-molecular trait associations may derive from any cell types present in bulk tissue samples, including whole blood that remains in the biopsies. We found that hsa-miR-122-5p, the miRNA that was most significantly (positively) associated with serum Alanine aminotransferase (ALT) and Glutamyltransferase (GT) in both skeletal muscle tissue and subcutaneous adipose tissue(Supplementary Figure2.7.26), was positively correlated with hemoglobin subunit beta (*HBB*, a blood specific gene[109], [110]) expression level in subcutaneous adipose tissue (Spearman's rho=0.22, two-sided test p-value=4.9×10-4; (Supplementary Figure2.7.27)). ALT and GT are mainly synthesized in the liver, and hsa-miR-122-5p is primarily expressed in liver[111], suggesting the possibility that ALT, GT, and hsa-miR-122-5p are released from hepatocytes into the circulating blood because of liver damage caused by ectopic fat deposition and insulin resistance.

2.4 Discussion

In this study, using mRNA and miRNA expression levels and DNA methylation levels measured in two T2D-relevant tissues (skeletal muscle and subcutaneous adipose) in combination with genotype data, we identified similar proportions of mRNA, miRNA and DNAme sites having ≥ 1 *cis*-QTLs. Integrating these QTLs with T2D GWAS variants using colocalization analysis, we identified a total of 15 unique mRNAs and 177 unique DNAme sites that had QTLs colocalized with T2D GWAS variants in the two tissue together. We did not identify any miRNAs that had QTLs colocalized with T2D GWAS variants in either tissue. Using mRNA and miRNA expression levels and DNA methylation levels in combination with T2D-relevant physiological traits, we identified mRNAs, miRNAs and DNAme sites that were associated with physiological traits. For every tested physiological trait, we observed the trend that most of the associations with mRNA, miRNA or DNAme sites were significant in one of the two tissues, not both. Seven physiological traits (BMI, relative fat mass, waist, fasting serum insulin, HOMA, fasting serum C peptide, fasting serum C peptide 30min) had more mRNAs significant in both tissues than expected by chance.

We provided several lines of evidence for the hypothesis that compared to mRNAs, miRNAs may be under stronger selective pressure and therefore have a lower level of *cis*-QTL regulation. First, of all tested mRNAs or miRNAs, there was a smaller proportion of miRNAs having *cis*-QTLs than mRNAs. Second, this trend persisted when we compared the proportions of mRNAs or miRNAs with *cis*-QTLs at different read count levels, suggesting that a lower proportion of miRNAs having *cis*-QTLs than mRNAs was not due to differences in the power to detect a QTL. Third, the lead variants of miRNA *cis*-QTLs had lower MAF than the lead variants of mRNA *cis*-eQTLs, suggesting stronger constraints on genetic variants affecting miRNA than mRNA.

Several programs[83], [112]–[115] are available to perform colocalization analysis between GWAS and QTL associations, including Enloc[83]. FastEnloc[83], [84] is the latest implementation of Enloc that runs faster. FastEnloc allows for testing colocalization between a GWAS signal and each independent eQTL of a molecular trait by using the posterior probabilities of multiple eQTLs identified by DAP[82]. GTEx[31] extensively compared

colocalization methods and decided to use the approach of detecting multiple QTLs with DAP followed by testing for colocalization with Enloc as their primary approach. They found that this approach was able to capture secondary eQTLs colocalized with GWAS associations and had well-controlled type 1 error.

In the colocalization analysis, we found T2D GWAS variant rs516946 was colocalized with the *ANK1* eQTL in both skeletal muscle and subcutaneous adipose tissues. A GWAS meta-analysis of European individuals (Mahajan et al.)[33] reported three conditionally independent signals for T2D at this locus, the primary signal rs13262861, and two secondary signals rs148766658 and rs4736819. Mahajan et al.[33] highlighted that *NKX6-3*, the *cis*-eGene of rs13263861 in pancreatic islet may be responsible for the T2D predisposition at this locus. A T2D GWAS meta-analysis in East Asian individuals (Spracklen et al.)[100] reported two T2D GWAS signals (rs33981001 and rs62508166) at this locus. Spracklen et al.[100] found that T2D signal rs33981001 was colocalized with *NKX6.3* eQTLs in pancreatic islet and T2D signal r62508166 was colocalized with *ANK1* eQTLs in subcutaneous adipose and skeletal muscle tissue, respectively. As rs516946 was in high LD with rs62508166 (R^2 =0.94 and D'= 1 in 1000G Phase 3 Asian panel; R^2=0.85 and D'=0.95 in FUSION), the colocalization between T2D GWAS variant rs516946 and the *ANK1* eQTL we observed in our data was consistent with Spracklen et al. Taken together, the findings from the Mahajan et al., Spracklen et al. and our work suggest that the multiple GWAS signals in this locus might contribute to T2D susceptibility by affecting different genes in different tissues and that the T2D signal tagged by rs516946 may act by changing the expression level of *ANK1* in skeletal muscle and subcutaneous adipose tissues.

Tissue/cell type composition is associated with the levels of physiological traits and molecular traits[27], [116], [117], and therefore can strongly impact and/or confound the physiological trait-molecular trait analysis. To account for the potential confounding effects introduced by tissue/cell type compositions, we used two approaches: estimating tissue or cell type proportions by using an external reference transcriptome and estimating surrogate variables. Compared to adjusting for the base set of covariates (without adjusting for tissue/cell type composition), adjusting for tissue/cell type composition using either approach

substantially reduced the number of significant physiological trait-molecular trait associations, suggesting the broad impact of the potential confounding effects of tissue/cell type on molecular trait associations with physiological traits. Comparing the results using different tissue/cell type composition, although the physiological trait-molecular trait associations overall showed concordant effect directions, the number of significant associations differed. In skeletal muscle tissue, using surrogate variables yielded less significant associations than using the tissue/fiber type estimates. In subcutaneous adipose tissue, using surrogate variables yielded the least number of significant associations, followed by the 17-component approach, and the 5-component approach. As our knowledge of the interplay between physiological and molecular traits at the cell-type level is still at a very primitive stage, we cannot tease out the spurious physiological trait-molecular trait associations driven by tissue/cell type composition heterogeneity across samples as well as the false negatives caused by overcorrection. Overall, these results emphasize the importance of taking into account tissue/cell type composition and also pose the pressing need for single-cell data, with which better composition estimates can be generated.

EIF4EBP1 displayed highly consistent associations with physiological traits in both direction and strength in skeletal muscle and subcutaneous adipose tissues. Higher *EIF4EBP1* was associated with a beneficial physiological trait profile (lower BMI, waist, C peptide, fasting serum insulin; higher HDL and Matsuda index). The mammalian target of rapamycin (mTOR) is a master regulator of cell growth and plays a pivotal role in metabolic processes in skeletal muscle, adipose and liver tissues upon postprandial elevation of insulin levels[118]. The EIF4EBP1 protein, once phosphorylated by the stimulation of mToR, stimulates protein synthesis[119]. Tsai et al.[120] studied the changes of *EIF4EBP1* upon high-fat challenge in mice. With four mice in each group, they observed that mRNA expression level of *EIF4EBP1* was significantly decreased in HFD-fed male skeletal muscle (p-value < 0.001) and adipose (p-value < 0.01) tissues, but not in female skeletal muscle or adipose tissues. We did not observe the expression level of *EIF4EBP1* differentiated by sex, nor did we find differences in the association between *EIF4EBP1* and the physiological traits between males and females in either tissue.

Of the genes whose eQTL colocalized with T2D GWAS variants, *INHBB* was the only one for which we observed significant associations with physiological traits, and also for which we have generated experimental evidence that the T2D risk allele causes higher rates of transcription than the non-risk allele. Several lines of evidence have shown the role of adipose *INHBB* in obesity and insulin resistance. *INHBB* was down-regulated by diet-induced weight loss (p-value <0.001) in the subcutaneous adipose samples from 24 patients[105]. Hoggard et al.[121] showed that *INHBB* mRNA was reduced in the 24h-fasted mice when compared with the fed controls (p-value $\leq$ 0.01), and increased 12h after refeeding (p-value $\leq$ 0.01) with eight mice in each condition. They[121] also showed that in differentiated 3T3-L1 adipocytes (each condition with three replicates), insulin increased the expression of *INHBB* (p-value $\leq$ 0.05), while dexamethasone decreased the expression of *INHBB* (p-value $\leq$ 0.001) when compared with untreated control cells. The various lines of evidence and our results suggest that *INHBB* may not only respond to physiological changes but also mediate the genetic risk underlying the T2D GWAS variant rs11688682. We also note that the detected physiological trait-molecular trait associations may exist in any tissue or cell types present in the biopsies, as can be seen from the hsa-miR-122R-5p associations with ALT and GT in both tissues. Pirola et al.[111] discovered that hsa-miR-122 was upregulated (p-value $\leq$ 0.05) either in simple steatosis (SS) or non-alcoholic steatohepatitis (NASH) in a case-control study with 48 participants and replicated the associations in a larger validation cohort with 96 participants. One possible explanation is that ALT, GT, and hsa-miR-122-5p are released from hepatocytes into the circulating blood because of liver damage caused by ectopic fat deposition and insulin resistance. We also examined the chromatin states of the hsa-miR-122-5p flanking region across a variety of tissue or cell types (Figure2.7.27). hsa-miR-122-5p resides in an active enhancer region in liver, a strong transcription region in skeletal muscle and subcutaneous adipose tissues and a few other tissues or cell types, and a repressed polycomb region in the rest of the tissue or cell types. Thus, we cannot rule out the possibility that hsa-miR-122-5p is expressed in skeletal muscle tissue and subcutaneous adipose tissue at a lower level compared to liver and the associations exist in cells inherent to skeletal muscle tissue and subcutaneous adipose tissue.

The current colocalization results has several limitations. First, the eQTLs used for the colocalization analysis were identified in samples from Finnish participants, whereas the T2D GWAS summary statistics were derived from participants of a broader European-ancestry. Finns are less genetically similar to other European-ancestry individuals as compared to individuals within other European-ancestries[122]–[124], our colocalization analysis may have failed to capture genes underlying T2D GWAS variants that occur as very low frequency in Finnish. Second, we used marginal T2D associations for the colocalization analysis. For multi-signal T2D loci, testing for colocalization between each of the T2D signals and QTL associations may enable us to identify more colocalized QTLs and discover additional molecular mechanisms. Most of the independent signals in a GWAS locus discovered in large-scale GWAS meta-analyses have been separated using approximate conditional analysis[125], which heavily depends on the genetic similarity between the participant studies and the reference panel. The separated GWAS signals may not reflect the multiple causal variants underlying the locus, which may further influence the colocalization results based on them. Closer examination of the association patterns of the multiple GWAS signals and colocalized eQTLs is necessary to evaluate the evidence for colocalizations.

The multi-omic data in this study provides rich opportunities for other analyses. We can test for colocalization between QTLs of different types of molecular traits to gain insights into shared causal variants between miRNA and mRNA or between DNAme sites and mRNA. We can use causal inference tests to untangle how molecular traits may be causally related to each other. We can apply mediation analysis to look for genes or DNA methylation sites that may mediate the effect of a GWAS variant on a disease or trait.

In summary, we generated a multi-omic QTL catalog by applying QTL analyses to mRNA and miRNA expression and DNA methylation levels. Integrating this catalog with T2D GWAS signals, we identified potential mediator mRNAs and DNAme sites for T2D loci, providing strong candidates for further functional follow-up. This multi-omic QTL resource also provides the scientific community opportunities to functionally annotate their genetic variants of interest and investigate the interplay between the genome, epigenome, and

transcriptome.

2.5 **Data availability**

QTL associations will be made publically available once the manuscript is accepted.

2.6 **My contributions**

This project resulted from the efforts of many individuals from the FUSION tissue biopsy group over the years. FUSION tissue biopsy group collected tissue biopsies of skeletal muscle and subcutaneous adipose, performed genotype array, mRNA-seq, miRNA-seq, and DNA methylation array experiments, and processed data generated from these experiments. I participated in the quality control of the subcutaneous adipose mRNA-seq data, and the miRNA-seq data for both tissues. Swarooparani Vadlamudi and Dr. Karen Mohlke performed the transcriptional reporter assays.

I performed the *cis*-eQTL detection and colocalization analyses. Dr. Leland Taylor, Dr. Anne Jackson, and I contributed to the tissue/cell type heterogeneity adjustment approaches for the physiological trait-molecular trait association analysis. Dr. Leland Taylor estimated the tissue/cell type compositions for skeletal muscle tissue samples. Dr. Anne Jackson performed the surrogate variable analysis for both tissues and estimated the tissue/cell type compositions using the single-nuclei cell data for the subcutaneous adipose tissues (17-component approach). I estimated the tissue/cell type compositions using publicly available RNA-seq data for the subcutaneous adipose tissues (5-component approach). Dr. Anne Jackson and I worked together on the analysis of identifying molecular traits associated with physiological traits, where Dr. Anne Jackson performed the association tests, and I analyzed and interpreted the results. Except for the method for the transcriptional reporter assay, I wrote all the rest of the manuscript with the guidance of Dr. Laura Scott and created all the figures.

2.7 **Supplementary figures**

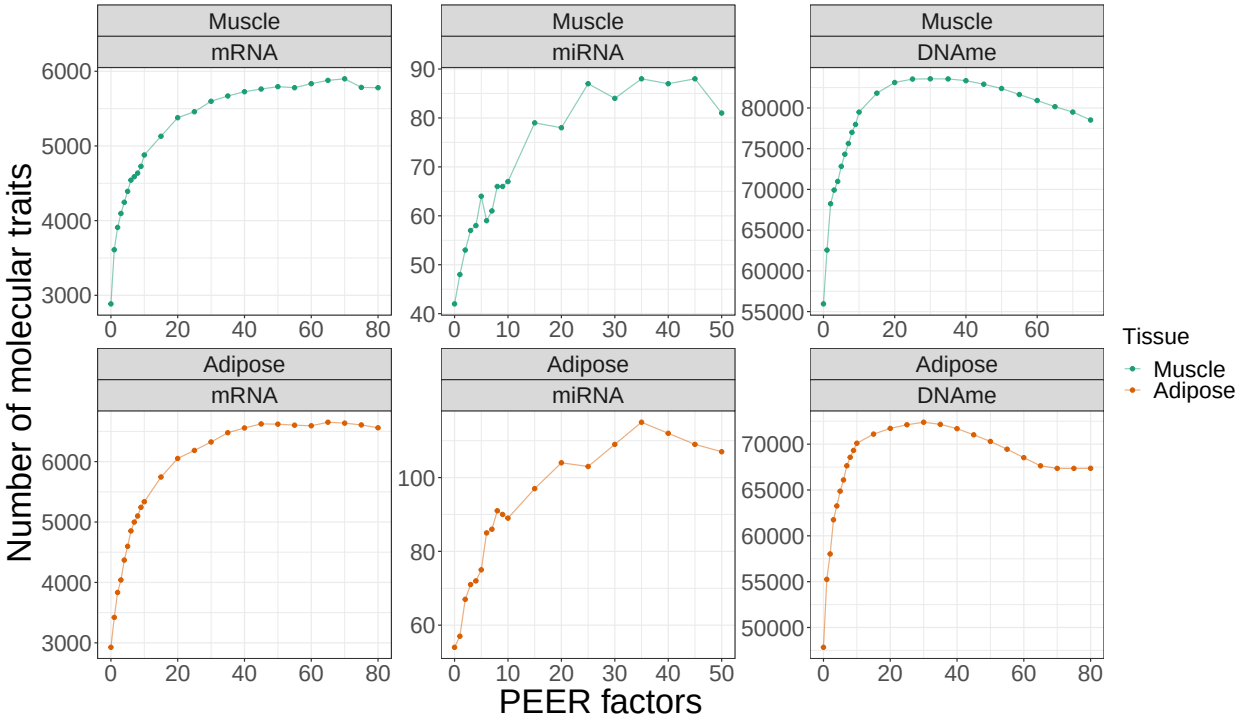


Figure 2.7.1. Scatterplots of the number of mRNAs, miRNAs and DNAme sites with ≥ 1 QTL at FDR≤ 1 % as a function of the number of PEER factors included as covariates.

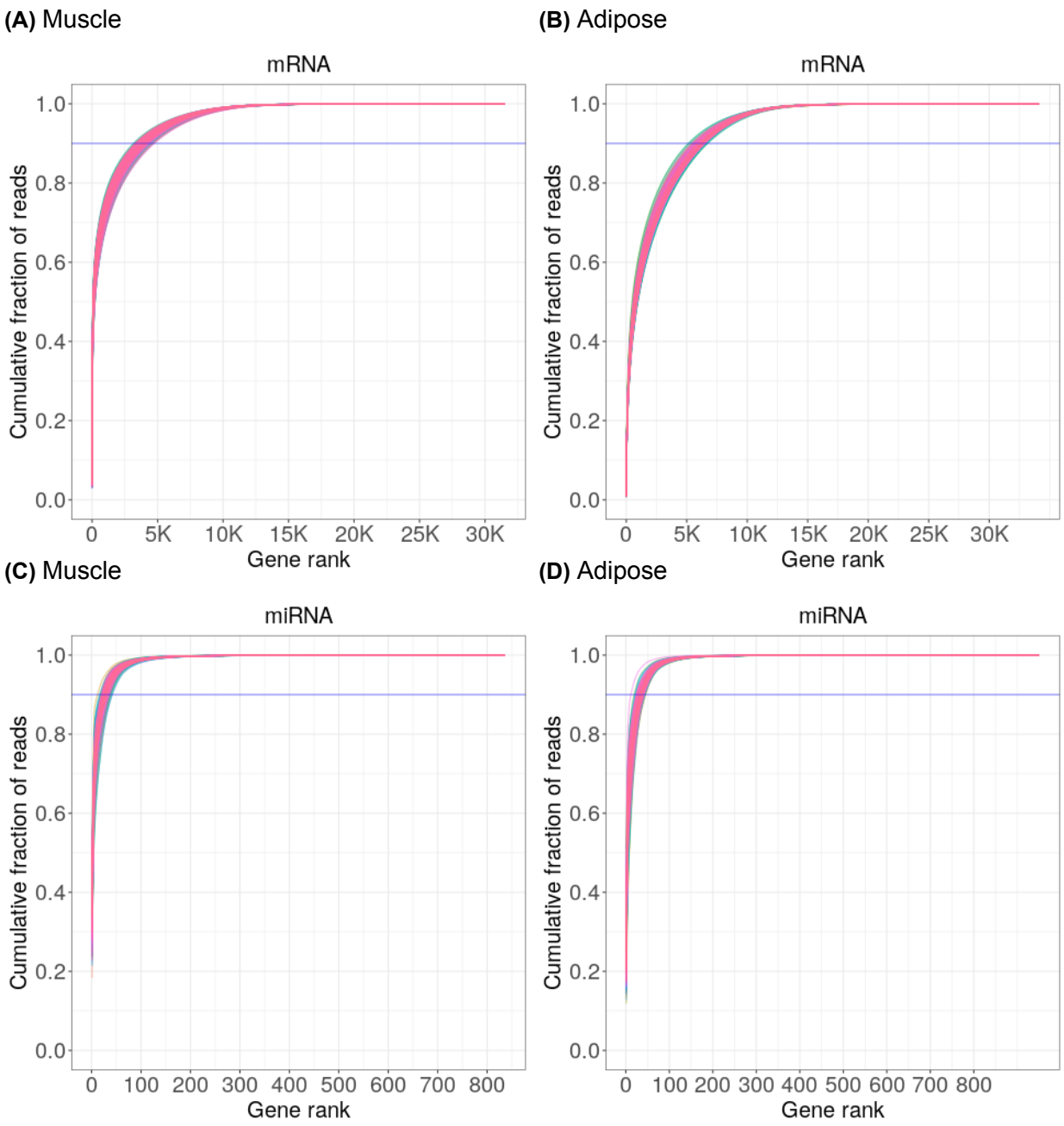


Figure 2.7.2. Cumulative fraction of reads as a function of the cumulative count of genes. Genes are ordered descendingly by read counts.

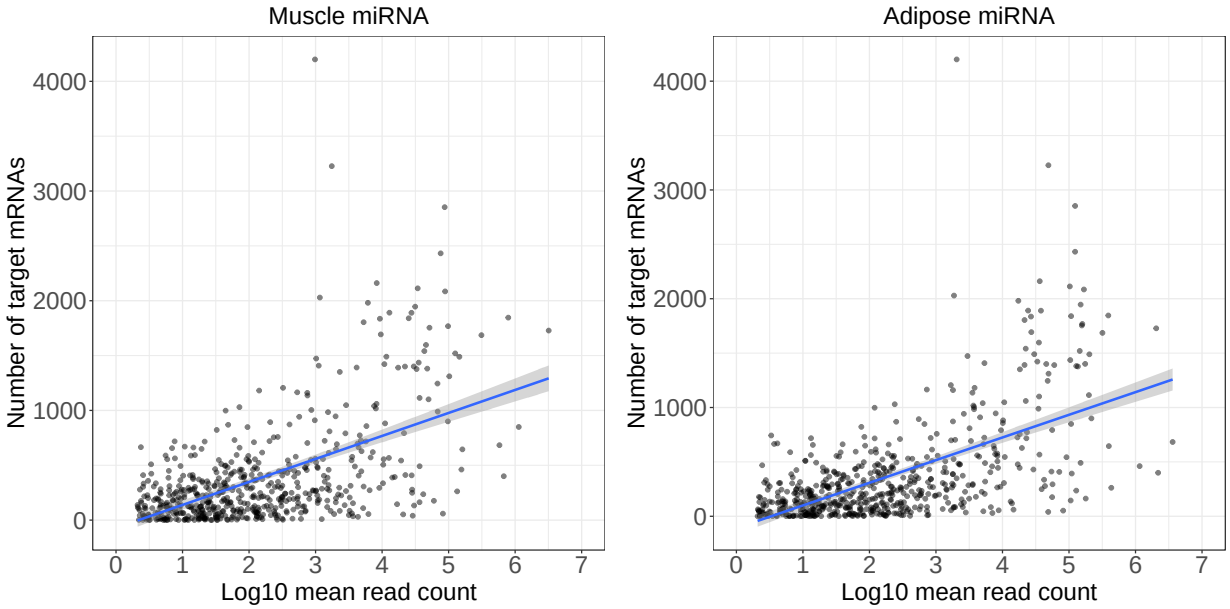


Figure 2.7.3. Relationship between the number of predicted target mRNAs and miRNA log10 mean read count

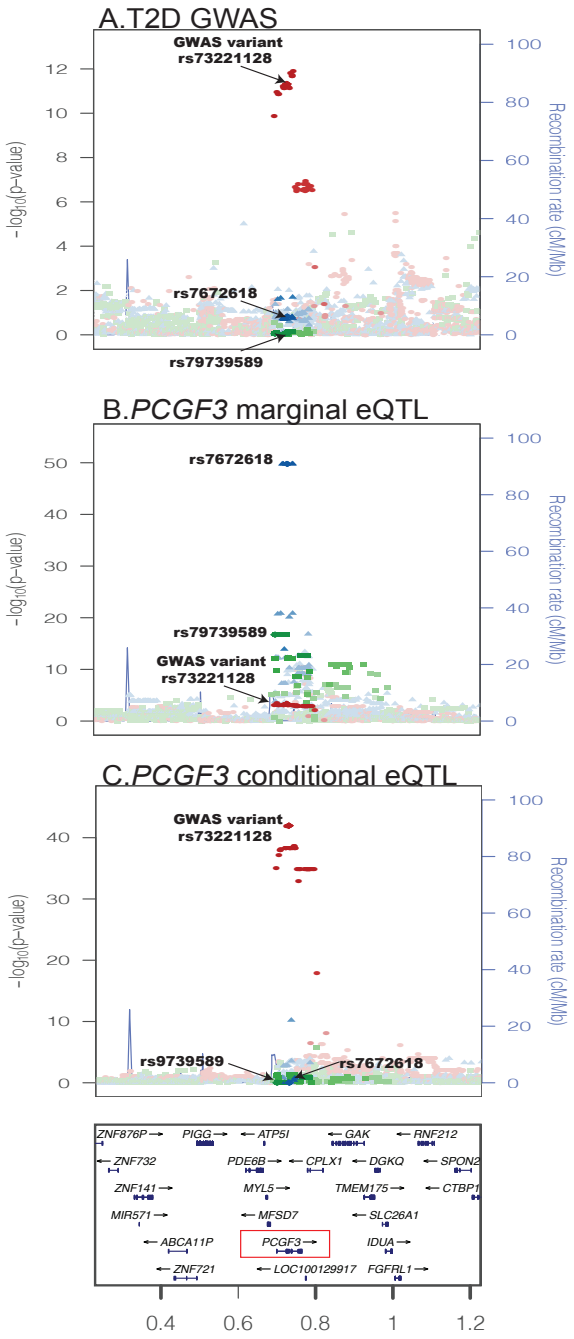


Figure 2.7.4. T2D GWAS signal is colocalized with the secondary eQTL for *PCGF3* in subcutaneous adipose tissue. Regional plots are colored by three independent eQTLs (represented by lead eQTL variants rs7672618, rs73221128, rs79739589) present in the FUSION data using FUSION LD. (A). Regional association plot for T2D meta-analysis from Mahajan et al.[33] at the rs73221128 locus (p-value = 4.5×10^{-12}); (B). Marginal eQTL association plot for *PCGF3* expression level. Marginal rs73221128-*PCGF3* association p-value = 2.59×10^{-4}. rs73221128 is in low LD R^2 (0.03) with the variant (rs7672618) that had the most significant marginal association with *PCGF3* expression (p-value = 2.87×10^{-47}); (C). After adjusting for rs7672618 and rs79739589, the T2D GWAS variant rs73221128 is more significantly associated with *PCGF3* (p-value = 1.10×10^{-42}).

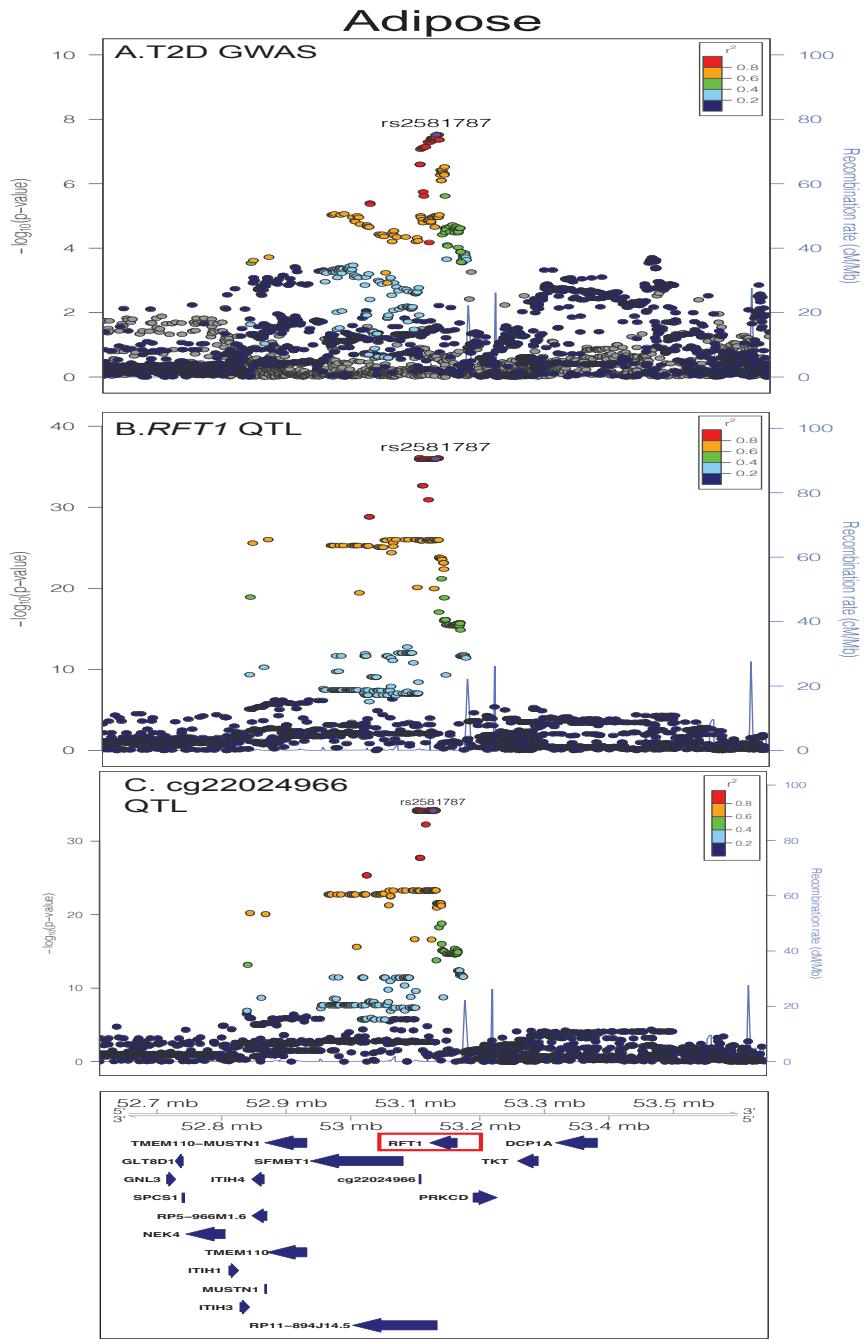


Figure 2.7.5. T2D GWAS signal rs2581787 is colocalized with the eQTL for *RFT1* and its nearby DNAme site cg22024966 in subcutaneous adipose tissue. Regional plots are colored LD R^2 with T2D GWAS variant rs2581787 using FUSION LD. (A). Regional association plot for T2D meta-analysis from Mahajan et al.[33] at the rs2581787 locus (p-value = 3.00×10^{-8}); (B). Marginal eQTL association plot for *RFT1* expression level. Marginal rs2581787-*RFT1* association p-value = 1.06×10^{-36}; (C). Marginal mQTL association plot for cg22024966 methylation level. Marginal rs2581787-cg22024966 association p-value = 6.56×10^{-35}.

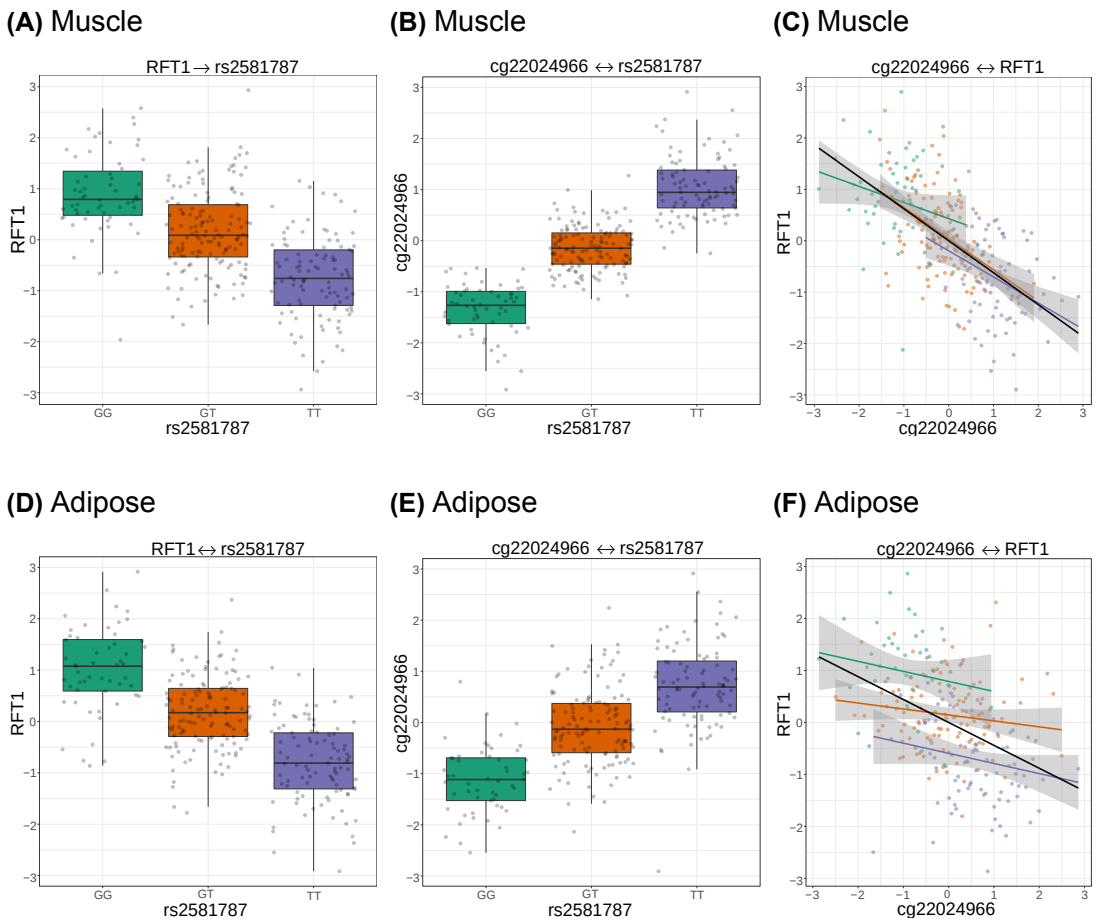
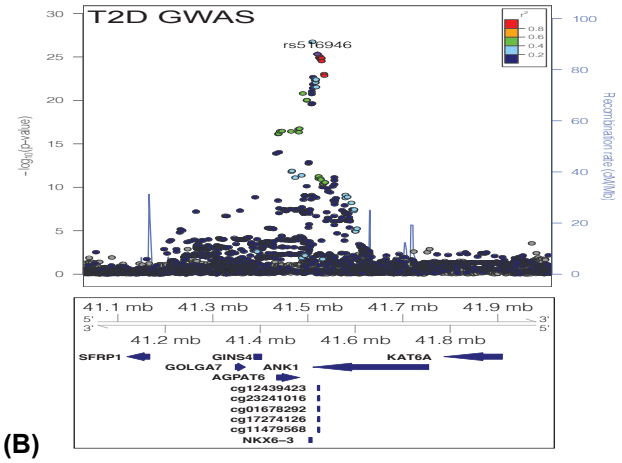
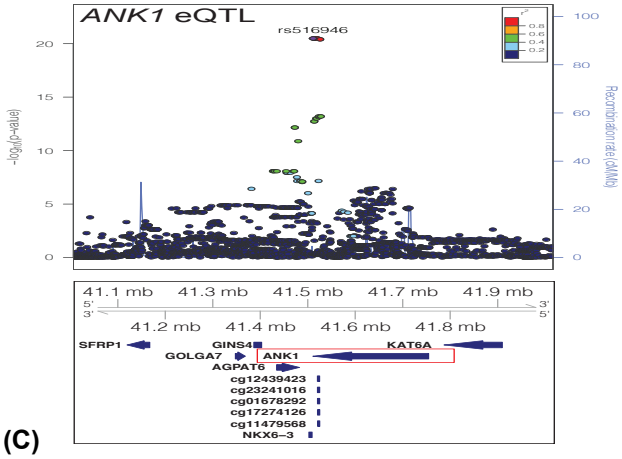


Figure 2.7.6. Effects of rs2581787 on *RFT1* and its nearby DNAme site cg22024966. (A). Box plot of residual *RFT1* expression levels by rs2581787 genotype in skeletal muscle tissue; (B). Box plot of residual cg22024966 methylation level by rs2581787 genotype in skeletal muscle tissue; (C). Scatter plot of residual *RFT1* expression (adjusted for PEER factors used in QTL mapping; y-axis) and residual cg22024966 methylation level (adjusted for PEER factors used in QTL mapping; x-axis) in skeletal muscle tissue, colored by rs2581787 genotypes. Linear regression lines for the relationship overall (black) and within each rs2581787 genotype (GG, green; GT orange; TT, purple; (D), (E), (F) are the same figures for subcutaneous adipose tissue.

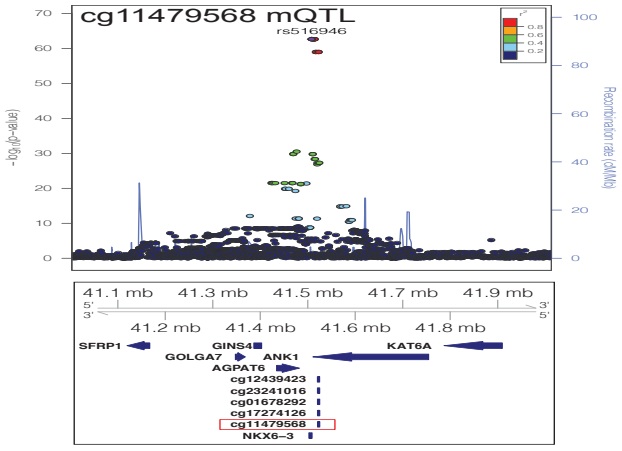
(A)


T2D GWAS
rs516946
r²
0.8
0.6
0.4
0.2
-log10(p-value)
Recombination rate (cM/Mb)
41.1 mb
41.3 mb
41.5 mb
41.7 mb
41.9 mb
41.2 mb
41.4 mb
41.6 mb
41.8 mb
SFRP1
GINS4
KAT6A
GOLGA7
ANK1
AGPAT6
cg12439423
cg23241016
cg01678292
cg17274126
cg11479568
NKX6-3

(B)


ANK1 eQTL
rs516946
r²
0.8
0.6
0.4
0.2
-log10(p-value)
Recombination rate (cM/Mb)
41.1 mb
41.3 mb
41.5 mb
41.7 mb
41.9 mb
41.2 mb
41.4 mb
41.6 mb
41.8 mb
SFRP1
GINS4
KAT6A
GOLGA7
ANK1
AGPAT6
cg12439423
cg23241016
cg01678292
cg17274126
cg11479568
NKX6-3

(C)


cg11479568 mQTL
rs516946
r²
0.8
0.6
0.4
0.2
-log10(p-value)
Recombination rate (cM/Mb)
41.1 mb
41.3 mb
41.5 mb
41.7 mb
41.9 mb
41.2 mb
41.4 mb
41.6 mb
41.8 mb
SFRP1
GINS4
KAT6A
GOLGA7
ANK1
AGPAT6
cg12439423
cg23241016
cg01678292
cg17274126
cg11479568
NKX6-3

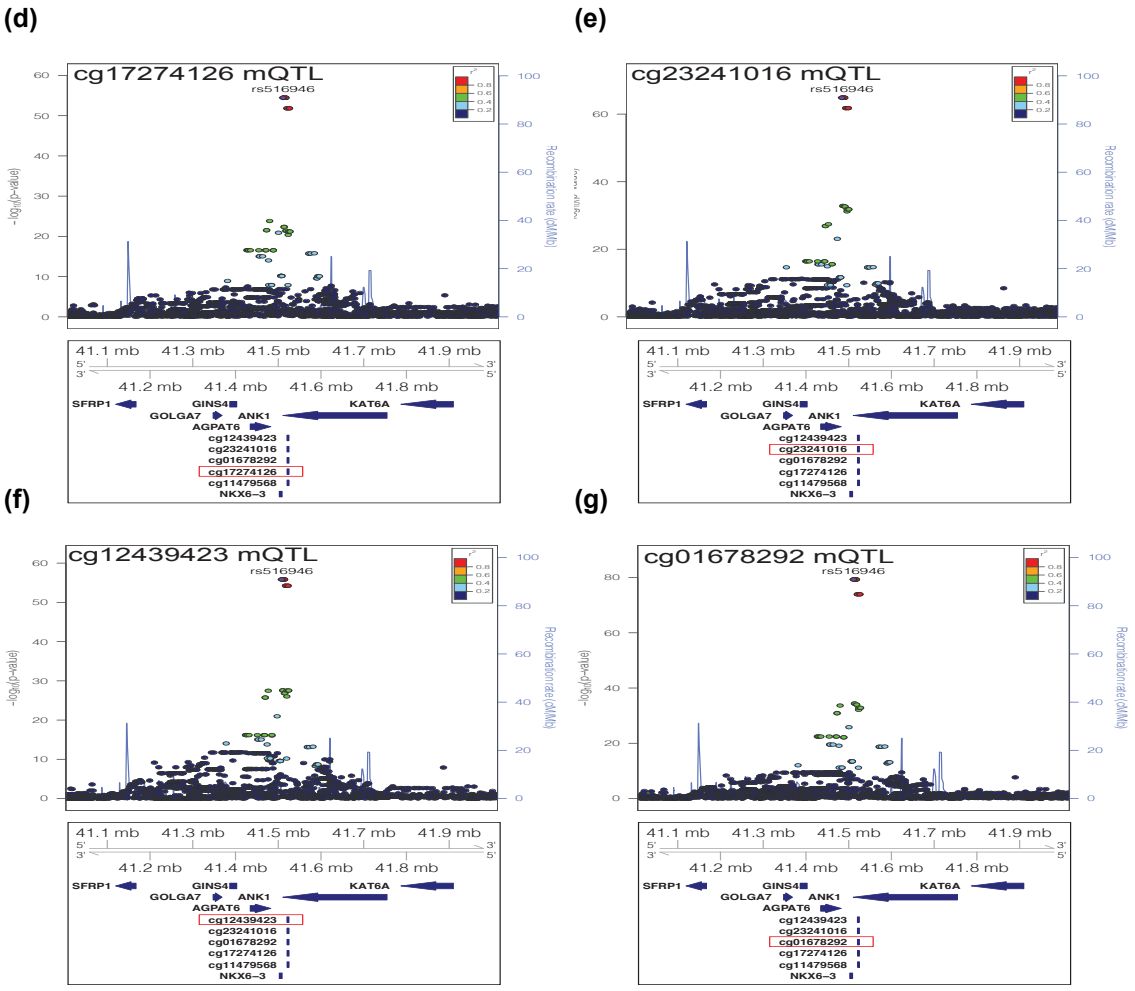


Figure 2.7.7. T2D GWAS signal rs516946 is colocalized with QTLs of *ANK1* and its nearby DNAme sites in subcutaneous adipose tissue. Regional plots are colored LD R^2 with T2D GWAS variant rs516946 using FUSION LD. (A). Regional association plot for T2D meta-analysis from Mahajan et al.[33] at the rs516946 locus (p-value = 4.70×10^{-26}); (B). Marginal eQTL association plot for ANK1 expression levels. rs516946-*ANK1* association p-value = 3.28×10^{-21}; (C). Marginal mQTL association plot for cg11479568 methylation levels. rs516946-cg11479568 association p-value = 2.14×10^{-63}; (D). Marginal mQTL association plot for cg17274126 methylation levels. rs516946-cg17274126 association p-value = 3.51×10^{-55}; (E). Marginal mQTL association plot for cg23241016 methylation levels. rs516946-cg23241016 association p-value = 1.41×10^{-65}; (F). Marginal mQTL association plot for cg12439423 methylation levels. rs516946-cg12439423 association p-value = 1.43×10^{-56}; (G). Marginal mQTL association plot for cg01678292 methylation levels. rs516946-cg01678292 association p-value = 5.01×10^{-80}.

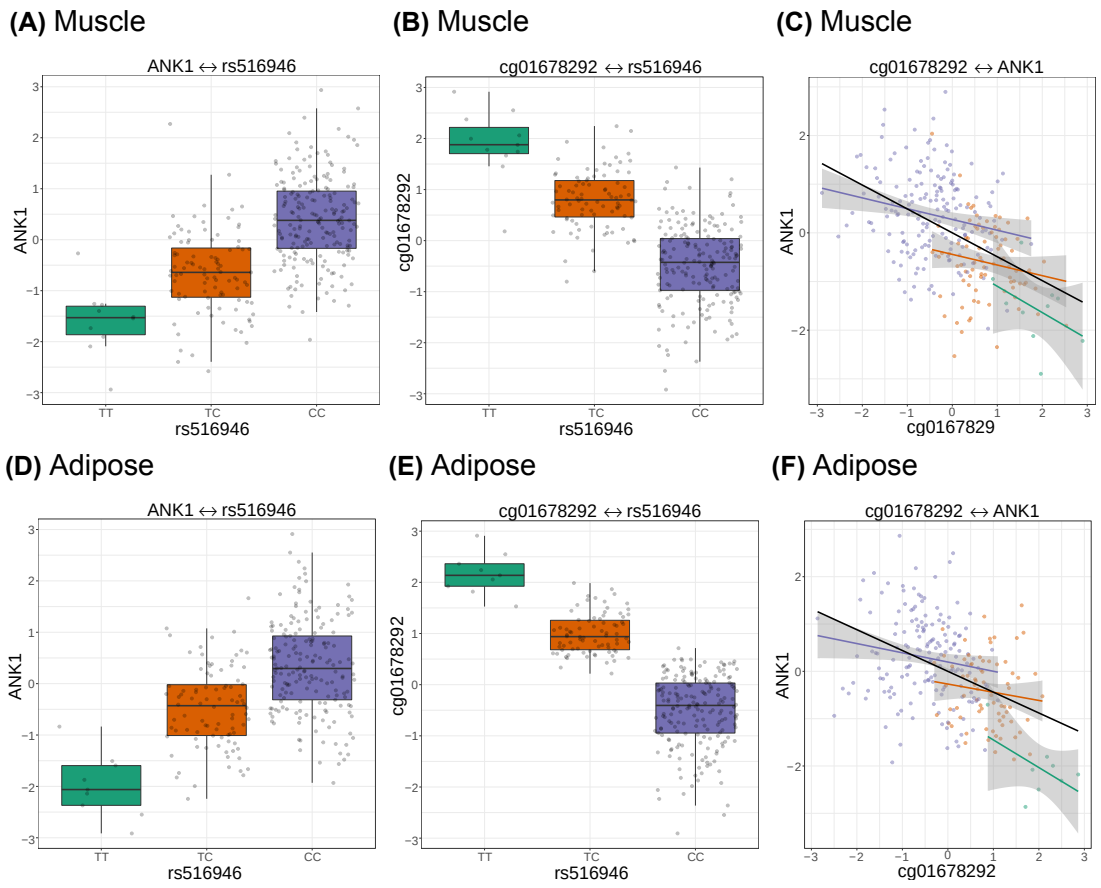


Figure 2.7.8. Effects of rs516946 on *ANK1* and its nearby DNAme site cg01678292. (A). Box plot of residual *ANK1* expression levels by rs516946 genotype; (B). Box plot of residual cg01678292 methylation level by rs516946 genotype; (C). Scatter plot of residual *ANK1* expression (adjusted for PEER factors used in QTL mapping; y-axis) and residual cg01678292 methylation level (adjusted for PEER factors used in QTL mapping; x-axis, colored by rs516946 genotypes. Linear regression lines for the relationship overall (black) and within each rs516946 genotype (TT, green; TC, orange; CC, purple). (D), (E), (F) are the same figures for subcutaneous adipose tissue.

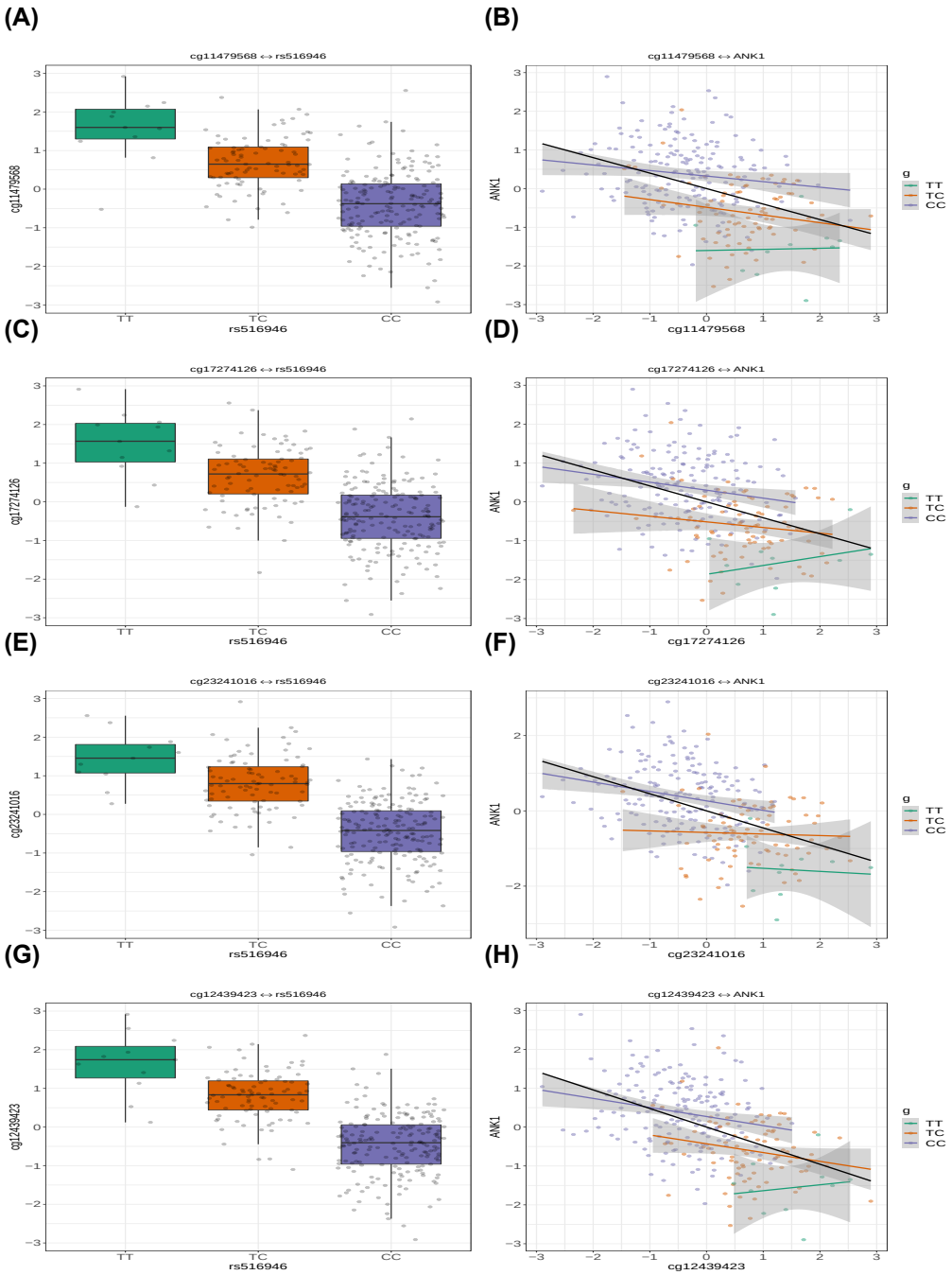


Figure 2.7.9. Effects of rs516946 on *ANK1* and its nearby DNAme sites in skeletal muscle tissue. (A). Box plot of residual cg11479568 methylation level by rs516946 genotype; (B). Scatter plot of residual *ANK1* expression (adjusted for PEER factors used in QTL mapping; y-axis) and residual cg11479568 methylation level (adjusted for PEER factors used in QTL mapping; x-axis, colored by rs516946 genotypes. (C). Box plot of residual cg17274126 methylation level by rs516946 genotype; (D). Scatter plot of residual *ANK1* expression and residual cg17274126 methylation level. (E). Box plot of residual cg23241016 methylation level by rs516946 genotype; (F). Scatter plot of residual *ANK1* expression and residual cg23241016 methylation level. (H). Box plot of residual cg12439423 methylation level by rs516946 genotype; (I). Scatter plot of residual *ANK1* expression and residual cg12439423 methylation level.

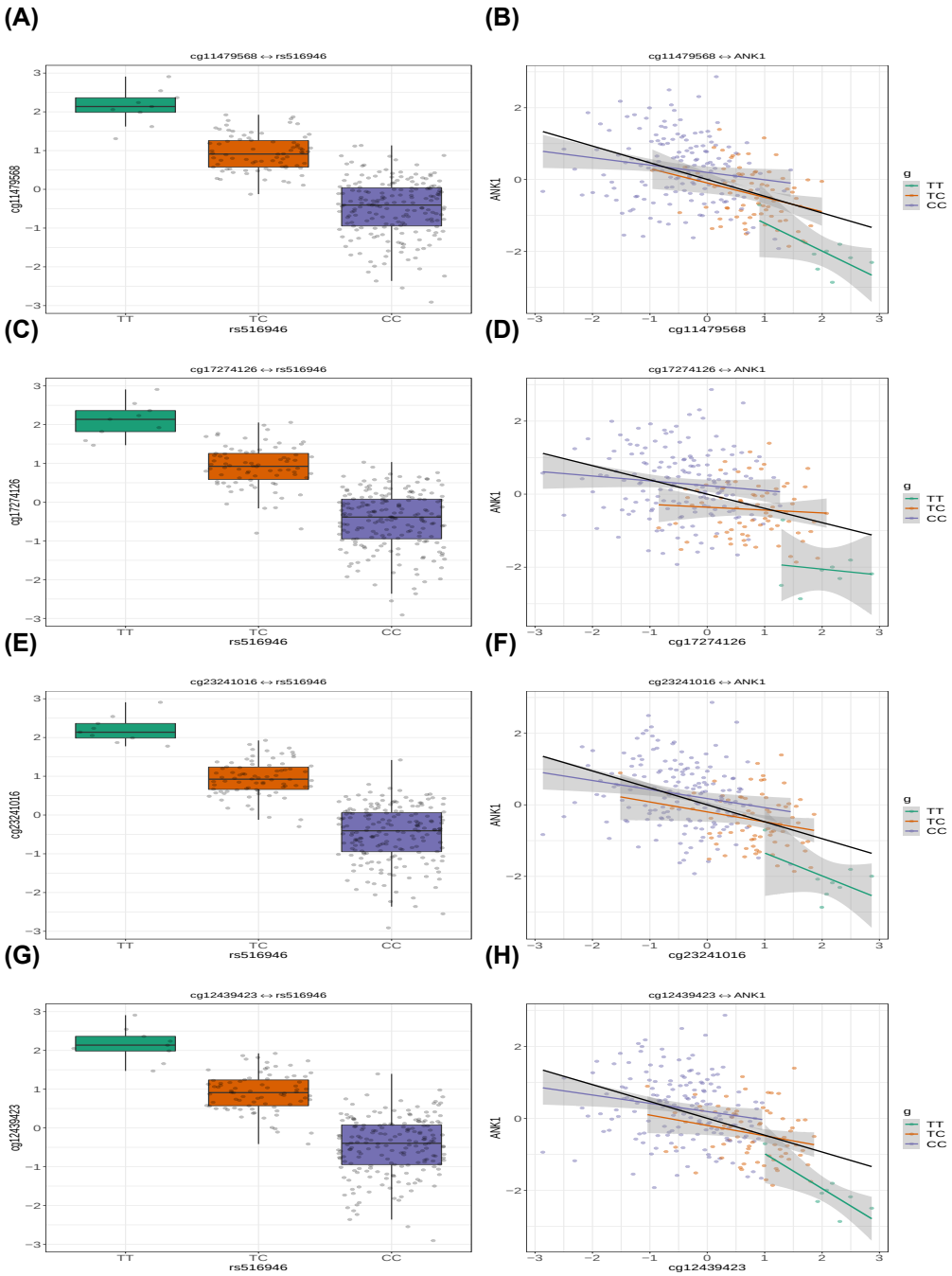
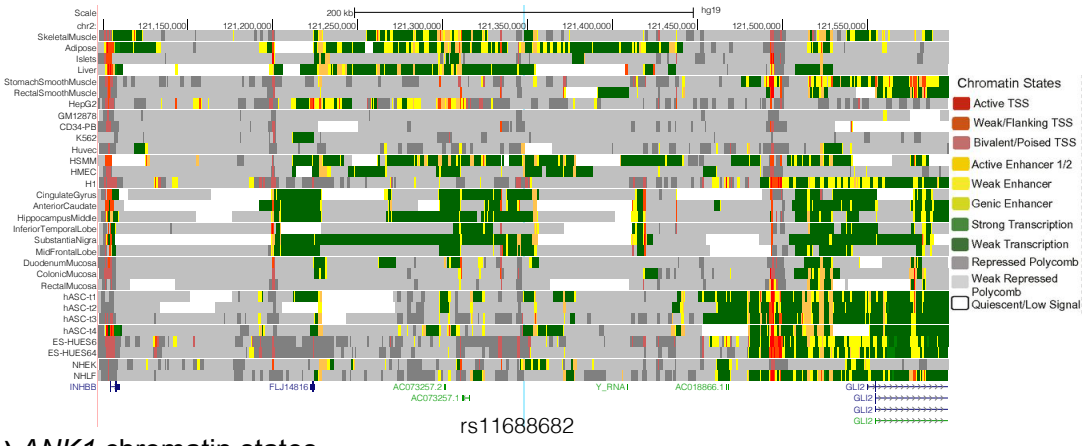
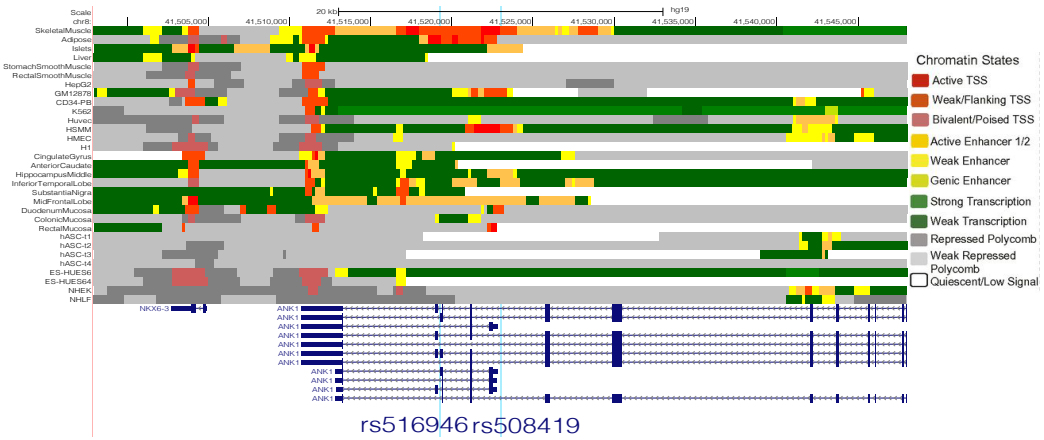


Figure 2.7.10. Effects of rs516946 on *ANK1* and its nearby DNAme sites in subcutaneous adipose tissue. (A). Box plot of residual cg11479568 methylation level by rs516946 genotype; (B). Scatter plot of residual *ANK1* expression (adjusted for PEER factors used in QTL mapping; y-axis) and residual cg11479568 methylation level (adjusted for PEER factors used in QTL mapping; x-axis, colored by rs516946 genotypes. (C). Box plot of residual cg17274126 methylation level by rs516946 genotype; (D). Scatter plot of residual *ANK1* expression and residual cg17274126 methylation level. (E). Box plot of residual cg23241016 methylation level by rs516946 genotype; (F). Scatter plot of residual *ANK1* expression and residual cg23241016 methylation level. (H). Box plot of residual cg12439423 methylation level by rs516946 genotype; (I). Scatter plot of residual *ANK1* expression and residual cg12439423 methylation level.

(a) *INHBB* chromatin states



(b) *ANK1* chromatin states



(c) *RFT1* chromatin states

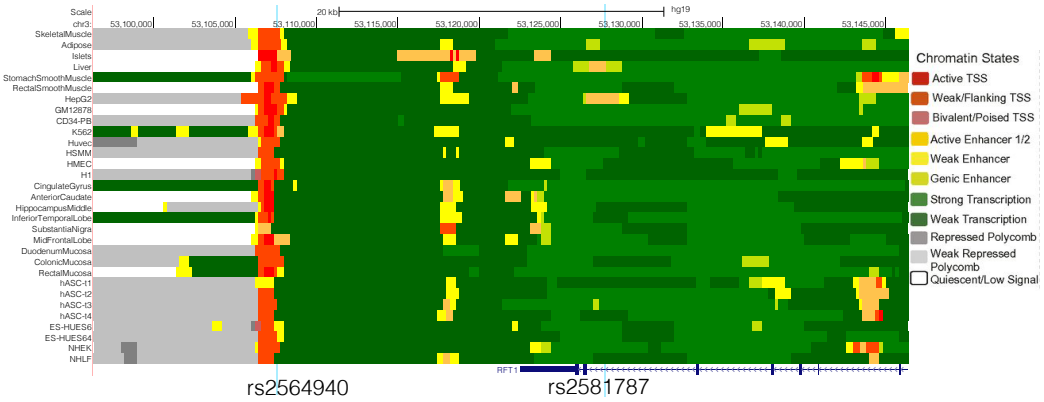


Figure 2.7.11. UCSC genome browser view of chromatin states (described in Varshney et al.[85]) near *INHBB*, *ANK1* and *RFT1* in diverse tissue and cell types.

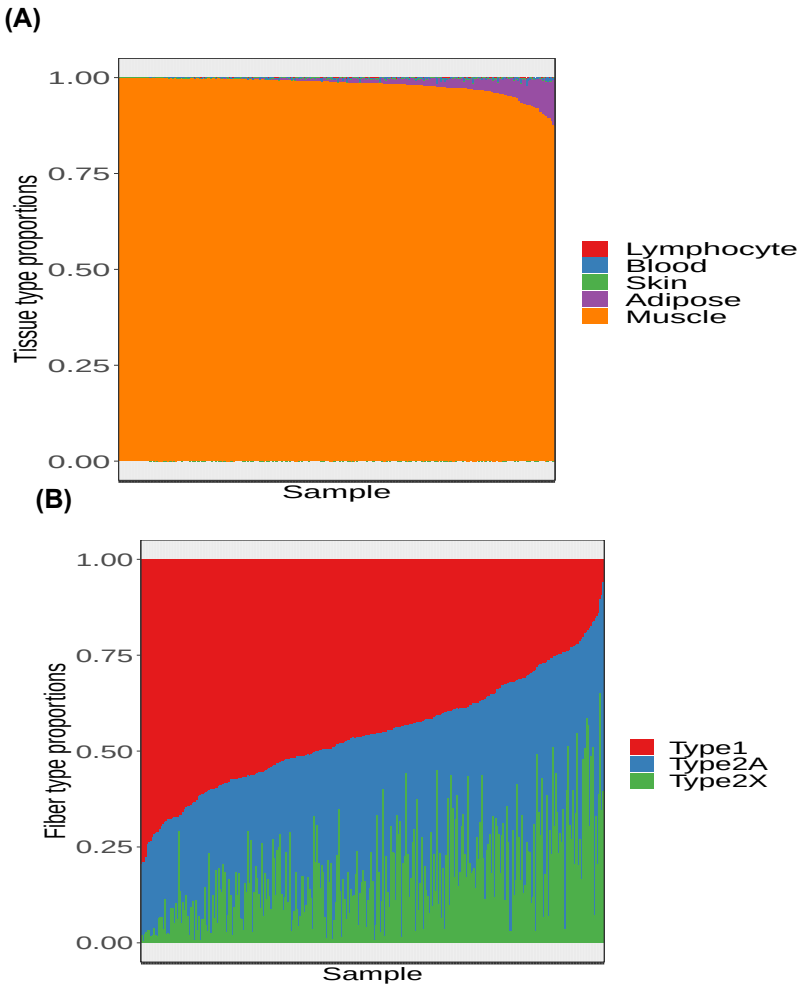


Figure 2.7.12. Tissue/cell-type proportion estimates using the tissue/fiber type approach for skeletal muscle tissue samples. (A). estimated proportions for lymphocytes(Lymphocyte), whole blood(Blood), skin not sun exposed (Skin), subcutaneous adipose (Adipose) and skeletal muscle (Muscle). (B). estimated proportions for Type 1 muscle fiber(Type 1), Type 2A muscle fiber(Type 2A), Type 2X muscle fiber(Type 2X).

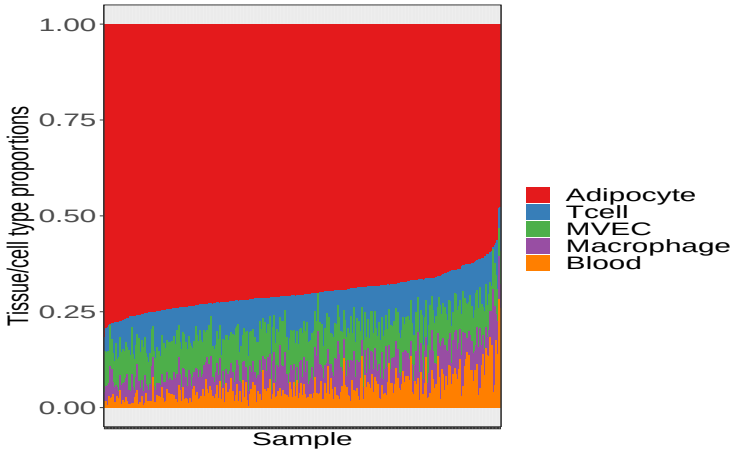


Figure 2.7.13. Tissue/cell-type proportion estimates using the 5-component approach for subcutaneous adipose tissue samples. Estimated proportions for adipocytes (Adipocyte), CD4+ T cells (Tcell), microvascular endothelial cells (MVEC), macrophages(Macrophage) and whole blood(Blood).

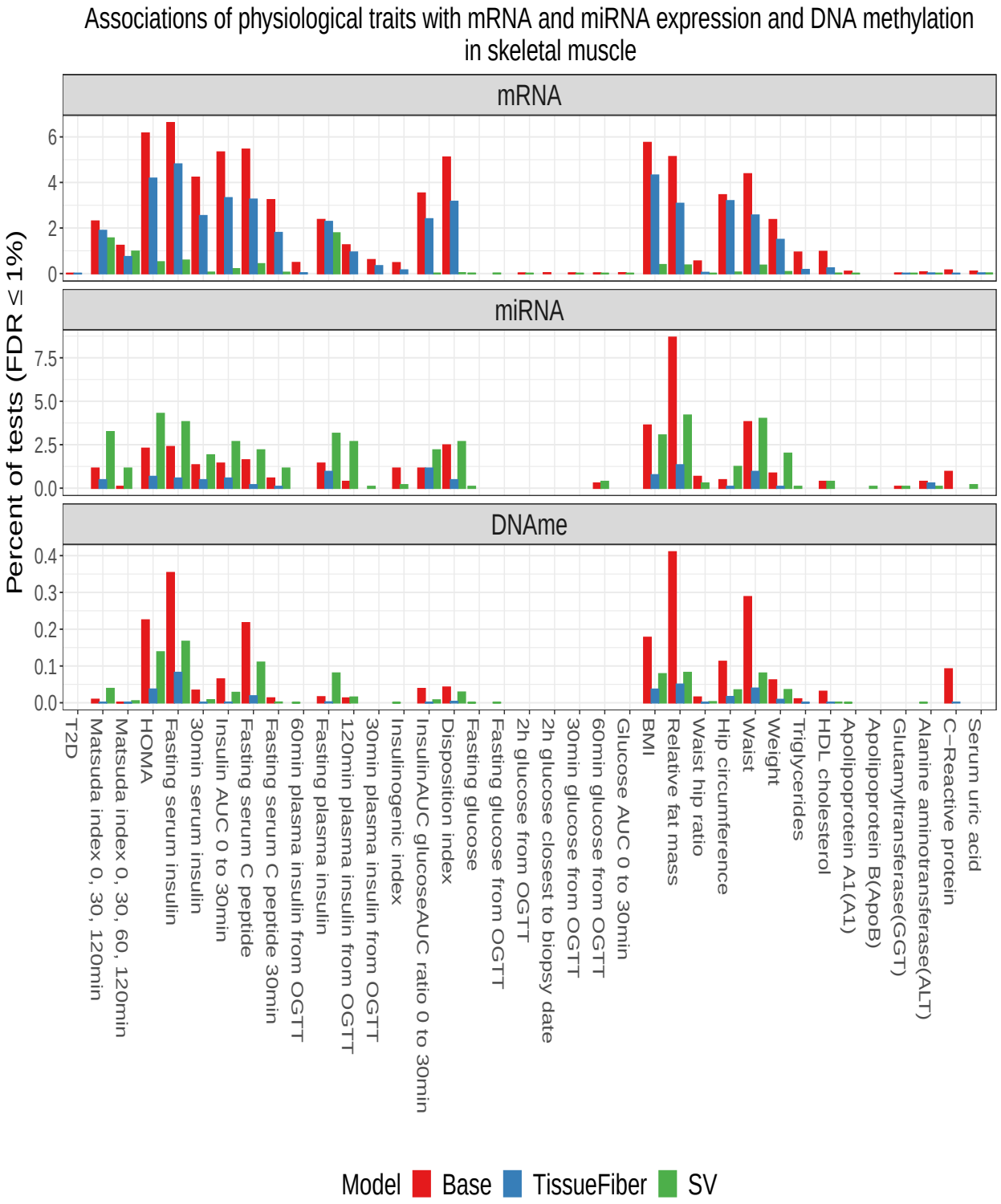


Figure 2.7.14. Percent of mRNAs/miRNAs/DNAme sites associated with the levels of physiological traits in skeletal muscle tissue at FDR$\leq$ 1% using different models. Models used different approaches to adjust for tissue/cell-type composition. Base: used a base set of covariates, without adjustment for composition. TissueFiber: used a base set of covariates and estimates of five tissue types and three muscle fiber types as the adjustment for composition. SV: used a base set of covariates and surrogate variables as the adjustment for composition.

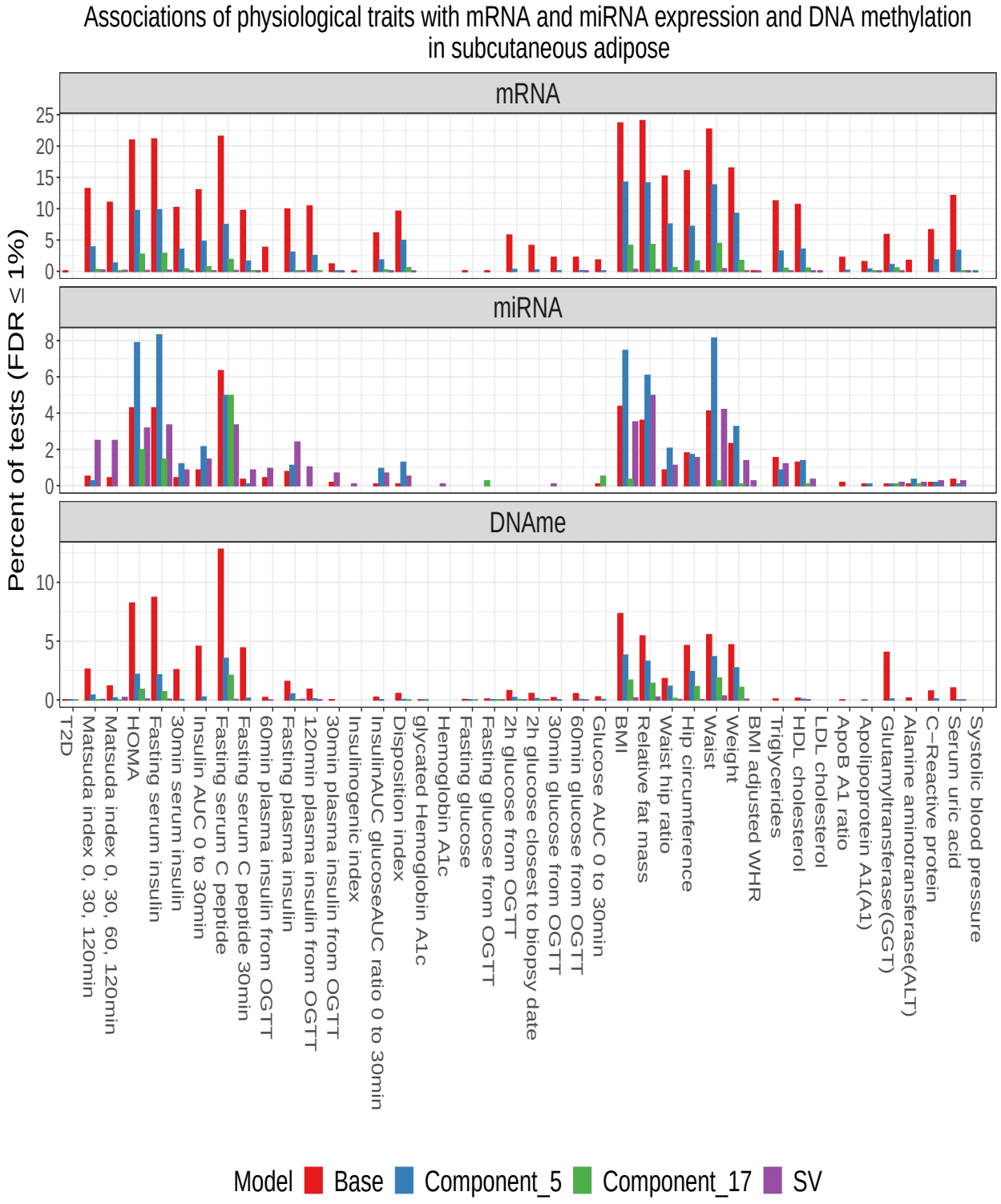


Figure 2.7.15. Percent of mRNAs/miRNAs/DNAme sites associated with the levels of physiological traits in subcutaneous adipose tissue at FDR$\leq$ 1% using different models. Models used different approaches to adjust for tissue/cell-type composition. Base: used a base set of covariates, without adjustment for composition. Component_5: used a base set of covariates and estimates of five components as the adjustment for composition. Component_17: used a base set of covariates and estimates of 17 components as the adjustment for composition. SV: used a base set of covariates and surrogate variables as the adjustment for composition.

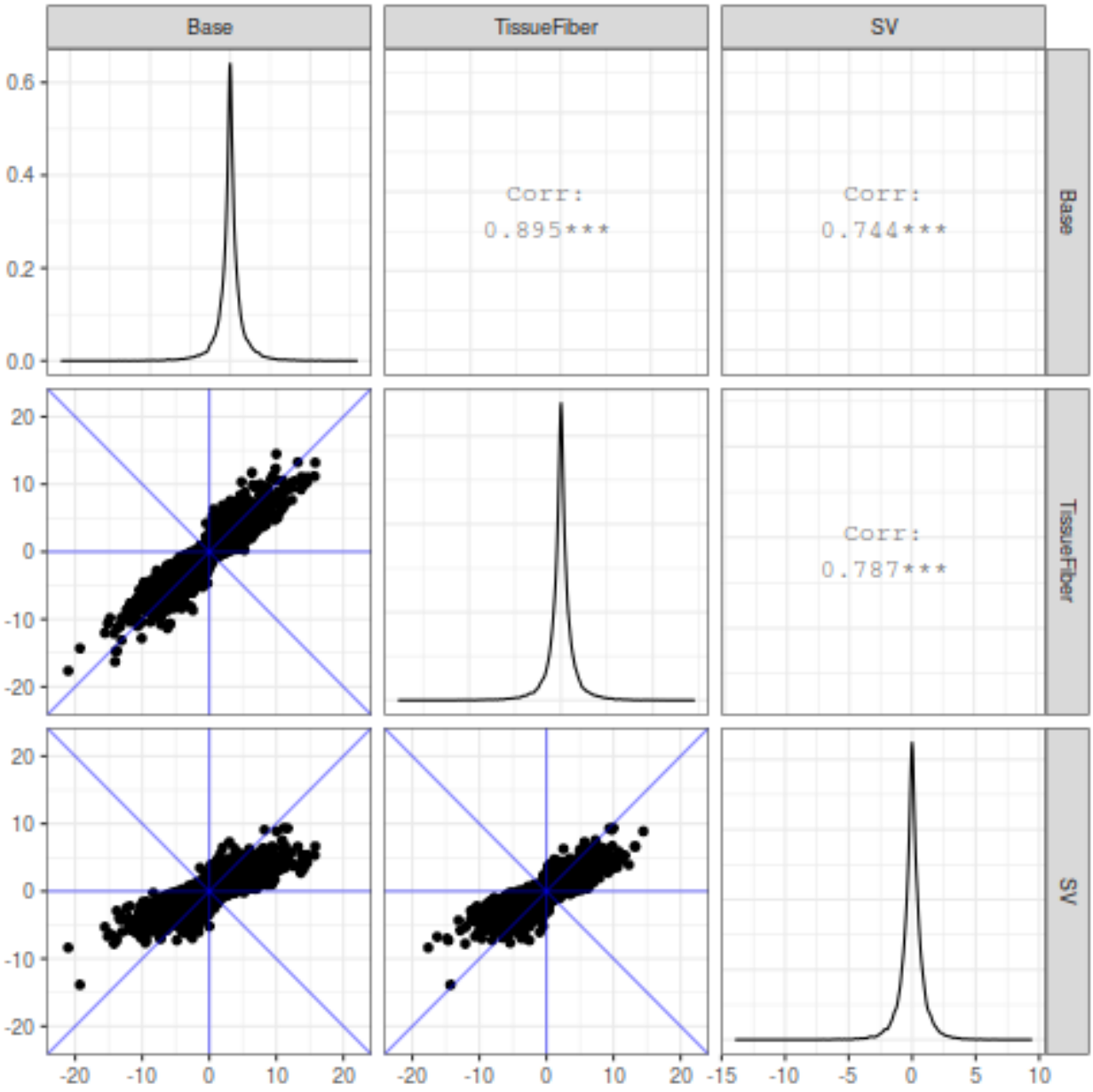


Figure 2.7.16. Pairwise scatterplot of -log10(p-value) of fasting serum insulin-mRNA associations between results using different models in skeletal muscle tissue. Models used different approaches to adjust for tissue/cell-type composition. Base: used a base set of covariates, without adjustment for composition. TissueFiber: used a base set of covariates and estimates of five tissue types and two three muscle fiber types as the adjustment for composition. SV: used a base set of covariates and surrogate variables as the adjustment for composition.

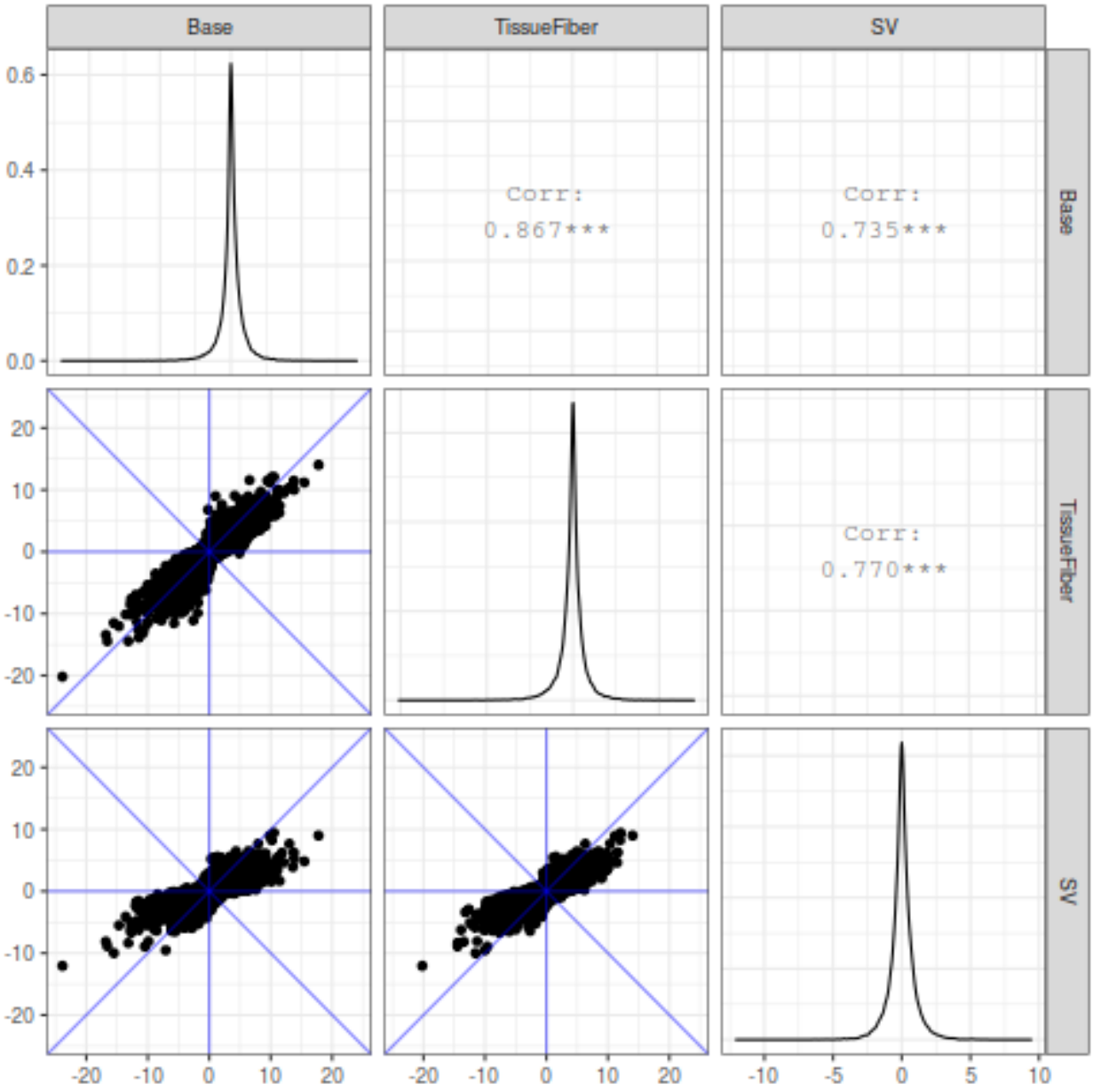


Figure 2.7.17. Pairwise scatterplot of -log10(p-value) of BMI-mRNA associations between results using different models in skeletal muscle tissue. Models used different approaches to adjust for tissue/cell-type composition. Base: used a base set of covariates, without adjustment for composition. TissueFiber: used a base set of covariates and estimates of five tissue types and two three muscle fiber types as the adjustment for composition. SV: used a base set of covariates and surrogate variables as the adjustment for composition.

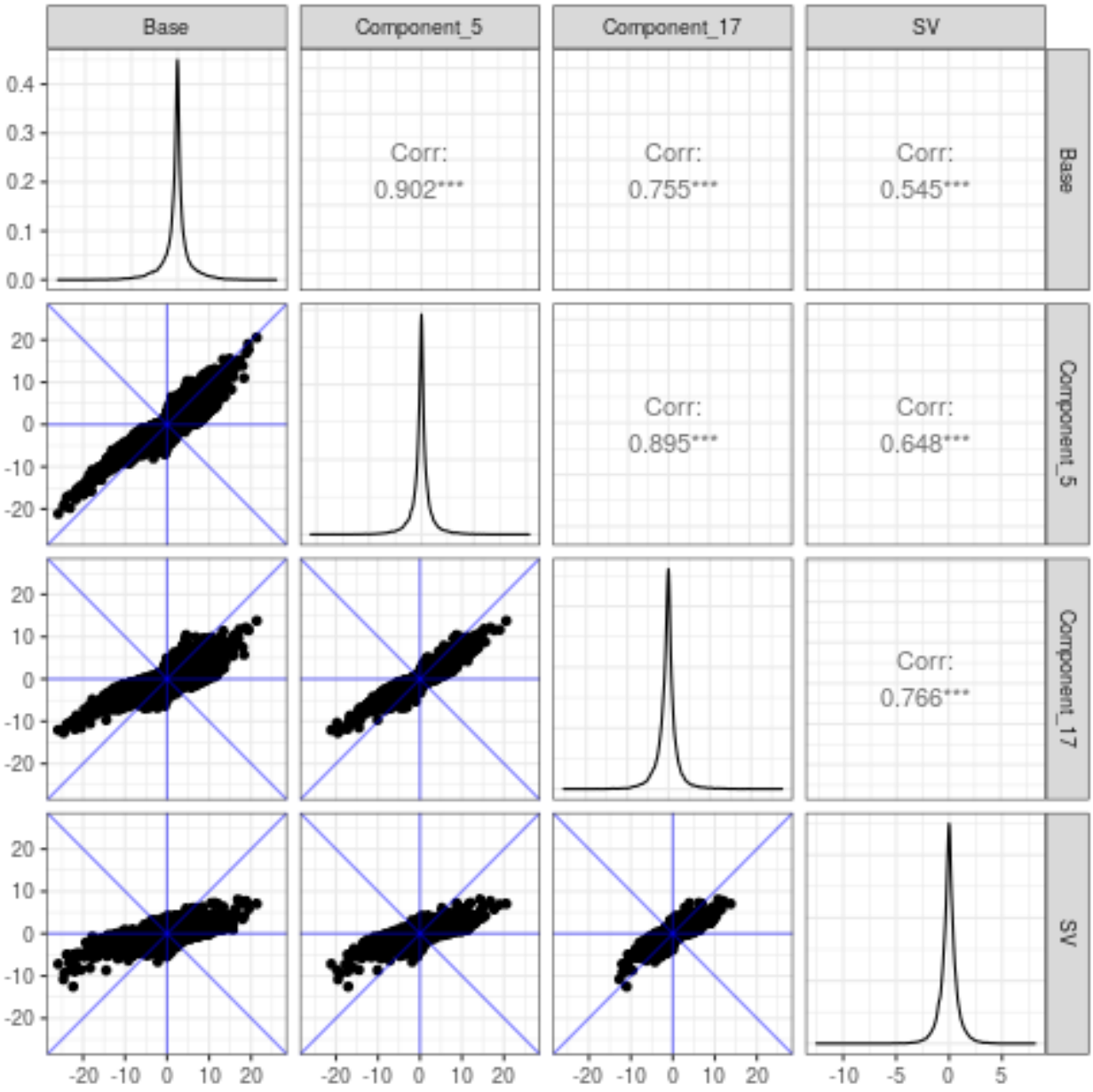


Figure 2.7.18. Pairwise scatterplot of -log10(p-value) of fasting serum insulin-mRNA associations between results using different models in subcutaneous adipose tissue. Models used different approaches to adjust for tissue/cell-type composition. Base: used a base set of covariates, without adjustment for composition. Component_5: used a base set of covariates and estimates of five components as the adjustment for composition. Component_17: used a base set of covariates and estimates of 17 components as the adjustment for composition. SV: used a base set of covariates and surrogate variables as the adjustment for composition.

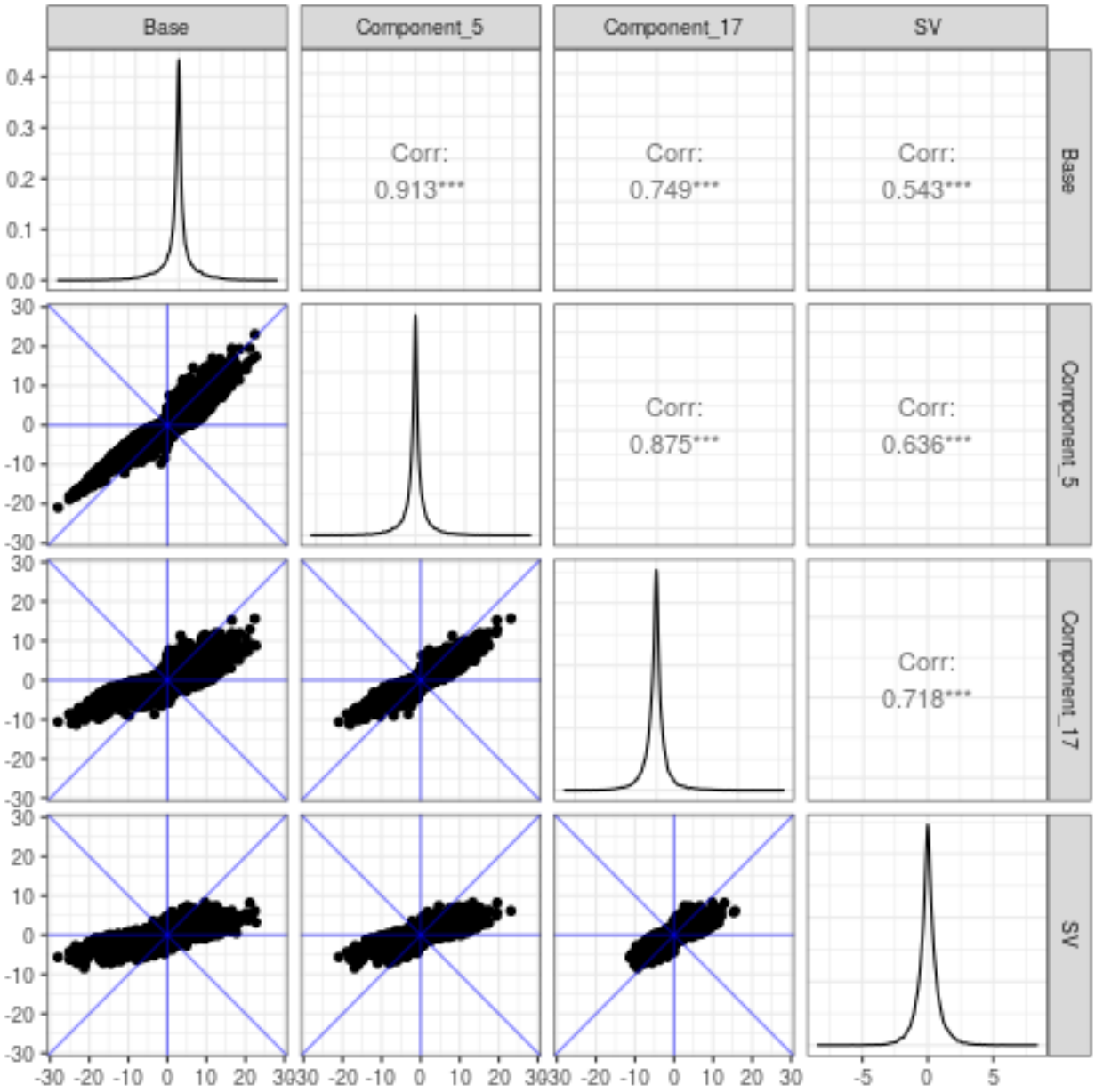


Figure 2.7.19. Pairwise scatterplot of -log10(p-value) of BMI-mRNA associations between results using different models in subcutaneous adipose tissue. Models used different approaches to adjust for tissue/cell-type composition. Base: used a base set of covariates, without adjustment for composition. Component_5: used a base set of covariates and estimates of five components as the adjustment for composition. Component_17: used a base set of covariates and estimates of 17 components as the adjustment for composition. SV: used a base set of covariates and surrogate variables as the adjustment for composition.

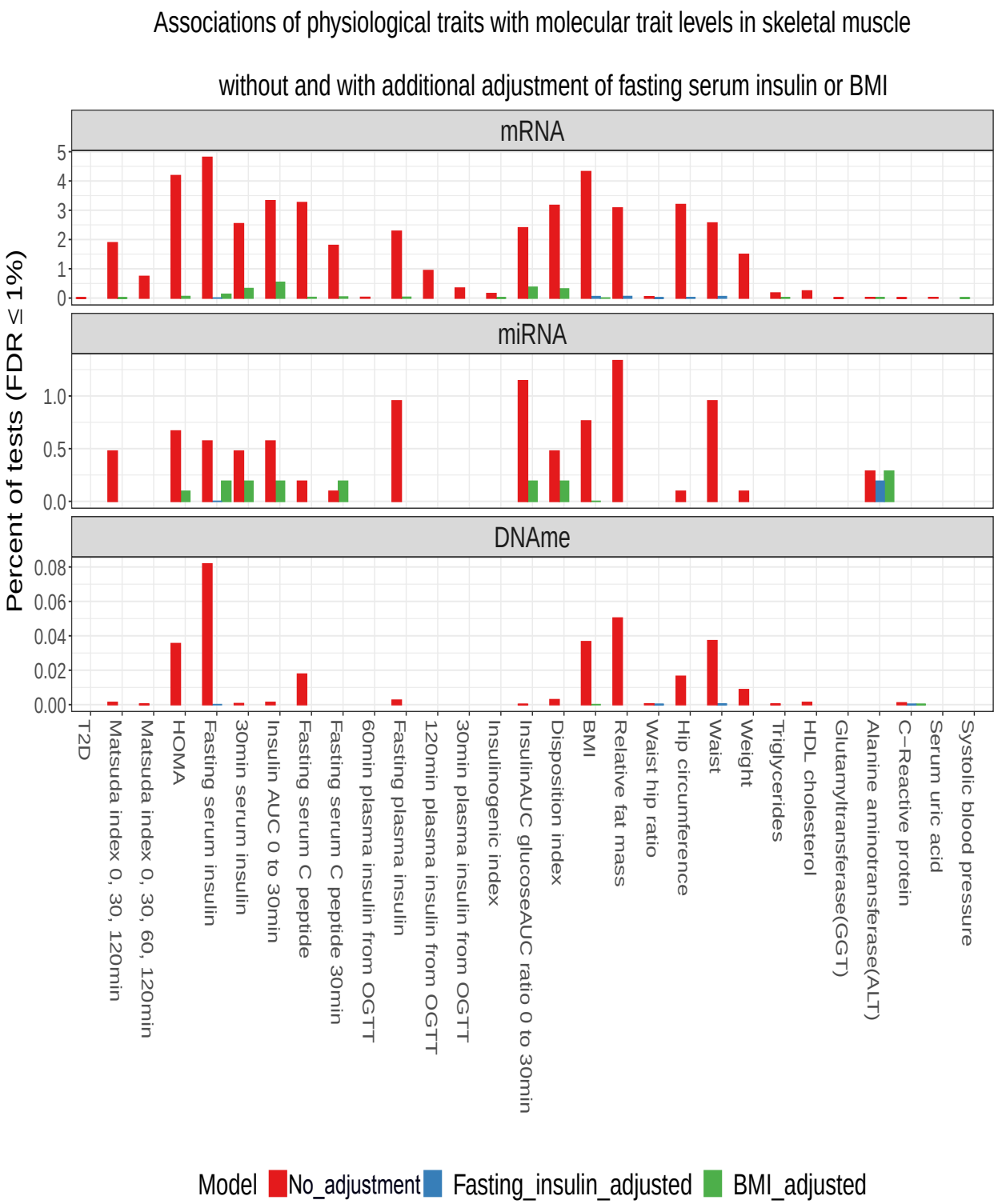


Figure 2.7.20. Percent of mRNAs/miRNAs/DNAme sites associated with the levels of physiological traits in skeletal muscle tissue without and with additional adjustment of fasting serum insulin or BMI. All models adjusted the tissue/cell-type proportion estimates obtained using the TissueFiber type approach.

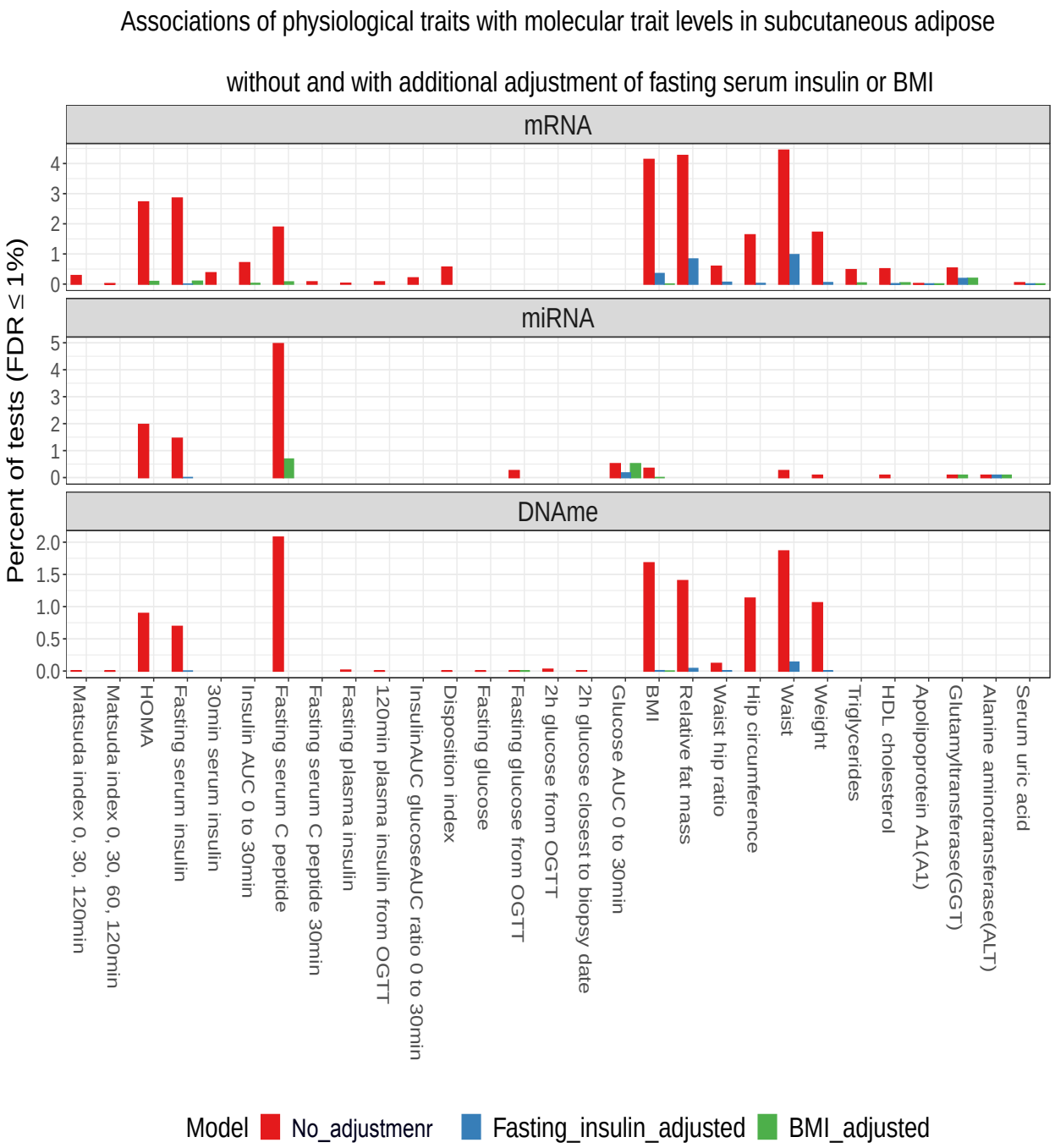


Figure 2.7.21. Percent of mRNAs/miRNAs/DNAme sites associated with the levels of physiological traits in subcutaneous adipose tissue without and with additional adjustment of fasting serum insulin or BMI. All models adjusted the tissue/cell-type proportion estimates obtained using the 17-component type approach.

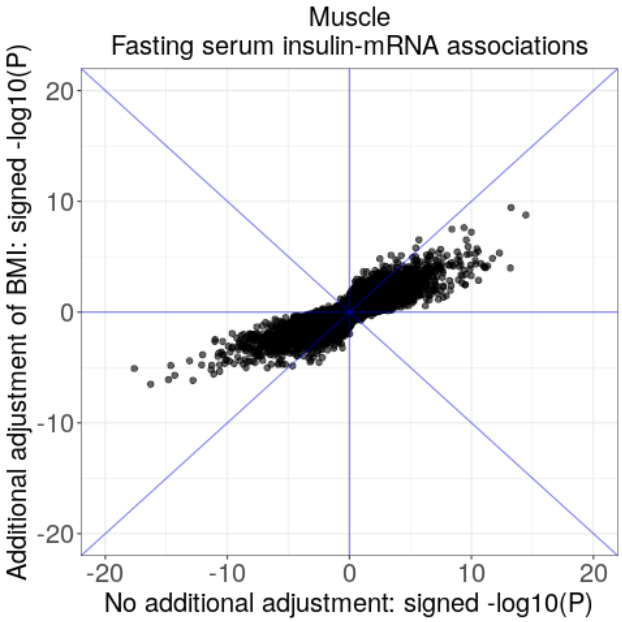


Figure 2.7.22. Effect of the additional adjustment of BMI on fasting serum insulin-mRNA associations in skeletal muscle tissue.. Scatterplot of -log10(p-value) for the associations between each mRNA and fasting serum insulin levels without adjusting for BMI (x-axis) and additionally adjusting for BMI (y-axis).

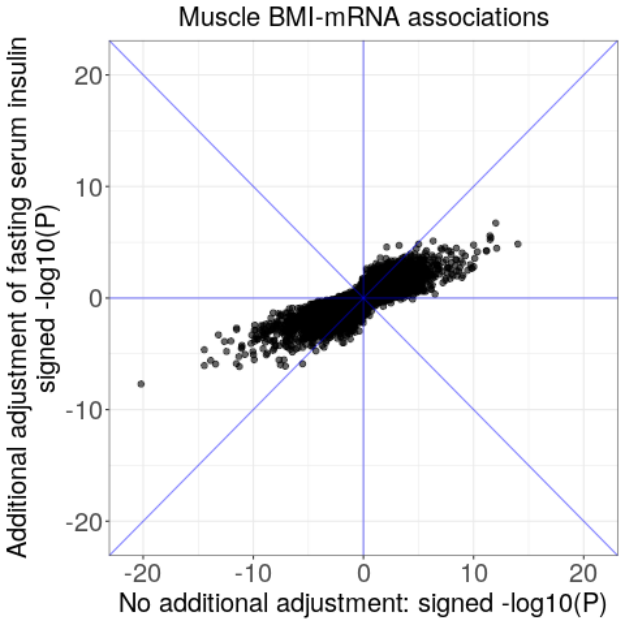


Figure 2.7.23. Effect of the additional adjustment of fasting serum insulin on BMI-mRNA associations in skeletal muscle tissue. Scatterplot of -log10(p-value) for the associations between each mRNA and BMI levels without adjusting for fasting serum insulin (x-axis) and additionally adjusting for fasting serum insulin (y-axis).

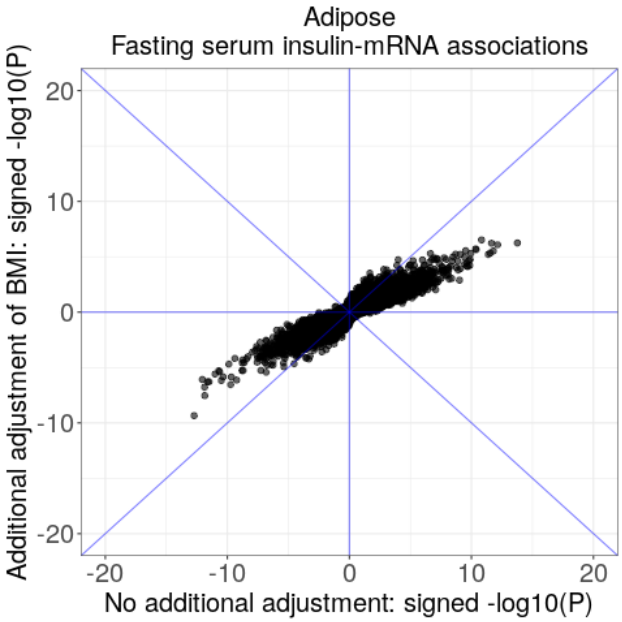


Figure 2.7.24. Effect of the additional adjustment of BMI on fasting serum insulin-mRNA associations in subcutaneous adipose tissue. Scatterplot of -log10(p-value) for the associations between each mRNA and fasting serum insulin levels without adjusting for BMI (x-axis) and additionally adjusting for BMI (y-axis).

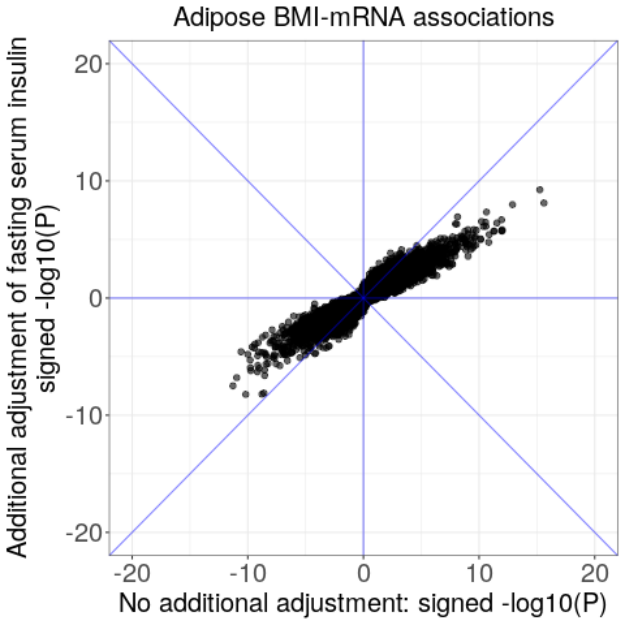


Figure 2.7.25. Effect of the additional adjustment of fasting serum insulin on BMI-mRNA associations in subcutaneous adipose tissue. Scatterplot of -log10(p-value) for the associations between each mRNA and BMI levels without adjusting for fasting serum insulin (x-axis) and additionally adjusting for fasting serum insulin (y-axis)

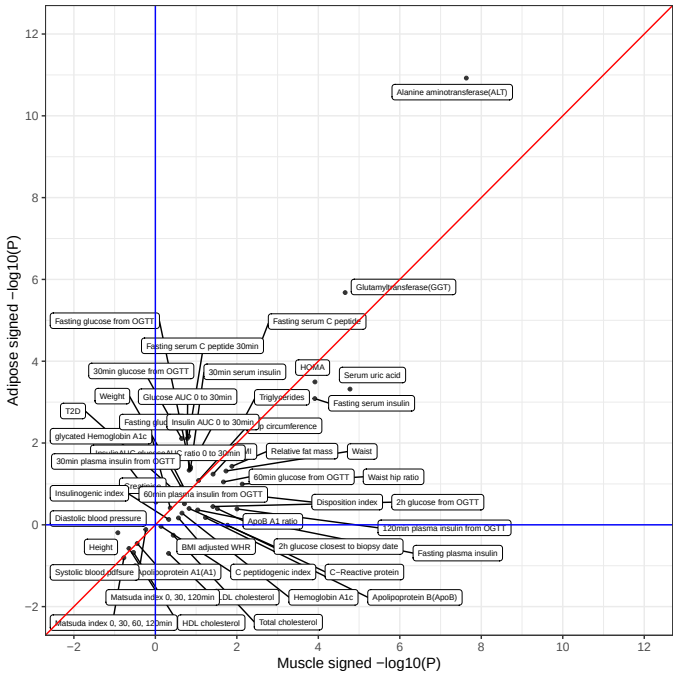


Figure 2.7.26. Associations of physiological traits with the hsa-miR-122-5p expression level in skeletal muscle and subcutaneous adipose tissues. Scatterplot of -log10(p-value) for the associations between each physiological trait and hsa-miR-122-5p expression levels in skeletal muscle(x-axis) and in subcutaneous adipose (y-axis).

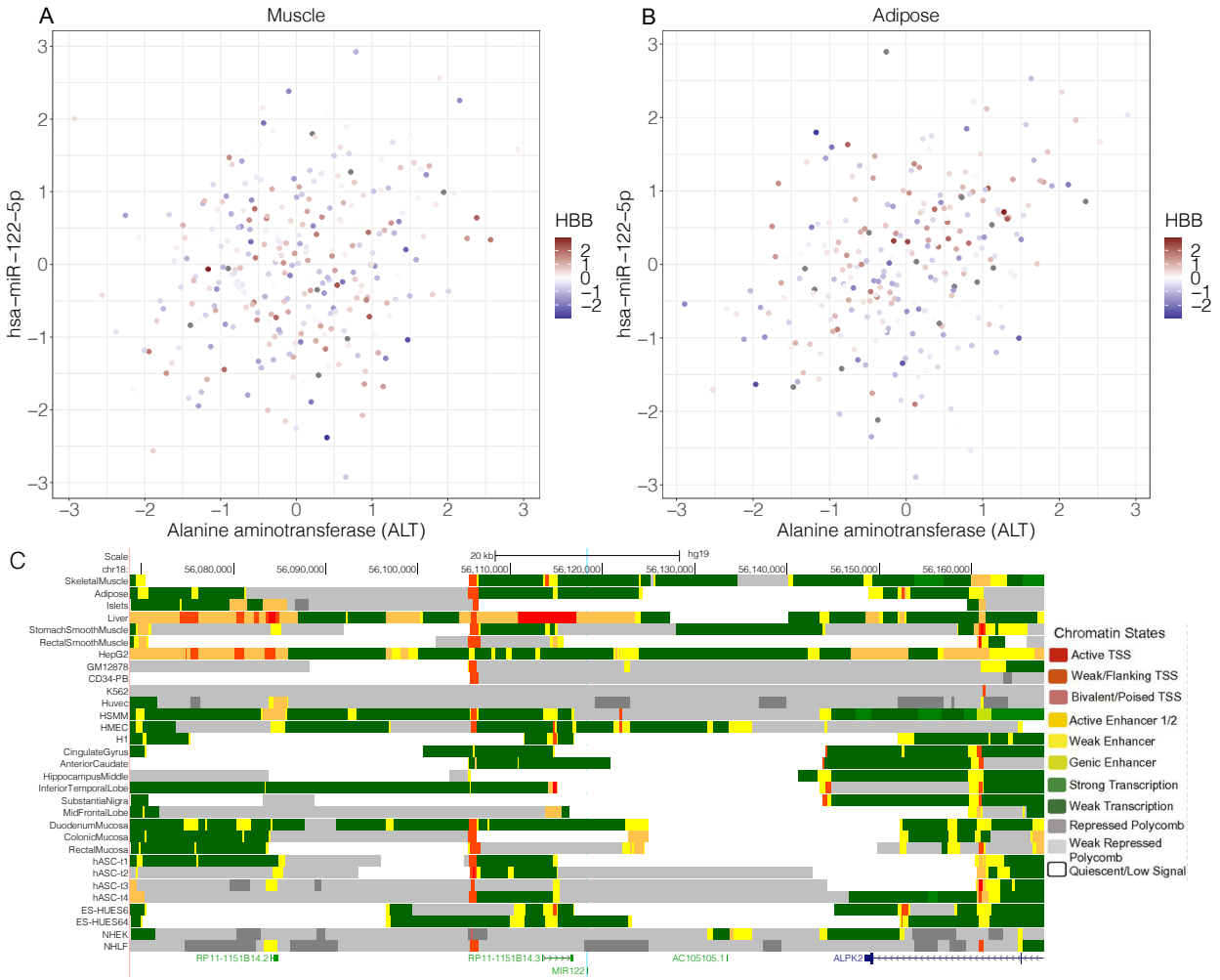


Figure 2.7.27. Association of hsa-miR-122-5p with Alanine aminotransferase (ALT) may reflect the cell-type heterogeneity in bulk-tissue biopsy samples (A).Scatterplot of ALT levels (x-axis) and hsa-miR-122-5p read per million mapped to miRNA (RPMMM) values (y-axis), colored by hemoglobin subunit beta (*HBB*) transcripts per million (TPM) values, in skeletal muscle tissue (Upper left) and subcutaneous adipose tissue (Upper Right). (B).UCSC genome browser view of chromatin states near hsa-miR-122-5p in diverse tissue and cell types.

2.8 **Supplementary Tables**

Table 2.8.1. Characterization of participants in the FUSION tissue biopsy study.

	Genotype	Muscle			Adipose		
		mRNA	DNAme	miRNA	mRNA	DNAme	miRNA
N	328	301	282	290	280	276	263
Sex = M (%)	182 (55.5%)	174 (57.8%)	159 (56.3%)	164(56.3%)	149(53.2%)	149(53.9%)	152(57.3%)
Age (mean ± sd)	59.97±7.80	59.91± 7.66	59.98±7.91	60.07±7.95	60.38±7.62	59.76±8.08	60.66±7.35
BMI (kg/m2; mean ± sd)	27.76±4.45	27.45± 4.13	27.63±4.27	27.63±4.24	27.63±4.31	28.01±4.53	27.77±4.32
Fasting Serum Insulin (mU/l; mean ± sd)	8.8±5.35	8.59± 5.20	8.82±5.42	8.8±5.38	8.49±4.75	8.98±5.47	8.93±5.49
Fasting Plasma Glucose (mmol/l; mean ± sd)	6.23±0.97	6.27± 0.78	6.19±0.72	6.22±0.73	6.23±1.01	6.24±1.02	6.27±1.03
Ever Smoker = Y (%)	48 (14.6%)	43 (14.3%)	41(14.5%)	42(14.4%)	42(15.0%)	42(15.2%)	40(15.0%)
Oral Glucose Tolerance Test Status (%)							
Normal Glucose Tolerance (NGT)	125 (38.1%)	108 (35.9%)	102(36.1%)	106(36.4%)	106(37.5%)	101(36.5%)	94(35.4%)
Impaired Fasting Glucose (IGF)	52(15.9%)	43 (13.4%)	47(16.6%)	51(17.5%)	41(14.5%)	47(17.0%)	38(14.3%)
Impaired Glucose Tolerance(IGT)	84(25.6%)	73 (24.3%)	80(28.3%)	74(25.4%)	79(28.0%)	75(27.1%)	76(28.6%)
Type 2 Diabetes (T2D)	67(20.4%)	77 (25.6%)	53(18.7%)	60(20.6%)	556(19.8%)	53(19.2%)	57(21.5%)

Table 2.8.2. T2D GWAS variants that were colocalized with eQTLs or mQTLs in muslce and adipose at RCP > 0.5

gene	variant	RCP	QTL rank	QTL chr	QTL pos	QTL REF	QTL ALT	QTL MAF	QTL beta	QTL p	GWAS EA	GWAS NEA	GWAS EAF	GWAS beta	GWAS p
ENSG00000011523	rs2723064	0.56	1	2	65279805	T	C	0.33	-0.36	4.04E-36	T	C	0.62	0.051	3.90E-15
ENSG00000163083	rs11688682	1.00	1	2	121347612	G	C	0.25	-0.27	1.51E-08	C	G	0.27	-0.058	1.40E-14
ENSG00000163933	rs2581787	0.63	1	3	53127677	G	T	0.44	-0.53	2.57E-30	T	G	0.56	0.036	3.00E-08
ENSG00000029534	rs516946	0.60	1	8	41519248	T	C	0.19	0.51	9.71E-28	T	C	0.23	-0.08	4.70E-26
ENSG00000141699	rs684214	0.70	1	17	40696915	C	T	0.18	0.22	3.39E-11	T	C	0.28	0.05	3.50E-12
ENSG00000185619	rs73221128	0.51	3	4	728111	C	T	0.04	-0.91	1.06E-33	T	C	0.045	0.11	4.80E-12
ENSG00000002726	rs7794796	0.60	1	7	150540196	C	T	0.39	0.51	1.76E-09	T	C	0.33	0.04	2.50E-09
ENSG00000149084	rs1061810	0.58	1	11	43877934	C	A	0.30	-0.75	4.56E-86	A	C	0.29	0.05	8.50E-13
ENSG00000163545	rs10157145	0.76	1	1	205261963	T	C	0.49	-0.16	9.72E-10	T	C	0.49	-0.036	1.10E-08
ENSG00000011523	rs2723065	0.66	1	2	65279414	A	G	0.33	-0.18	2.14E-19	A	G	0.62	0.051	4.50E-15
ENSG00000163083	rs11688682	0.79	1	2	121347612	G	C	0.25	-0.52	2.68E-06	C	G	0.27	-0.058	1.40E-14
ENSG00000160801	rs11926707	0.77	1	3	46925539	T	C	0.36	-0.26	1.34E-11	T	C	0.37	-0.038	1.50E-08
ENSG00000163933	rs2581787	0.70	1	3	53127677	G	T	0.44	-0.48	1.06E-36	T	G	0.56	0.036	3.00E-08
ENSG00000233559	rs61061846	0.83	1	7	130458674	G	A	0.33	-0.26	3.27E-09	A	G	0.31	0.057	3.60E-16
ENSG00000029534	rs516946	0.57	1	8	41519248	T	C	0.19	0.60	3.28E-21	T	C	0.23	-0.08	4.70E-26
ENSG00000147874	rs62563594	0.62	1	9	19065862	T	C	0.34	-0.15	5.16E-21	T	C	0.6	-0.041	4.50E-10
ENSG00000107679	rs2280141	0.68	1	10	124193181	T	G	0.34	-0.09	2.17E-09	T	G	0.52	0.047	2.00E-13
ENSG00000115221	rs764729	0.59	1	2	161126732	A	C	0.18	-1.13	1.25E-43	A	C	0.26	-0.047	1.10E-10
ENSG00000185619	rs73221128	0.53	2	4	728111	C	T	0.04	-1.02	1.08E-42	T	C	0.045	0.11	4.80E-12
ENSG00000002726	rs62492368	0.88	1	7	150537635	G	A	0.38	0.73	5.53E-26	A	G	0.31	0.044	1.50E-10
ENSG00000149084	rs1061810	0.66	1	11	43877934	C	A	0.30	-0.74	2.48E-86	A	C	0.29	0.05	8.50E-13
ENSG00000064655	rs55966194	0.73	1	20	45599090	C	G	0.19	-0.51	2.21E-07	C	G	0.72	0.047	7.00E-11
cg12387154	rs2857605	0.67	3	6	31524851	C	T	0.24	0.39	1.18E-06	T	C	0.78	0.061	4.80E-15
cg13799504	rs28624681	0.77	1	9	139237902	T	C	0.28	0.86	3.64E-35	T	C	0.24	-0.077	1.20E-20
cg01678292	rs516946	0.63	1	8	41519248	T	C	0.19	-0.63	1.68E-42	T	C	0.23	-0.08	4.70E-26
cg12439423	rs516946	0.63	1	8	41519248	T	C	0.19	-0.52	1.93E-39	T	C	0.23	-0.08	4.70E-26
cg17274126	rs516946	0.63	1	8	41519248	T	C	0.19	-0.46	4.04E-25	T	C	0.23	-0.08	4.70E-26
cg11479568	rs516946	0.63	1	8	41519248	T	C	0.19	-0.50	7.77E-27	T	C	0.23	-0.08	4.70E-26
cg23241016	rs516946	0.61	1	8	41519248	T	C	0.19	-0.58	1.41E-34	T	C	0.23	-0.08	4.70E-26
cg12003463	rs516946	0.60	1	8	41519248	T	C	0.19	-0.26	3.03E-10	T	C	0.23	-0.08	4.70E-26
cg00328284	rs516946	0.52	1	8	41519248	T	C	0.19	-0.24	1.26E-08	T	C	0.23	-0.08	4.70E-26
cg27650870	rs516946	0.62	2	8	41519248	T	C	0.19	-0.34	1.14E-07	T	C	0.23	-0.08	4.70E-26
cg11023808	rs1206760	0.69	1	20	45582472	G	A	0.45	-0.46	1.35E-10	A	G	0.57	-0.041	1.60E-10
cg01981545	rs6937795	0.56	1	6	137291281	A	C	0.48	0.27	4.11E-09	A	C	0.53	0.048	6.50E-14
cg27467552	rs36138276	0.62	1	22	50422348	G	A	0.41	0.87	4.26E-44	A	G	0.5	-0.042	2.30E-10
cg21364723	rs36155743	0.54	1	22	50417483	C	T	0.41	0.48	1.63E-12	T	C	0.49	-0.042	4.20E-10
cg01346448	rs36138276	0.61	1	22	50422348	G	A	0.41	0.58	2.01E-24	A	G	0.5	-0.042	2.30E-10
cg21805149	rs1742546	0.61	1	14	91883499	G	A	0.28	-0.52	2.25E-13	A	G	0.57	0.037	8.40E-09
cg16477774	rs12789028	0.66	1	11	65326154	G	A	0.11	-0.60	5.38E-08	A	G	0.19	0.062	2.10E-14
cg06979164	rs2290203	0.51	1	15	91512067	G	A	0.25	-0.55	2.41E-14	A	G	0.2	0.061	8.70E-15
cg03372407	rs11858506	0.57	1	15	41831773	C	T	0.39	0.60	7.68E-20	T	C	0.64	-0.047	1.70E-12
cg23850205	rs8107967	0.71	1	19	7972615	A	G	0.49	0.59	1.28E-20	A	G	0.44	0.044	1.00E-11

Table 2.8.2 continued from previous page

gene	variant	RCP	qtl_rank	QTL_chr	QTL_pos	QTL_REF	QTL_ALT	MAF	QTL_beta	QTL_p	GWAS_EA	GWAS_NEA	GWAS_EAF	GWAS_beta	GWAS_p
cg08925307	rs2303700	0.79	2	19	7976529	T	C	0.33	0.29	3.48E-07	T	C	0.33	0.048	6.40E-12
cg04459751	rs10011174	0.64	1	4	153495515	G	A	0.38	0.94	2.20E-53	A	G	0.32	-0.051	9.20E-14
cg07814932	rs6668119	0.62	1	1	120439109	G	C	0.15	0.56	4.43E-11	C	G	0.11	0.081	4.80E-15
cg10446745	rs6668119	0.61	1	1	120439109	G	C	0.15	-0.61	1.53E-11	C	G	0.11	0.081	4.80E-15
cg01379234	rs2074314	0.53	1	11	17411821	C	T	0.46	-0.47	3.73E-26	T	C	0.63	-0.068	3.10E-25
cg26029265	rs10408163	0.59	1	19	47597102	T	C	0.36	0.25	2.10E-09	T	C	0.29	-0.045	2.40E-10
cg14189808	rs2241388	0.63	1	19	47572987	T	C	0.36	0.32	2.97E-08	T	C	0.29	-0.046	1.80E-10
cg27300045	rs62136856	0.62	1	19	47573527	A	G	0.36	0.39	2.67E-14	A	G	0.29	-0.045	1.10E-10
cg27408049	rs1572993	0.75	1	1	205045087	G	A	0.45	0.59	5.04E-32	A	G	0.43	0.037	1.00E-08
cg24610763	rs68137036	0.53	1	6	43820215	A	G	0.34	-0.32	6.75E-11	A	G	0.72	-0.049	3.90E-12
cg00859314	rs10426693	0.64	1	19	46147527	T	C	0.40	-0.49	9.22E-11	T	C	0.56	0.058	1.00E-18
cg15591645	rs10426693	0.62	1	19	46147527	T	C	0.40	-0.48	1.76E-16	T	C	0.56	0.058	1.00E-18
cg01691686	rs10426693	0.61	1	19	46147527	T	C	0.40	-0.33	8.54E-10	T	C	0.56	0.058	1.00E-18
cg14517983	rs10426693	0.61	1	19	46147527	T	C	0.40	-0.37	3.89E-07	T	C	0.56	0.058	1.00E-18
cg15737090	rs28433019	0.61	1	19	46153651	C	T	0.40	-0.60	8.10E-23	T	C	0.44	-0.058	7.60E-19
cg05289678	rs6977081	0.62	1	7	150542515	G	T	0.40	0.54	4.26E-15	T	G	0.33	0.039	9.50E-09
cg26475742	rs6977081	0.59	1	7	150542515	G	T	0.40	0.55	7.02E-16	T	G	0.33	0.039	9.50E-09
cg00668852	rs7794796	0.53	1	7	150540196	C	T	0.39	-0.31	2.98E-10	T	C	0.33	0.04	2.50E-09
cg06221570	rs7794796	0.67	1	7	150540196	C	T	0.39	0.45	6.92E-12	T	C	0.33	0.04	2.50E-09
cg22512663	rs6743795	0.64	1	2	161122134	A	G	0.20	-0.33	2.57E-06	A	G	0.28	-0.047	4.80E-11
cg02329928	rs3757974	0.67	1	8	145545546	A	G	0.29	0.70	2.22E-34	A	G	0.62	-0.051	1.30E-13
cg23097878	rs11038678	0.64	1	11	45858522	C	A	0.42	0.37	1.62E-13	A	C	0.52	-0.035	4.90E-08
cg13554586	rs11187129	0.90	1	10	94429907	T	C	0.45	0.76	1.22E-35	T	C	0.57	0.11	4.60E-60
cg23009123	rs11187129	0.85	1	10	94429907	T	C	0.45	0.75	1.64E-34	T	C	0.57	0.11	4.60E-60
cg17928459	rs11187129	0.85	1	10	94429907	T	C	0.45	0.55	6.90E-21	T	C	0.57	0.11	4.60E-60
cg09001573	rs11187129	0.72	1	10	94429907	T	C	0.45	0.36	1.83E-14	T	C	0.57	0.11	4.60E-60
cg24787755	rs11187129	0.90	1	10	94429907	T	C	0.45	0.35	4.19E-14	T	C	0.57	0.11	4.60E-60
cg00653997	rs12325400	0.55	1	16	30023786	C	G	0.37	-0.44	9.98E-12	C	G	0.6	-0.042	1.70E-10
cg14689537	rs12219514	0.90	1	10	94466439	A	G	0.45	0.58	1.05E-20	A	G	0.56	0.11	4.60E-61
cg16049864	rs896854	0.69	1	8	95960511	T	C	0.46	-0.91	3.75E-51	T	C	0.5	0.05	4.00E-15
cg12838385	rs896854	0.66	1	8	95960511	T	C	0.46	-0.76	7.09E-33	T	C	0.5	0.05	4.00E-15
cg22283921	rs896854	0.59	1	8	95960511	T	C	0.46	0.94	2.50E-45	T	C	0.5	0.05	4.00E-15
cg05986745	rs73167315	0.71	1	13	31026830	A	T	0.32	0.59	7.61E-23	A	T	0.72	0.04	3.70E-08
cg04251828	rs111852127	0.59	1	16	75249170	T	A	0.08	0.62	2.86E-16	A	T	0.077	-0.13	2.30E-26
cg12751941	rs11642612	0.50	1	16	30030195	A	C	0.37	-0.35	2.74E-08	A	C	0.6	-0.042	1.80E-10
cg18599843	rs12778642	0.65	1	10	94464307	G	T	0.45	0.34	5.15E-11	T	G	0.44	-0.11	1.30E-61
cg25506282	rs12778642	0.62	1	10	94464307	G	T	0.45	0.66	2.65E-28	T	G	0.44	-0.11	1.30E-61
cg15432903	rs5215	0.54	1	11	17408630	C	T	0.49	0.29	3.26E-14	T	C	0.63	-0.07	2.00E-26
cg23343264	rs116861488	0.79	2	12	118401849	G	A	0.18	-0.53	6.96E-12	A	G	0.14	0.052	1.40E-08
cg01386425	rs74855230	0.72	2	12	118401220	C	T	0.18	-0.67	1.62E-09	T	C	0.14	0.052	1.40E-08
cg22386930	rs34845373	0.55	1	2	25635771	A	G	0.23	-0.42	4.88E-08	A	G	0.73	0.04	4.30E-08
cg00379635	rs12987881	0.73	1	2	25638408	C	T	0.23	1.25	1.16E-71	T	C	0.27	-0.039	4.50E-08

Table 2.8.2 continued from previous page

gene	variant	RCP	qtl_rank	QTL_chr	QTL_pos	QTL_REF	QTL_ALT	MAF	QTL_beta	QTL_p	GWAS_EA	GWAS_NEA	GWAS_EAF	GWAS_beta	GWAS_p
cg03275851	rs13092876	0.62	1	3	185495320	G	A	0.29	-0.61	6.29E-19	A	G	0.32	0.11	5.10E-58
cg19595750	rs3764049	0.66	1	12	133087707	C	G	0.33	-0.54	8.83E-25	C	G	0.68	-0.048	1.90E-11
cg03030267	rs4810145	0.91	1	20	57396495	T	C	0.38	-0.62	3.33E-24	T	C	0.48	-0.045	4.40E-12
cg13280882	rs555754	0.58	1	6	160769423	G	A	0.50	0.76	1.03E-31	A	G	0.48	-0.037	4.00E-09
cg15344192	rs11688682	1.00	1	2	121347612	G	C	0.25	0.52	1.51E-15	C	G	0.27	-0.058	1.40E-14
cg14231073	rs11688682	1.00	1	2	121347612	G	C	0.25	1.07	9.37E-69	C	G	0.27	-0.058	1.40E-14
cg22826063	rs917195	1.00	1	7	30728452	C	T	0.20	-0.82	3.28E-37	T	C	0.23	-0.051	5.60E-11
cg00907998	rs11257655	0.99	1	10	12307894	C	T	0.28	-0.38	2.61E-11	T	C	0.22	0.09	3.70E-32
cg19574696	rs4709746	0.96	1	6	164133001	C	T	0.08	0.55	8.35E-13	T	C	0.13	-0.056	5.00E-09
cg19435526	rs4804833	0.93	1	19	7970635	A	G	0.39	-0.24	1.11E-10	A	G	0.39	0.047	1.10E-12
cg04167856	rs56348580	0.92	1	12	121432117	G	C	0.28	-0.95	1.98E-40	C	G	0.31	-0.062	3.80E-19
cg15728109	rs11257655	0.92	1	10	12307894	C	T	0.28	0.51	7.79E-15	T	C	0.22	0.09	3.70E-32
cg16531156	rs11257655	0.90	1	10	12307894	C	T	0.28	-0.41	9.96E-09	T	C	0.22	0.09	3.70E-32
cg02010481	rs1513272	0.90	1	7	28200097	C	T	0.43	-0.50	1.10E-18	T	C	0.49	-0.092	5.30E-48
cg02430063	rs10408179	0.85	1	19	46157004	T	C	0.42	-0.49	1.02E-11	T	C	0.56	0.059	2.90E-19
cg12840540	rs35318451	0.81	1	12	133068484	G	A	0.34	0.40	9.98E-12	A	G	0.33	0.049	2.60E-12
cg13564020	rs17109256	0.81	1	14	79939993	G	A	0.24	-1.07	4.91E-49	A	G	0.22	0.057	1.90E-13
cg11953941	rs62492368	0.76	1	7	150537635	G	A	0.38	0.42	4.52E-09	A	G	0.31	0.044	1.50E-10
cg16465430	rs17122782	0.72	1	14	23289189	T	C	0.15	0.68	4.11E-24	T	C	0.77	-0.043	2.30E-08
cg00048149	rs1398676	0.72	1	12	26459420	C	T	0.31	-0.94	3.98E-67	T	C	0.25	0.047	1.30E-10
cg21330313	rs878521	0.71	1	7	44255643	G	A	0.19	0.50	2.87E-12	A	G	0.25	0.057	1.60E-14
cg10655499	rs7640294	0.69	1	3	53130913	C	A	0.44	-0.25	2.44E-13	A	C	0.56	0.036	3.00E-08
cg22024966	rs2581787	0.69	1	3	53127677	G	T	0.44	1.01	3.14E-90	T	G	0.56	0.036	3.00E-08
cg22190077	rs7970193	0.69	1	12	27963301	G	A	0.18	0.68	2.40E-19	A	G	0.19	-0.074	4.90E-20
cg18383835	rs2280141	0.68	1	10	124193181	T	G	0.34	-0.33	2.64E-15	T	G	0.52	0.047	2.00E-13
cg06542216	rs10097617	0.66	1	8	95961626	T	C	0.45	-0.78	5.64E-40	T	C	0.48	0.051	1.10E-15
cg23172400	rs2879813	0.65	1	8	95960947	A	G	0.45	-0.60	1.05E-32	A	G	0.48	0.051	1.70E-15
cg13393036	rs2879813	0.65	1	8	95960947	A	G	0.45	-0.75	3.35E-50	A	G	0.48	0.051	1.70E-15
cg21721566	rs56348580	0.64	2	12	121432117	G	C	0.28	-0.39	2.07E-10	C	G	0.31	-0.062	3.80E-19
cg20039814	rs10097617	0.62	1	8	95961626	T	C	0.45	-0.79	4.57E-53	T	C	0.48	0.051	1.10E-15
cg09323728	rs10097617	0.62	1	8	95961626	T	C	0.45	-0.68	1.19E-43	T	C	0.48	0.051	1.10E-15
cg18059933	rs2879813	0.62	1	8	95960947	A	G	0.45	-0.78	2.66E-58	A	G	0.48	0.051	1.70E-15
cg23890800	rs35318451	0.62	1	12	133068484	G	A	0.34	-0.36	1.04E-09	A	G	0.33	0.049	2.60E-12
cg15262952	rs35105141	0.61	1	16	30057148	C	T	0.36	0.38	4.57E-07	T	C	0.4	0.042	1.50E-10
cg05614952	rs35318451	0.60	1	12	133068484	G	A	0.34	0.33	1.51E-07	A	G	0.33	0.049	2.60E-12
cg06015834	rs35105141	0.58	1	16	30057148	C	T	0.36	-0.69	5.47E-22	T	C	0.4	0.042	1.50E-10
cg27655716	rs36098511	0.57	1	12	133080449	A	T	0.34	0.43	2.61E-15	A	T	0.67	-0.049	2.60E-12
cg26312217	rs917195	0.57	2	7	30728452	C	T	0.20	-0.40	1.40E-07	T	C	0.23	-0.051	5.60E-11
cg15482002	rs9828772	0.55	1	3	129333182	C	G	0.07	0.54	1.30E-07	C	G	0.9	0.059	4.20E-08
cg03575602	rs11257655	1.00	1	10	12307894	C	T	0.28	0.36	2.26E-15	T	C	0.22	0.09	3.70E-32
cg10894156	rs11257655	1.00	1	10	12307894	C	T	0.28	-0.63	8.44E-18	T	C	0.22	0.09	3.70E-32
cg25354617	rs362307	0.99	1	4	3241845	C	T	0.06	-0.80	6.86E-15	T	C	0.077	0.074	1.10E-09

Table 2.8.2 continued from previous page

gene	variant	RCP	qtl_rank	QTL_chr	QTL_pos	QTL_REF	QTL_ALT	MAF	QTL_beta	QTL_p	GWAS_EA	GWAS_NEA	GWAS_EAF	GWAS_beta	GWAS_p
cg07161603	rs13262861	0.98	2	8	41508577	C	A	0.12	-0.60	2.76E-07	A	C	0.17	-0.094	1.80E-27
cg01033600	rs4804833	0.98	2	19	7970635	A	G	0.39	-0.40	1.77E-08	A	G	0.39	0.047	1.10E-12
cg20670582	rs4709746	0.95	1	6	164133001	C	T	0.08	1.16	1.47E-33	T	C	0.13	-0.056	5.00E-09
cg14353998	rs56348580	0.91	2	12	121432117	G	C	0.28	0.46	1.17E-13	C	G	0.31	-0.062	3.80E-19
cg24317972	rs11257655	0.88	2	10	12307894	C	T	0.28	0.33	5.17E-09	T	C	0.22	0.09	3.70E-32
cg04198914	rs10908278	0.83	1	17	36099952	T	A	0.37	-0.77	1.09E-31	A	T	0.52	-0.074	3.10E-30
cg07688604	rs35318451	0.79	1	12	133068484	G	A	0.34	-0.38	1.80E-13	A	G	0.33	0.049	2.60E-12
cg15043029	rs62492368	0.77	2	7	150537635	G	A	0.38	0.35	1.43E-06	A	G	0.31	0.044	1.50E-10
cg14350257	rs28429551	0.65	1	9	139243334	T	A	0.29	-0.57	1.10E-20	A	T	0.75	0.076	4.80E-21
cg05423304	rs11496066	0.64	2	7	102486254	T	C	0.20	0.31	7.56E-07	T	C	0.82	0.047	1.20E-08
cg02414922	rs28641468	0.62	1	9	139239585	T	C	0.29	-0.46	1.91E-12	T	C	0.25	-0.076	4.20E-21
cg25694349	rs35318451	0.60	1	12	133068484	G	A	0.34	-0.40	1.52E-14	A	G	0.33	0.049	2.60E-12
cg24933060	rs4729854	0.74	1	7	102383663	T	A	0.38	-0.67	1.52E-14	A	T	0.48	0.037	3.30E-08
cg01678292	rs516946	0.62	1	8	41519248	T	C	0.19	-1.23	5.01E-80	T	C	0.23	-0.08	4.70E-26
cg12439423	rs516946	0.62	1	8	41519248	T	C	0.19	-1.05	1.43E-56	T	C	0.23	-0.08	4.70E-26
cg23241016	rs516946	0.62	1	8	41519248	T	C	0.19	-1.15	1.41E-65	T	C	0.23	-0.08	4.70E-26
cg11479568	rs516946	0.61	1	8	41519248	T	C	0.19	-1.11	2.14E-63	T	C	0.23	-0.08	4.70E-26
cg17274126	rs516946	0.61	1	8	41519248	T	C	0.19	-0.95	3.51E-55	T	C	0.23	-0.08	4.70E-26
cg17420165	rs1206760	0.68	1	20	45582472	G	A	0.45	0.59	3.93E-20	A	G	0.57	-0.041	1.60E-10
cg02010152	rs9275614	0.60	1	6	32684257	A	G	0.16	0.60	3.21E-09	A	G	0.88	-0.074	3.70E-14
cg01493678	rs9275611	0.63	1	6	32683763	G	A	0.16	0.69	1.47E-12	A	G	0.12	0.076	5.50E-15
cg15672654	rs137862	0.51	1	22	50446550	C	A	0.42	-0.55	3.05E-22	A	C	0.51	0.04	1.40E-09
cg00090674	rs137864	0.52	1	22	50446988	C	T	0.42	-0.43	7.21E-20	T	C	0.51	0.04	8.80E-10
cg08241514	rs5771069	0.52	1	22	50435480	A	G	0.40	-0.41	1.82E-11	A	G	0.49	-0.041	6.00E-10
cg27491509	rs137845	0.57	1	22	50439430	A	G	0.41	-0.66	8.14E-29	A	G	0.49	-0.04	7.30E-10
cg01548456	rs36155743	0.56	1	22	50417483	C	T	0.41	-0.42	1.88E-18	T	C	0.49	-0.042	4.20E-10
cg21364723	rs36138276	0.55	1	22	50422348	G	A	0.41	0.73	1.61E-31	A	G	0.5	-0.042	2.30E-10
cg01464473	rs9873519	0.82	2	3	124921457	C	T	0.37	-0.36	1.11E-06	T	C	0.54	0.039	1.40E-09
cg05256313	rs9870956	0.64	1	3	124925881	C	T	0.33	0.65	3.02E-21	T	C	0.43	-0.037	1.00E-08
cg10768996	rs4951182	0.56	1	1	205236233	A	C	0.35	-0.22	2.14E-09	A	C	0.55	-0.036	3.00E-08
cg21805149	rs11621425	0.50	1	14	91906186	C	G	0.28	-0.44	2.31E-14	C	G	0.42	-0.036	2.50E-08
cg07029024	rs12789028	0.72	1	11	65326154	G	A	0.11	-0.48	1.15E-09	A	G	0.19	0.062	2.10E-14
cg03372407	rs11858506	0.56	1	15	41831773	C	T	0.39	0.75	4.44E-26	T	C	0.64	-0.047	1.70E-12
cg06979164	rs8032722	0.73	1	15	91522070	T	C	0.27	-0.75	1.06E-26	T	C	0.76	-0.057	1.40E-14
cg10446745	rs6668119	0.56	1	1	120439109	G	C	0.15	-0.59	6.30E-11	C	G	0.11	0.081	4.80E-15
cg17799449	rs11639412	0.51	1	15	64112634	T	A	0.45	0.78	3.71E-36	A	T	0.57	-0.039	2.10E-09
cg17413945	rs68137036	0.67	1	6	43820215	A	G	0.34	-0.31	1.67E-09	A	G	0.72	-0.049	3.90E-12
cg01184401	rs68137036	0.55	1	6	43820215	A	G	0.34	-0.33	1.63E-08	A	G	0.72	-0.049	3.90E-12
cg02430063	rs35816837	0.65	1	19	46148903	C	A	0.40	-0.73	2.15E-23	A	C	0.44	-0.058	8.70E-19
cg15737090	rs10426693	0.60	1	19	46147527	T	C	0.40	-0.54	2.43E-14	T	C	0.56	0.058	1.00E-18
cg15591645	rs10426693	0.59	1	19	46147527	T	C	0.40	-0.39	9.25E-08	T	C	0.56	0.058	1.00E-18
cg23850205	rs2115107	0.97	1	19	7968168	G	A	0.37	-0.79	5.16E-27	A	G	0.39	0.047	1.90E-12

Table 2.8.2 continued from previous page

gene	variant	RCP	qtl_rank	QTL_chr	QTL_pos	QTL_REF	QTL_ALT	MAF	QTL_beta	QTL_p	GWAS_EA	GWAS_NEA	GWAS_EAF	GWAS_beta	GWAS_p
cg17602887	rs73167313	0.60	1	13	31019580	C	T	0.32	-0.37	1.89E-07	T	C	0.28	-0.04	3.70E-08
cg23097878	rs7945565	0.57	1	11	45878992	A	G	0.42	0.45	6.07E-28	A	G	0.49	0.035	4.90E-08
cg20670582	rs17630640	0.83	1	6	164107529	A	G	0.08	0.36	4.02E-08	A	G	0.87	0.053	3.40E-08
cg11211307	rs3757971	0.63	1	8	145545949	T	C	0.29	0.36	9.44E-12	T	C	0.62	-0.051	1.00E-13
cg27423010	rs3757969	0.53	1	8	145551199	C	G	0.29	0.68	5.43E-32	C	G	0.63	-0.052	1.30E-13
cg00706536	rs12938909	0.71	1	17	40787764	C	G	0.18	0.52	3.52E-09	C	G	0.72	-0.048	1.20E-11
cg17928459	rs11187129	0.63	1	10	94429907	T	C	0.45	0.39	1.22E-12	T	C	0.57	0.11	4.60E-60
cg09001573	rs11187129	0.57	1	10	94429907	T	C	0.45	0.32	1.08E-11	T	C	0.57	0.11	4.60E-60
cg13554586	rs11187129	0.51	1	10	94429907	T	C	0.45	0.29	1.06E-07	T	C	0.57	0.11	4.60E-60
cg06015834	rs12325400	0.76	1	16	30023786	C	G	0.37	-0.67	7.25E-20	C	G	0.6	-0.042	1.70E-10
cg01283141	rs12444108	0.66	1	16	30027694	G	A	0.37	-0.36	1.41E-11	A	G	0.4	0.042	1.70E-10
cg24787755	rs12219514	0.66	1	10	94466439	A	G	0.45	0.35	2.71E-14	A	G	0.56	0.11	4.60E-61
cg12838385	rs896854	0.61	1	8	95960511	T	C	0.46	-0.95	5.43E-58	T	C	0.5	0.05	4.00E-15
cg18059933	rs896854	0.59	1	8	95960511	T	C	0.46	-1.02	1.06E-81	T	C	0.5	0.05	4.00E-15
cg01714284	rs56376363	0.55	1	4	185726914	T	C	0.08	-0.52	1.13E-08	T	C	0.85	0.067	2.10E-13
cg01899937	rs56376363	0.53	1	4	185726914	T	C	0.08	0.68	2.31E-18	T	C	0.85	0.067	2.10E-13
cg18599843	rs12778642	0.63	1	10	94464307	G	T	0.45	0.31	4.18E-08	T	G	0.44	-0.11	1.30E-61
cg07298363	rs17211038	0.74	1	5	52118488	C	T	0.15	-0.39	6.14E-10	T	C	0.16	0.05	1.00E-08
cg13187651	rs12936169	0.61	2	17	40824823	G	A	0.18	0.47	5.27E-06	A	G	0.28	0.048	3.00E-11
cg03977449	rs35602018	0.71	1	2	25635264	G	A	0.23	-0.47	7.07E-20	A	G	0.27	-0.04	4.60E-08
cg00379635	rs35602018	0.71	1	2	25635264	G	A	0.23	0.89	1.79E-60	A	G	0.27	-0.04	4.60E-08
cg03030267	rs4810145	0.77	1	20	57396495	T	C	0.38	-0.40	1.62E-09	T	C	0.48	-0.045	4.40E-12
cg14689537	rs5015480	0.53	1	10	94465559	C	T	0.45	0.45	1.96E-18	T	C	0.41	-0.11	2.70E-62
cg07715834	rs11688682	1.00	1	2	121347612	G	C	0.25	0.54	1.95E-15	C	G	0.27	-0.058	1.40E-14
cg25756780	rs11688682	1.00	1	2	121347612	G	C	0.25	-0.66	1.11E-23	C	G	0.27	-0.058	1.40E-14
cg01653701	rs11688682	1.00	1	2	121347612	G	C	0.25	0.50	5.15E-12	C	G	0.27	-0.058	1.40E-14
cg00421221	rs11688682	1.00	1	2	121347612	G	C	0.25	0.38	2.43E-09	C	G	0.27	-0.058	1.40E-14
cg14231073	rs11688682	1.00	1	2	121347612	G	C	0.25	1.01	4.96E-79	C	G	0.27	-0.058	1.40E-14
cg26035105	rs11688682	1.00	1	2	121347612	G	C	0.25	0.52	1.27E-11	C	G	0.27	-0.058	1.40E-14
cg15344192	rs11688682	1.00	1	2	121347612	G	C	0.25	0.44	4.83E-20	C	G	0.27	-0.058	1.40E-14
cg26406689	rs11688682	1.00	1	2	121347612	G	C	0.25	0.33	2.65E-11	C	G	0.27	-0.058	1.40E-14
cg27114644	rs11688682	0.99	1	2	121347612	G	C	0.25	-0.45	8.12E-12	C	G	0.27	-0.058	1.40E-14
cg20067049	rs11688682	0.99	1	2	121347612	G	C	0.25	-0.60	4.13E-17	C	G	0.27	-0.058	1.40E-14
cg02793858	rs11688682	0.99	1	2	121347612	G	C	0.25	0.40	3.23E-12	C	G	0.27	-0.058	1.40E-14
cg04167856	rs56348580	0.93	1	12	121432117	G	C	0.28	-0.98	2.02E-44	C	G	0.31	-0.062	3.80E-19
cg24950598	rs11688682	0.90	2	2	121347612	G	C	0.25	0.49	9.65E-14	C	G	0.27	-0.058	1.40E-14
cg06627114	rs72926932	0.89	1	18	53050646	A	C	0.08	-0.42	8.87E-08	A	C	0.92	-0.083	3.60E-13
cg24610763	rs9472138	0.80	1	6	43811762	C	T	0.34	-0.39	2.16E-12	T	C	0.29	0.051	6.90E-13
cg24796450	rs9873618	0.78	1	3	170733076	G	A	0.26	-0.41	1.30E-19	A	G	0.29	-0.066	8.50E-21
cg18625956	rs9472138	0.76	1	6	43811762	C	T	0.34	-0.49	3.55E-16	T	C	0.29	0.051	6.90E-13
cg09255149	rs867489	0.76	1	20	48833957	C	T	0.48	-0.86	1.47E-44	T	C	0.46	-0.043	2.70E-11
cg25658765	rs3811978	0.73	1	5	52100489	A	G	0.15	-0.44	2.18E-06	A	G	0.83	-0.053	4.20E-10

Table 2.8.2 continued from previous page

gene	variant	RCP	qtl_rank	QTL_chr	QTL_pos	QTL_REF	QTL_ALT	MAF	QTL_beta	QTL_p	GWAS_EA	GWAS_NEA	GWAS_EAF	GWAS_beta	GWAS_p
cg18828459	rs9369425	0.71	1	6	43810974	G	A	0.34	-0.38	2.56E-09	A	G	0.71	-0.051	6.90E-13
cg17746527	rs867489	0.71	1	20	48833957	C	T	0.48	-0.22	4.14E-07	T	C	0.46	-0.043	2.70E-11
cg17975832	rs7970193	0.71	1	12	27963301	G	A	0.18	0.79	4.38E-27	A	G	0.19	-0.074	4.90E-20
cg22190077	rs10771372	0.70	1	12	27962260	C	T	0.18	1.02	2.25E-47	T	C	0.19	-0.074	5.50E-20
cg22024966	rs7640294	0.67	1	3	53130913	C	A	0.44	0.82	6.56E-35	A	C	0.56	0.036	3.00E-08
cg13393036	rs2879813	0.62	1	8	95960947	A	G	0.45	-0.92	3.62E-68	A	G	0.48	0.051	1.70E-15
cg20039814	rs2879813	0.62	1	8	95960947	A	G	0.45	-0.97	4.14E-75	A	G	0.48	0.051	1.70E-15
cg06542216	rs2879813	0.62	1	8	95960947	A	G	0.45	-0.94	4.92E-63	A	G	0.48	0.051	1.70E-15
cg03175975	rs11688682	0.62	1	2	121347612	G	C	0.25	0.36	6.52E-07	C	G	0.27	-0.058	1.40E-14
cg09323728	rs2879813	0.62	1	8	95960947	A	G	0.45	-0.82	2.72E-63	A	G	0.48	0.051	1.70E-15
cg03713592	rs77464186	0.61	1	11	72460398	A	C	0.22	0.81	6.18E-26	A	C	0.84	0.11	2.30E-33
cg23172400	rs2879813	0.61	1	8	95960947	A	G	0.45	-0.78	7.52E-54	A	G	0.48	0.051	1.70E-15
cg16049864	rs2879813	0.60	1	8	95960947	A	G	0.45	-1.05	6.43E-79	A	G	0.48	0.051	1.70E-15
cg18006637	rs1903002	0.57	1	4	89740894	C	G	0.44	-0.34	1.03E-09	C	G	0.5	-0.036	3.00E-08
cg18063878	rs11039307	0.57	1	11	47611152	C	T	0.43	-0.22	1.82E-09	T	C	0.41	0.037	9.20E-09
cg26933147	rs1493694	0.56	1	1	120526982	C	T	0.14	-0.60	9.25E-08	T	C	0.11	0.084	2.10E-16
cg21752471	rs329122	0.52	1	5	133864599	G	A	0.40	-0.23	5.56E-07	A	G	0.43	0.037	9.20E-09
cg15438478	rs2028150	0.51	1	2	65655012	C	G	0.31	-0.88	6.10E-58	C	G	0.6	0.052	3.10E-15
cg14624731	rs11688682	1.00	2	2	121347612	G	C	0.25	0.40	8.64E-17	C	G	0.27	-0.058	1.40E-14
cg22826063	rs917195	1.00	1	7	30728452	C	T	0.20	-1.31	1.71E-69	T	C	0.23	-0.051	5.60E-11
cg10894156	rs11257655	1.00	1	10	12307894	C	T	0.28	-0.61	3.71E-30	T	C	0.22	0.09	3.70E-32
cg07161603	rs13262861	0.98	2	8	41508577	C	A	0.12	-0.59	1.06E-08	A	C	0.17	-0.094	1.80E-27
cg20677018	rs13262861	0.98	2	8	41508577	C	A	0.12	-0.60	1.28E-07	A	C	0.17	-0.094	1.80E-27
cg19435526	rs4804833	0.97	1	19	7970635	A	G	0.39	-0.67	6.15E-29	A	G	0.39	0.047	1.10E-12
cg17254229	rs11688682	0.97	2	2	121347612	G	C	0.25	0.21	4.47E-07	C	G	0.27	-0.058	1.40E-14
cg08957513	rs72926932	0.88	2	18	53050646	A	C	0.08	-0.51	2.68E-06	A	C	0.92	-0.083	3.60E-13
cg19846096	rs35318451	0.80	2	12	133068484	G	A	0.34	-0.49	6.08E-12	A	G	0.33	0.049	2.60E-12
cg14213590	rs28429551	0.67	2	9	139243334	T	A	0.29	0.40	6.95E-11	A	T	0.75	0.076	4.80E-21
cg13799504	rs3935875	0.67	1	9	139238824	A	G	0.29	0.53	1.42E-23	A	G	0.25	-0.076	4.20E-21
cg18071195	rs2581787	0.66	1	3	53127677	G	T	0.44	1.13	1.04E-101	T	G	0.56	0.036	3.00E-08
cg20214067	rs35318451	0.64	1	12	133068484	G	A	0.34	-0.38	2.45E-10	A	G	0.33	0.049	2.60E-12
cg19209729	rs28562046	0.52	2	9	139241595	C	G	0.29	-0.30	1.29E-07	C	G	0.25	-0.076	3.40E-21
cg05532283	rs28562046	0.52	2	9	139241595	C	G	0.29	-0.33	5.51E-08	C	G	0.25	-0.076	3.40E-21

Table 2.8.3. Sample sizes for physiological trait associations with each type of molecular traits

Trait	Muscle mRNA	Muscle DNAme	Muscle miRNA	Adipose mRNA	Adipose DNAme	Adipose miRNA
Matsuda index 0, 30, 60, 120min	151	140	154	132	138	126
60min plasma insulin from OGTT	154	143	157	135	141	128
Matsuda index 0, 30, 120min	167	155	170	147	155	140
30min plasma insulin from OGTT	170	158	173	150	158	143
120min plasma insulin from OGTT	172	161	176	153	161	144
Fasting plasma insulin	172	161	176	153	161	144
T2D	185	170	179	176	169	162
Hemoglobin A1c	197	185	196	178	184	169
60min glucose from OGTT	275	255	264	251	247	235
Disposition index	291	271	280	272	266	254
Insulin AUC 0 to 30min	291	271	280	272	266	254
Insulinogenic index	291	271	280	272	266	254
InsulinAUC glucoseAUC ratio 0 to 30min	291	271	280	272	266	254
30min serum insulin	291	271	280	272	266	254
C peptidogenic index	292	272	281	273	267	255
Fasting serum C peptide 30min	292	272	281	273	267	255
30min glucose from OGTT	295	275	284	275	270	258
Glucose AUC 0 to 30min	295	275	284	275	270	258
glycated Hemoglobin A1c	300	281	289	279	275	262
Diastolic blood pressure	300	281	289	278	274	262
2h glucose from OGTT	300	281	289	279	275	262
Systolic blood pressure	300	281	289	278	274	262
ApoB A1 ratio	301	282	290	280	276	263
BMI	301	282	290	280	276	263
Fasting serum C peptide	301	282	290	280	276	263
HDL cholesterol	301	282	290	280	276	263
LDL cholesterol	301	282	290	280	276	263
Total cholesterol	301	282	290	280	276	263
Creatinine	301	282	290	280	276	263
Triglycerides	301	282	290	280	276	263
Fasting glucose from OGTT	301	282	290	280	276	263
2h glucose closest to biopsy date	301	282	290	280	276	263
Fasting glucose	301	282	290	280	276	263
Height	301	282	290	280	276	263
Hip circumference	300	281	289	280	275	263
HOMA	301	282	290	280	276	263
Relative fat mass	300	281	289	280	275	263
Alanine aminotransferase(ALT)	301	282	290	280	276	263
Glutamyltransferase(GGT)	301	282	290	280	276	263
C-Reactive protein	301	282	290	280	276	263
Fasting serum insulin	301	282	290	280	276	263
Apolipoprotein A1(A1)	301	282	290	280	276	263
Apolipoprotein B(ApoB)	301	282	290	280	276	263
Serum uric acid	301	282	290	280	276	263
Waist	300	281	289	280	275	263
Weight	301	282	290	280	276	263
Waist hip ratio	300	281	289	280	275	263
BMI adjusted WHR	300	281	289	280	275	263

Table 2.8.4. Number of molecular traits significantly associated with physiological traits in muscle and/or adipose.

Molecular trait type	Physiological trait	Number of significant associations in muscle	Number of significant associations in adipose	Number of associations significant in both tissues	Different direction	Same direction	Fisher test p-value
mRNA	BMI	288	311	19	2	17	1.72E-10
mRNA	Relative fat mass	211	337	17	1	16	2.51E-10
mRNA	Waist	157	360	14	0	14	6.08E-09
mRNA	Fasting serum insulin	350	226	13	2	11	2.44E-06
mRNA	HOMA	296	213	11	2	9	8.38E-06
mRNA	Fasting serum C peptide	246	152	7	2	5	0.00024535
mRNA	Hip circumference	189	109	4	0	4	0.0048009
mRNA	Insulin AUC 0 to 30min	225	71	4	0	4	0.00190729
mRNA	Disposition index	197	50	2	0	2	0.0415669
mRNA	Fasting serum C peptide 30min	117	10	2	0	2	0.00064419
mRNA	InsulinAUC glucoseAUC ratio 0 to 30min	142	17	2	0	2	0.0027991
mRNA	30min serum insulin	166	39	2	0	2	0.01914321
DNAme	BMI	44	811	1	0	1	0.05014214
DNAme	Relative fat mass	43	881	1	0	1	0.05315112
DNAme	Waist	40	1314	1	0	1	0.07299754
miRNA	Alanine aminotransferase(ALT)	3	1	1	0	1	0.00308008
miRNA	Glutamyltransferase(GGT)	2	1	1	0	1	0.00205339
miRNA	Fasting serum insulin	6	8	1	0	1	0.04840223
mRNA	Triglycerides	14	44	1	0	1	0.02002371
mRNA	Matsuda index 0, 30, 120min	141	27	1	0	1	0.1177083
mRNA	Weight	81	131	1	0	1	0.29482222
DNAme	Disposition index	9	2	0	0	0	1
DNAme	Fasting serum C peptide	20	363	0	0	0	1
DNAme	HDL cholesterol	5	2	0	0	0	1
DNAme	Creatinine	0	1	0	0	0	1
DNAme	Triglycerides	4	0	0	0	0	1
DNAme	Fasting glucose from OGTT	0	6	0	0	0	1
DNAme	2h glucose from OGTT	0	19	0	0	0	1
DNAme	60min glucose from OGTT	1	0	0	0	0	1

Table 2.8.4 continued from previous page

Molecular trait type	Physiological trait	Number of significant associations in muscle	Number of significant associations in adipose	Number of associations significant in both tissues	Different direction	Same direction	Fisher test p-value
DNAme	2h glucose closest to biopsy date	0	4	0	0	0	1
DNAme	Glucose AUC 0 to 30min	0	1	0	0	0	1
DNAme	Fasting glucose	0	3	0	0	0	1
DNAme	Hip circumference	12	292	0	0	0	1
DNAme	HOMA	36	421	0	0	0	1
DNAme	Insulin AUC 0 to 30min	7	0	0	0	0	1
DNAme	InsulinAUC glucoseAUC ratio 0 to 30min	4	0	0	0	0	1
DNAme	Matsuda index 0, 30, 120min	5	10	0	0	0	1
DNAme	Matsuda index 0, 30, 60, 120min	2	3	0	0	0	1
DNAme	Fasting plasma insulin	6	13	0	0	0	1
DNAme	120min plasma insulin from OGTT	1	1	0	0	0	1
DNAme	Glutamyltransferase(GGT)	0	1	0	0	0	1
DNAme	C-Reactive protein	5	0	0	0	0	1
DNAme	Fasting serum insulin	65	344	0	0	0	1
DNAme	30min serum insulin	4	0	0	0	0	1
DNAme	T2D	0	0	0	0	0	1
DNAme	Weight	15	424	0	0	0	1
DNAme	Waist hip ratio	3	112	0	0	0	1
miRNA	glycated Hemoglobin A1c	1	0	0	0	0	1
miRNA	BMI	6	4	0	0	0	1
miRNA	Disposition index	5	1	0	0	0	1
miRNA	Fasting serum C peptide	4	18	0	0	0	1
miRNA	Fasting serum C peptide 30min	2	0	0	0	0	1
miRNA	HDL cholesterol	1	2	0	0	0	1
miRNA	Triglycerides	0	1	0	0	0	1
miRNA	Fasting glucose from OGTT	0	3	0	0	0	1
miRNA	60min glucose from OGTT	3	0	0	0	0	1
miRNA	Glucose AUC 0 to 30min	0	6	0	0	0	1
miRNA	Hip circumference	1	0	0	0	0	1
miRNA	HOMA	6	9	0	0	0	1

Table 2.8.4 continued from previous page

Molecular trait type	Physiological trait	Number of significant associations in muscle	Number of significant associations in adipose	Number of associations significant in both tissues	Different direction	Same direction	Fisher test p-value
miRNA	Insulin AUC 0 to 30min	6	1	0	0	0	1
miRNA	Insulinogenic index	1	0	0	0	0	1
miRNA	InsulinAUC glucoseAUC ratio 0 to 30min	6	0	0	0	0	1
miRNA	Matsuda index 0, 30, 120min	5	0	0	0	0	1
miRNA	Fasting plasma insulin	5	0	0	0	0	1
miRNA	120min plasma insulin from OGTT	1	0	0	0	0	1
miRNA	Relative fat mass	8	0	0	0	0	1
miRNA	30min serum insulin	5	0	0	0	0	1
miRNA	Serum uric acid	1	0	0	0	0	1
miRNA	T2D	1	0	0	0	0	1
miRNA	Waist	5	3	0	0	0	1
miRNA	Weight	1	1	0	0	0	1
miRNA	Waist hip ratio	0	1	0	0	0	1
mRNA	ApoB A1 ratio	2	2	0	0	0	1
mRNA	C peptidogenic index	1	0	0	0	0	1
mRNA	Total cholesterol	1	0	0	0	0	1
mRNA	HDL cholesterol	16	43	0	0	0	1
mRNA	Creatinine	1	0	0	0	0	1
mRNA	Fasting glucose from OGTT	1	0	0	0	0	1
mRNA	2h glucose from OGTT	1	0	0	0	0	1
mRNA	60min glucose from OGTT	1	0	0	0	0	1
mRNA	2h glucose closest to biopsy date	1	0	0	0	0	1
mRNA	Glucose AUC 0 to 30min	1	0	0	0	0	1
mRNA	Height	2	0	0	0	0	1
mRNA	Insulinogenic index	16	0	0	0	0	1
mRNA	Matsuda index 0, 30, 60, 120min	53	4	0	0	0	1
mRNA	Fasting plasma insulin	156	7	0	0	0	1
mRNA	120min plasma insulin from OGTT	63	10	0	0	0	1
mRNA	30min plasma insulin from OGTT	26	0	0	0	0	1
mRNA	60min plasma insulin from OGTT	7	0	0	0	0	1

Table 2.8.4 continued from previous page

Molecular trait type	Physiological trait	Number of significant associations in muscle	Number of significant associations in adipose	Number of associations significant in both tissues	Different direction	Same direction	Fisher test p-value
mRNA	Alanine aminotransferase(ALT)	4	1	0	0	0	1
mRNA	Glutamyltransferase(GGT)	3	27	0	0	0	1
mRNA	C-Reactive protein	2	0	0	0	0	1
mRNA	Apolipoprotein A1(A1)	3	5	0	0	0	1
mRNA	Serum uric acid	5	7	0	0	0	1
mRNA	Systolic blood pressure	1	0	0	0	0	1
mRNA	T2D	1	0	0	0	0	1
mRNA	Waist hip ratio	9	40	0	0	0	1

Table 2.8.5. *INHBB* associations with physiological traits

Physiological trait	Base model		Model adjusted for fasting serum insulin		Additionally adjusing for waist	
	Coefficient	P-values	Coefficient	P-values	Coefficient	P-values
Fasting serum insulin	0.32	1.72E-07	NA	NA	0.27	2.71E-05
HOMA	0.31	3.77E-07	-0.02	9.54E-01	0.27	5.03E-05
Fasting serum C peptide	0.27	1.46E-05	-0.01	9.24E-01	0.21	1.28E-03
Fasting plasma insulin	0.33	1.23E-04	0.28	5.65E-01	0.27	2.37E-03
Insulin AUC 0 to 30min	0.23	1.61E-04	0.04	6.15E-01	0.19	2.04E-03
Matsuda index 0, 30, 120min	-0.35	2.23E-04	-0.11	6.58E-01	-0.28	2.78E-03
Waist	0.26	3.10E-04	0.13	6.92E-02	NA	NA
Relative fat mass	0.37	3.28E-04	0.19	6.97E-02	0.18	4.48E-01
Matsuda index 0, 30, 60, 120min	-0.36	4.87E-04	0.00	9.96E-01	-0.28	6.32E-03
30min serum insulin	0.20	8.08E-04	0.04	5.98E-01	0.16	6.20E-03
HDL cholesterol	-0.20	1.01E-03	-0.10	9.88E-02	-0.17	5.05E-03
Disposition index	0.19	1.04E-03	0.01	8.85E-01	0.15	1.30E-02
Triglycerides	0.20	1.55E-03	0.08	2.40E-01	0.16	9.44E-03
BMI	0.21	1.82E-03	0.07	3.42E-01	0.02	8.62E-01
InsulinAUC glucoseAUC ratio 0 to 30min	0.18	1.96E-03	0.02	8.03E-01	0.15	1.22E-02
120min plasma insulin from OGTT	0.28	2.38E-03	0.08	5.69E-01	0.23	1.54E-02
Fasting serum C peptide 30min	0.18	3.04E-03	0.01	8.40E-01	0.14	1.85E-02
Waist hip ratio	0.22	6.47E-03	0.13	1.06E-01	0.05	6.03E-01
Hip circumference	0.16	9.34E-03	0.05	4.09E-01	-0.07	5.29E-01
C-Reactive protein	0.14	2.05E-02	0.11	5.07E-02	0.07	2.41E-01
Weight	0.17	2.97E-02	0.02	8.36E-01	-0.17	1.87E-01
Systolic blood pcmb$_{r}essure$	0.12	4.77E-02	0.11	4.14E-02	0.08	1.79E-01
2h glucose from OGTT	0.12	5.75E-02	0.07	2.59E-01	0.09	1.32E-01
Apolipoprotein A1(A1)	-0.11	5.79E-02	-0.04	5.09E-01	-0.09	1.03E-01
ApoB A1 ratio	0.11	6.34E-02	0.06	3.18E-01	0.10	8.72E-02
2h glucose closest to biopsy date	0.11	7.39E-02	0.06	2.91E-01	0.09	1.48E-01
Alanine aminotransferase(ALT)	0.11	7.65E-02	0.02	7.33E-01	0.10	8.14E-02
Diastolic blood pressure	0.10	8.96E-02	0.07	2.01E-01	0.07	2.13E-01
Insulinogenic index	0.09	1.04E-01	0.01	9.00E-01	0.07	2.44E-01
60min plasma insulin from OGTT	0.15	1.08E-01	-0.15	2.18E-01	0.10	2.91E-01
30min plasma insulin from OGTT	0.14	1.27E-01	-0.06	5.44E-01	0.11	1.91E-01
Glucose AUC 0 to 30min	0.10	1.40E-01	0.03	6.57E-01	0.08	2.11E-01
BMI adjusted WHR	0.11	1.61E-01	0.10	1.48E-01	0.05	5.29E-01
30min glucose from OGTT	0.09	1.82E-01	0.03	6.04E-01	0.07	2.36E-01
Creatinine	-0.09	1.85E-01	-0.07	2.91E-01	-0.06	3.64E-01
Serum uric acid	0.08	2.25E-01	0.02	7.03E-01	0.01	8.31E-01
Fasting glucose from OGTT	0.07	2.43E-01	-0.02	7.13E-01	0.05	4.29E-01
T2D	0.19	2.43E-01	0.05	7.42E-01	0.14	4.05E-01
Apolipoprotein B(ApoB)	0.06	3.06E-01	0.04	4.43E-01	0.06	3.16E-01
60min glucose from OGTT	0.07	3.14E-01	0.02	7.32E-01	0.05	4.98E-01
Fasting glucose	0.06	3.36E-01	-0.03	6.32E-01	0.03	5.85E-01

Table 2.8.5 continued from previous page

Physiological trait	Base model		Model adjusted for fasting serum insulin		Additionally adjusing for waist	
	Coefficient	P-values	Coefficient	P-values	Coefficient	P-values
Hemoglobin A1c	0.06	4.25E-01	0.01	9.36E-01	0.05	5.35E-01
Glutamyltransferase(GGT)	0.04	4.66E-01	0.01	9.26E-01	0.01	8.93E-01
Total cholesterol	-0.04	4.70E-01	-0.02	7.64E-01	-0.03	5.77E-01
C peptidogenic index	0.04	4.72E-01	0.00	9.77E-01	0.03	6.42E-01
Height	-0.06	5.15E-01	-0.03	7.44E-01	-0.06	4.42E-01
LDL cholesterol	-0.02	6.56E-01	0.00	9.53E-01	-0.01	8.08E-01
Glycated Hemoglobin A1c	-0.01	9.15E-01	-0.06	2.92E-01	-0.03	6.06E-01

Chapter 3

A Subcutaneous Adipose Tissue eQTL Meta-analysis from 2256 European Individuals

3.1 Introduction

Based on visually distinguishable tissue color, human adipose tissues can be classified into white, brown, beige, and pink adipose[126]. As the predominant form (80%)[127] of adipose tissue in adults, white adipose tissue mainly exists under the skin as subcutaneous adipose tissue or inside the abdominal cavity as visceral adipose tissue. White adipose tissue encompasses adipocytes and the stromal-vascular fraction that includes heterogeneous cell populations such as preadipocytes, endothelial cells, pericytes, monocytes, macrophages, fibroblasts, and red blood cells. Besides providing physical protection and preventing heat loss[128], white adipose tissue also ensures sufficient energy status by storing free fatty acids (FFAs) in the fed state and releasing FFA during the fasting state[129].

For many years, white adipose tissue was considered only a reservoir for energy storage. In the last two decades, the paracrine and endocrine capacities of white adipose tissue have received increasing attention, and it has been found to release various protein, lipid, and nucleic acid factors[40]. Leptin and adiponectin are two types of hormones primarily produced and secreted from white adipose tissue, regulating energy metabolism and immunity[40]. Leptin inhibits food intake through central nervous system[130] and increases

insulin sensitivity by decreasing adiposity and lipotoxicity[131]. Adiponectin levels are lower in people with obesity, insulin resistance and type 2 diabetes (T2D)[132], [133]. Prolonged energy excess triggers white adipose tissue expansion by increasing adipocyte size (hypertrophy) and number (hyperplasia), resulting in increased body mass and obesity[134], [135]. When the storage capacity of white adipose tissue is approaching its limit, further energy overloads to ectopic tissues (e.g., skeletal muscle, liver, pancreas)[134], [135]. As the excessive energy continues to store in white adipose tissue and ectopic tissues, white adipose tissue undergoes deleterious effects such as inflammation, hypoxia, altered hormone secretion, and becomes dysfunctional[136], [137]. White adipose tissue dysfunction and ectopic lipid accumulation in turn lead to systemic insulin resistance (IR), promoting obesity-associated cardiometabolic disorder[136], [137]. While visceral adipose tissue has historically been considered a major culprit in the development of obesity and its related metabolic consequences, the role of subcutaneous adipose tissue has gained increasing attention[138]. It has been suggested that subcutaneous adipose tissue distribution in the upper body is detrimental to the development of type 2 diabetes (T2D) and cardiovascular diseases (CVD) while the distribution in the lower body may be protective[40], [42].

Cardiometabolic diseases (CMDs) such as T2D, obesity, and CVD are partially caused by genetic factors[139]. Studying the genetic regulation of gene expression in subcutaneous adipose can generate insights into the molecular mechanisms underlying the genetic predisposition to obesity and cardiometabolic disorders; such studies may also expand our knowledge on how genetic factors affect gene functions by influencing their expression levels in non-disease conditions.

The typical approach to discover genetic regulators of gene expression levels is to identify expression quantitative trait loci (eQTL) by testing for associations between gene expression and genetic variations. Usually eQTLs are detected by considering one variant at a time (single-variant model), using a linear regression model to test for the association between the expression level of a gene and a genetic variant. This approach often reveals a set of variants, each of which is statistically associated with the expression level of a gene.

However, the associations of this set of variants may be driven by the same causal variant or by multiple causal variants in the locus. One commonly used approach to identify the multiple variants with independent effects on a given gene is to use conditional analysis. The genetic variant showing the strongest statistical evidence for association in a locus without conditioning on any other genetic variants is considered the primary eQTL variant. The genetic variants that show statistical significance after adjusting for the previously identified QTL variants are considered secondary eQTLs.

Based on a thorough literature search, I found seven single-study eQTL analyses have identified eQTLs in human subcutaneous adipose tissues with a sample size ranging from 63 to 855[28], [29], [31], [44]–[47], [65]. Nearly all participants in these studies are of European ancestry, except that 15% of GTEx samples are from individuals of non-European ancestry (12.9% are of African-American ancestry and 1.3% are of East-Asian ancestry). All of the studies have detected *cis*-eQTLs (eQTLs that are within a certain distance to a gene). Using the marginal eQTL association model (test one variant at a time), the seven studies have identified that 4.5% to 68.3% of the tested genes have *cis*-eQTLs with different significance thresholds. Two of the seven studies (Raulerson et al.[29] and The GTEx Consortium[96]) have identified conditionally independent eQTLs. Raulerson et al.[29] performed one round of conditional analysis (conditioning on the first eQTL variant) and detected up to two eQTLs per gene, while GTEx[96] performed multiple rounds of conditional analysis. Raulerson et al. and the GTEx Consortium identified genes with eQTLs colocalizated with CMD GWAS loci. Notably, Raulerson et al. identified 21 genes whose secondary eQTLs, not primary eQTLs colocalizated with CMD GWAS loci.

Combining eQTL associations through meta-analysis to increase power is an effective strategy to identify genetic variants with modest or small effects on gene expression levels. Compared to eQTLs identified in individual studies, an expanded eQTL catalog identified in the meta-analysis provides a more comprehensive genetic architecture for gene expression levels and enables the discovery of additional genes involved in CMD mechanisms. Therefore, we combined the eQTL associations from TwinsUK, METSIM, GTEx, and FUSION and performed the largest RNA-seq based *cis*-eQTL meta-analysis in subcutaneous

adipose tissue (n=2256) to date. We identified genetic variants with independent effects on the gene expression level for each gene through conditional eQTL meta-analysis and used this eQTL catalog to generate insights into the potential target genes underlying the genetic associations for cardiometabolic diseases.

3.2 Methods

3.2.1 TwinsUK sample collection, genotype and RNA-seq data

TwinsUK had RNA-seq based gene expression data available for 804 subcutaneous adipose tissue biopsies, taken from Caucasian female twins recruited through the TwinsUK Adult twin registry[140], [141]. The punch biopsies of subcutaneous adipose tissue were taken from a sun-protected area in the sub-umbilical region. Array genotyping was also performed on these samples, on a combination of the HumanHap300, HumanHap610Q, 1M-Duo and 1.2M Duo Illumina arrays. Genotyping and imputation procedures using the Haplotype Reference Consortium (HRC) reference panel[68] were described[142], [143]. Poly(A)-selected RNA samples were prepared using the Illumina TruSeq directional mRNA-seq library protocol and sequenced on a HiSeq 2000 machine with 49-bp paired end reads. RNA-seq reads were aligned to the hg19 reference genome using STAR version 2.4.0.1[144] using the GENCODE v19 annotations[145]. Detailed RNA sample collection, RNA isolation, mRNA-seq, and quality control procedures were fully described[142], [146].

3.2.2 METSIM sample collection, genotype and RNA-seq data

The METSIM study consists of 10,197 males of Finnish ancestry from Kuopio, Finland[28]. Genotypes were measured using the Illumina OmniExpress BeadChip. Detailed genotyping and imputation procedures using the HRC panel[68] were described[28]. Subcutaneous adipose tissue biopsies taken from an area near the umbilicus were available for two subsets of the METSIM participants.

The first subset (METSIM-1) had a total of 550 needle biopsy samples. Poly(A)-selected RNA samples were prepared using the Illumina TruSeq RNA Sample Preparation Kit v2 and sequenced on a HiSeq 2000 sequencing machine. RNA-seq generated 50bp paired-

end reads with an average sequencing depth of 45 million reads per sample. RNA-seq reads were aligned to the hg19 reference genome using STAR version 2.4.2a[144] using the GENCODE v19 annotations[145]. Detailed procedures of sample collection, RNA isolation, mRNA-seq, and quality control were described in Raulerson et al.[29]. Raulerson et al. estimated the proportions of subcutaneous adipose tissue, whole blood, skeletal muscle tissue, and lymphocytes for each of the 550 samples[29]. They found the *cis*-eQTL results using samples with > 50% subcutaneous adipose (n=434) tissue had the most significant variant-gene pairs, the most significant *cis*-eQTL variants, and the strongest associations for known eQTLs for *KLF14*, *ADIPOQ*, and *CDH13*, compared to those using the full sample set (n=550) or using samples with adipose tissue proportion > 75% (n=387). They decided to use the 434 samples with > 50% subcutaneous adipose tissue (*ADIPOQ* expression levels $\geq$ 150 CPM adjusted for TMM) for cis-eQTL detection. In the current study, we further excluded eight of the 434 samples that overlapped with FUSION samples and included the remaining 426 samples from METSIM-1 in analysis.

The second subset (METSIM-2) had 420 surgical biopsy samples. Poly(A)-selected RNA samples were prepared using the Sciclone G3 NGS and NGSx Workstation. RNA samples were sequenced using the Illumina NovaSeq 6000 S4XP at the High Throughput Sequencing Core at the University of North Carolina at Chapel Hill. RNA-seq generated 150bp paired-end reads with an average sequencing depth of 42.6 million reads per sample. RNA-seq reads were aligned to the hg19 reference genome using STAR version 2.7.2a[144] using the GENCODE v19 annotations[145]. Detailed procedures of sample collection, RNA isolation, mRNA-seq, and quality control were described (Brotman et al., manuscript in preparation).

3.2.3 **GTEx v8 release sample collection, genotype and RNA-seq data**

GTEx v8 release collected 663 subcutaneous adipose tissue biopsies from the lower legs of post-mortem donors by surgical incision and performed RNA-seq on these samples. Of the 663 samples, 581 samples had genotype data from whole-genome sequencing (WGS). WGS-based genotyping was performed using Illumina HiSeq 2000 to a median depth of 32$\times$, as previously described[96]. Poly(A)-selected RNA samples were prepared

using the Illumina TruSeqTM unstranded RNA-seq protocol and sequenced using HiSeq 2000 or HiSeq 2500 machines. RNA-seq generated 76bp paired-end reads with a median coverage of about 83 million reads. Detailed procedures of RNA sequencing and quality control have been reported[96], [97]. As the gene expression and genotype data in the GTEx v8 release were in NCBI build GRCh38, while the other studies used NCBI build GRCh37, we processed the GTEx v8 data to match the other studies. For genotype data, we lifted over the GTEx v8 VCF from GRCh38 to GRCh37 using a reference file with the GRCh38 variants and corresponding GRCh37 variant positions. We replaced the GRCh38 variant information with the GRCh37 variant position in the VCF. We removed variants that were not able to be lifted over. We matched genes in the gene expression files between GTEx v8 and other studies by ENSEMBL gene IDs.

3.2.4 **FUSION sample collection, genotype and RNA-seq data**

FUSION collected 296 subcutaneous adipose tissue biopsies from 331 Finnish participants. The biopsies were taken from an area 5 to 10 cm lateral of the umbilicus by a surgical scalpel. Genotypes was measured using `HumanOmni2.5-4v1_H` or InfiniumOmni2-5Exome-8v1-3 BeadChip arrays. Detailed genotyping and imputation procedures using the HRC panel[68] have been described previously[26], [27] and in chapter 2. Poly(A)-selected RNA samples were prepared using the Illumina TruSeq directional mRNA-seq library protocol and sequenced on HiSeq sequencing machines. RNA-seq generated 100bp paired-end reads with a depth of > 80 million reads per sample. Detailed procedures have been described previously[26], [27] and in chapter 2. RNA-seq data were aligned to the human reference genome GRCh37 using STAR v2.5.3a[144] using the GENCODE v19 annotations[145]. Array genotypes were imputed to the HRC reference panel. The biopsy and experimental characteristics were summarized in Table3.3.1.

3.2.5 **Quality control filtering of genes and samples**

We developed a harmonized protocol from gene expression level quantification to within-study *cis*-eQTL mapping to minimize bias introduced by different analysis procedures across studies. For each study, we used the QTLtools quan function[79] to quantify the

gene expression levels. We retained genes with five or more counts per million (CPMs) in $\geq$ 25% of individuals in each study. To normalize for library size, we adjusted read counts for each gene for trimmed mean of M values(TMM)[147]. We inverse-normalized the TMM-normalised gene CPMs and used them in downstream analyses. To exclude samples with high blood contamination, we filtered out samples that likely had a low percentage of adipocytes, the characteristic cell type of adipose[148]. The previous publication based on the METSIM-1 samples[29] found that the *cis*-eQTL results using samples with adipose tissue proportions > 50% (n=434) had the most significant variant-gene pairs, the most significant *cis*-eQTL variants, and the strongest associations for known eQTLs for *KLF14*, *ADIPOQ*, and *CDH13*, compared to those using the full sample set (n=550) or using samples with adipose tissue proportion > 75% (n=387). The 434 samples with adipose tissue proportion > 50% had *ADIPOQ* (an adipocyte-specific gene[149]) expression levels > 150 CPM adjusted for TMM. Therefore, for the present study we removed samples with $\leq$ 150 CPM adjusted for TMM for *ADIPOQ* expression levels from each group. We filtered genetic variants for imputation $R^2 \geq 0.5$ and minor allele frequency (MAF) $\geq$ 0.01 in each study.

3.2.6 **PEER factor analysis**

To account for unknown biological and technical factors that may contribute to the gene expression levels, we performed factor analysis of gene expression levels using PEER v1.0[77] and included the estimated PEER factors as covariates in QTL mapping. To facilitate future examinations of BMI effects on eQTL detection, we adjusted for BMI from inverse normalized gene expression levels, and then inverse normalized the BMI-adjusted residuals. We used the inverse normalized BMI-adjusted residual gene expression levels to generate PEER factors. To detect *cis*-eQTLs, we tested for associations between the expression level of a gene and variants in the *cis* region (1Mb) for the gene using QTLtools[79]. we used a linear regression model with an additive genetic effect, adjusting for BMI and a specified number of PEER factors. To select the number of PEER factors that optimized *cis*-eQTL discovery, we generated PEER factors from zero to 100 with an increment of ten PEER factors, and compared proportion of genes with $\geq$ 1 eQTL across

models with 0, 10, 20, to 100 PEER factors. We define genes with $\geq$ 1 eQTL as eGenes. We used as covariates the largest number of PEER factors that resulted in $\geq$ 1% increase in the number of eGenes than the previous number of PEER factors.

3.2.7 *cis*-eQTL analysis

Finally, we performed *cis*-eQTL detection within each study using the number of PEER factors that optimized *cis*-eQTL discovery. We tested for *cis*-eQTLs among the genetic variants within 1Mb from the transcription start site (TSS) of a gene using the APEX store function[150], assuming an additive model of inheritance. For each variant-gene pair, we used a linear regression model to test for the association between the inverse-normalized gene expression levels and the variant dosages, adjusting for PEER factors, with and without additional adjustment for BMI. The eQTL results presented in the thesis are from the model with the adjustment for BMI.

For the use of conditional eQTL detection within each study or in a meta-analysis, we generated score statistic vectors of variant-gene associations and variance-covariance matrices between variants in a 2Mb-region around the TSS of a gene using the APEX store function[150].

3.2.8 Comparison of the conditional eQTL meta-analysis results obtained using the APEX conditional eQTL meta-analysis function and GCTA-COJO to those obtained using individual-level data for ten genes

To select software for genome-wide conditional eQTL meta-analysis, I compared the APEX conditional eQTL meta-analysis function[150], GCTA-COJO[125] and an individual-level data approach. I performed a comparison for ten genes on chromosome 22 in the meta-analysis of FUSION and GTEx v7, for which I had access to the individual-level data. The ten genes were randomly selected from genes with $\geq$ 2 eQTLs from a preliminary conditional-eQTL meta-analysis of TwinsUK, METSIM-1, and FUSION studies. For each analysis, I used a p-value $\leq 2.5 \times 10^{-6}$ as the threshold for inclusion for each variant in the model (a Bonferroni correction for testing 20,000 genes).

3.2.8.1 Analysis using the conditional eQTL meta-analysis function of APEX

I performed the single-variant eQTL mapping within each study using the single-variant eQTL mapping function of APEX[150] to generate the score statistics of marginal associations and variance-covariance matrices. Then I applied the conditional eQTL meta-analysis function of APEX to these marginal associations and variance-covariance matrices to identify conditional independent eQTLs for each gene.

3.2.8.2 Analysis using GCTA-COJO

I performed single-variant eQTL meta-analysis of FUSION and GTEx v7 data using the single-variant eQTL mapping meta-analysis function of APEX. Then I provided GCTA-COJO with the summary statistics (effect size, standard error, and p-value) of the single-variant eQTL meta-analysis results and HRC imputed genotypes of 10K randomly-selected unrelated UK Biobank (UKB) samples as a reference.

3.2.8.3 Analysis using individual-level data approach (gold standard)

As shown in Figure3.2.1, for each gene 1) I performed single-variant eQTL mapping within each study using FastQTL[151]; 2) I conducted a single-variant meta-analysis using the inverse-variance approach in METAL[152]; 3) I determined if the lead variant had a p-value less than the cut-off of 2.5×10^{-6}; 4) I included the dosage values of the lead variant as a covariate and performed a second round of single-variant eQTL mapping within each study to obtain the summary statistics of conditional associations; 5) I meta-analyzed the conditional associations using the inverse variance approach in METAL; 6) I repeated this process (steps 3 and 5) until no more variants passed the cut-off of p-value $\leq 2.5 \times 10^{-6}$.

I compared the number of eQTLs detected for each gene and the lead variant for each eQTL obtained using APEX or GCTA-COJO to those obtained using the individual-level data approach.

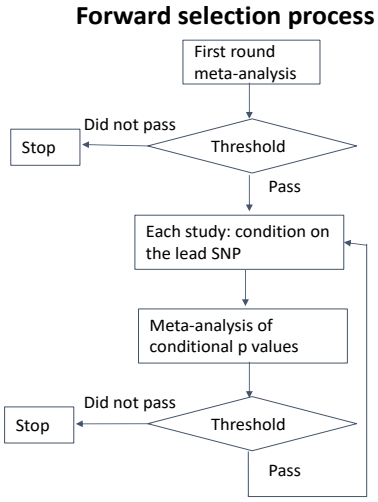


Figure 3.2.1. Workflow of conditional eQTL meta-analysis using individual-level data

3.2.9 Comparison of the conditional eQTL meta-analysis results obtained using the conditional eQTL meta-analysis function of APEX and GCTA-COJO on 538 chromosome 20 genes

To further compare conditional eQTL meta-analysis results obtained using the conditional eQTL meta-analysis function of APEX[150] and GCTA-COJO[125] for a larger set of genes and a meta-analysis of more studies, I separately applied them to detect conditional eQTLs for all 538 genes on chromosome 20 in a meta-analysis of three studies (TwinsUK, METSIM-1, and FUSION). I performed the analyses in the same way as described in section Comparison of the conditional eQTL meta-analysis results obtained using the APEX conditional eQTL meta-analysis function and GCTA-COJO to those obtained using individual-level data for ten genes. For the analysis with GCTA-COJO, we experimented with three reference panels, 10k samples from the larger METSIM cohort (almost the whole cohort), 10k and 20K randomly-selected unrelated samples from the UK Biobank(UKB). All of the three reference panels were imputed to the HRC reference panel. I used a p-value $\leq 2.5 \times 10^{-6}$ as the threshold for inclusion for each variant in the model (a Bonferroni correction for testing 20,000 genes).

3.2.10 **Genome-wide conditional eQTL meta-analysis**

I performed genome-wide conditional eQTL meta-analysis of TwinsUK, METSIM-1, METSIM-2, GTEx v8, and FUSION using the conditional eQTL meta-analysis function of APEX.

Part 1 Iterative conditional analysis to identify potentially independent eQTL signals:
I applied a forward selection process (Figure3.2.2) to each of the tested genes in parallel. For a given gene, in the first iteration, 1) I calculated the meta-analysis p-value for every variant in the *cis* region without conditioning on other variants. 2) I combined the meta-analysis unconditional p-values of all the variants tested in the first-round using a p-value combination method ACAT[153], [154] to calculate a gene-based p-value (denoted as ACAT p-value) to approximately account for the number of tested variants[150] and for the subsequent use in determining significant independent eQTLs in Part2. The null hypothesis of an ACAT p-value is that no remaining variant is associated with the gene expression level. 3) If the variant with the most significant meta-analysis p-value (lead variant) had a p-value ≤ 0.05, I included the lead variant in the conditioning eQTL variant list for the gene. The goal of using a lenient p-value threshold (≤ 0.05) is to make sure every gene that has a chance to have a significant ACAT-pvalue in the next round is included (see below for verification).

In the subsequent rounds, to avoid collinearity between the tested variant and the previously selected variant(s), I calculated the variance inflation factor (VIF) for each tested variant with the previously selected variant(s). As VIF > 10 indicates a high correlation, I tested variants with VIF $\leq$ 10. For each tested variant, I calculated the meta-analysis conditional p-value conditioning on the previously selected variant(s). If the meta-analysis conditional p-variant of the lead variant ≤ 0.05, I added the lead variant to the eQTL list for the given gene. I computed the gene-based p-value using the ACAT method based on the meta-analysis conditional p-values of all tested variants. I continued the rounds until no gene had a variant with meta-analysis conditional p-value ≤ 0.05.

Part 2 Determination of number of independent eQTLs signals for each gene using the ACAT-pvalues:

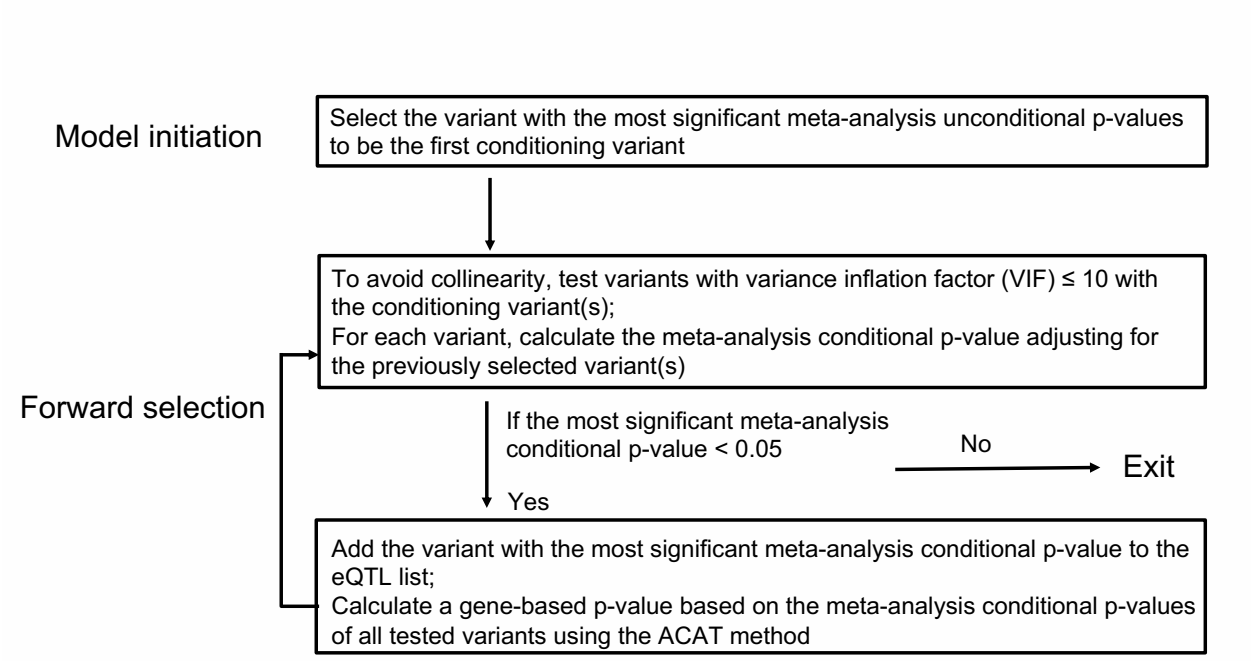


Figure 3.2.2. The forward-selection process in APEX used to identify conditional eQTLs for each gene in single-study analysis and meta-analysis.

After the forward selection process was completed for every gene, I determined how many eQTLs were significant for each gene (Figure3.2.3). To determine whether the primary eQTL of a gene was significant, I calculated the FDR with the Benjamini-Hochberg procedure based on the first round ACAT p-values (ACAT p-values of meta-analysis unconditional p-values for each gene). The primary eQTL of a given gene was considered significant if the first round ACAT p-value was $\leq$ 1% FDR. Next, for genes with significant primary eQTL signals, I calculated FDR using the second round ACTA p-values (ACAT p-values of meta-analysis p-values conditioning on the variants selected in the first iteration). The 2nd eQTL of a given gene was considered significant if the corresponding second round ACAT p-value $\leq$ 1% FDR. I repeated this process for each round until no gene was significant in the round. I verified that no gene would be selected to continue to the next stage by the threshold of ACAT p-value $\leq$ 1% FDR if the lead variant in the current round has meta-analysis p-value > 0.001.

After I identified eGenes and their independent eQTLs, I isolated each conditional eQTL association for the use of colocalization analysis by carrying out “all-but-one” conditional

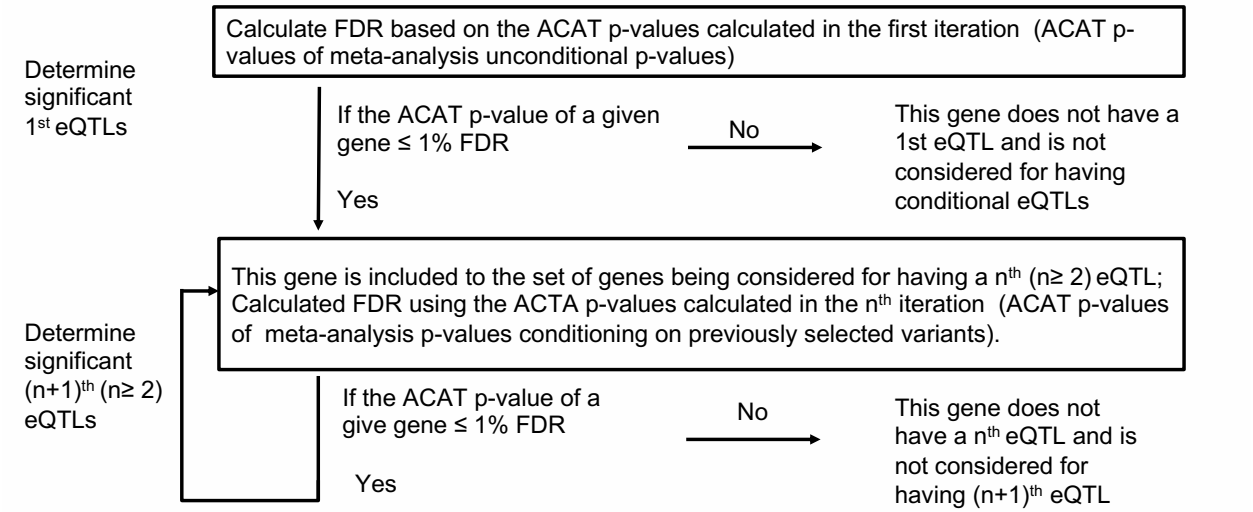


Figure 3.2.3. Post forward-selection procedure to determine significant conditional eQTLs using the ATAC-pvalues.

analyses using the APEX R interface. For each independent eQTL of a gene with $\geq$ 2 eQTLs, I calculated the meta-analysis effect sizes and p-values conditioning on the lead variants of all of the other independent eQTLs. In this way, we generated summary statistics for each independent eQTL, conditioning on all of the other independent eQTLs for a gene (denoted as "all-but-one" conditional analysis). For example, for a gene with three independent eQTLs, three "all-but-one" conditional analyses were done. The first analysis conditioned on the lead variants of the 2nd and 3rd eQTLs, the second analysis conditioned on the lead variants of the 1st and 3rd eQTLs, and the third analysis conditioned on the lead variants of the 1st and 2nd eQTLs.

3.2.11 **Colocalization analysis between genetic associations for cardiometabolic diseases and eQTLs for gene expression levels in subcutaneous adipose tissue**

I downloaded the GWAS summary statistics for seven cardiometabolic traits: T2D[33], BMI[155], waist-hip ratio (WHR)[155], BMI-adjusted waist-hip ratio (WHRadjBMI)[155], Coronary artery disease (CAD) [156], and fasting glucose and fasting insulin[104]. I used

coloc2[115] to quantify the probability that the genetic associations for GWAS traits and eQTL shared causal variants. For a GWAS signal-eQTL signal pair where the two signals are obtained from two different datasets of unrelated individuals, coloc2 assumes 1) the two signals are derived from the same ancestry, and 2) each signal has a single causal variant. Coloc2 first fine maps each signal and then integrates over the two posterior distributions to concurrently calculate the posterior probabilities of five hypotheses (H0, no association signal in either the GWAS or eQTL; H1, only the GWAS has an association signal; H2, only the eQTL has an association signal; H3, both datasets have an association signal, but they are not the same; H4, the GWAS and eQTL associations signals are colocalized).

I considered a GWAS signal and an eQTL signal to be colocalized if PP4 > 0.8. I performed colocalization analysis for genes present in at least one study and had $\geq$ 1 eQTL. For genes with 1 eQTL, I extracted summary statistics for variants located within 1 Mb flanking the lead GWAS variants from the downloaded marginal GWAS associations and from marginal eQTL associations to test for colocalization. For genes with $\geq$ 2 eQTLs, I extracted eQTL summary statistics from the "all-but-one" conditional analysis for each conditional eQTL. I applied the same colocalization analysis to each GWAS locus-conditional eQTL pair.

3.2.12 Colocalization between the separated WHRadjBMI GWAS locus near *ZNF664* and meta-analysis eQTL

For the WHRadjBMI GWAS locus near *ZNF664*, I explored whether using the conditional summary statistics for the GWAS associations will enable the identification of colocalization between additional eQTLs and the GWAS signals. Two distinct GWAS signals were found in the WHRadjBMI locus near *ZNF664*[157], represented by rs863750 and rs7133378 (R^2=0.016, D'=0.15 in European population), respectively. I separately computed the approximate conditional summary statistics for each signal, conditioning on the other using GCTA-COJO[125]. Using the summary statistics of the conditional GWAS associations and conditional eQTL associations, I tested for colocalization for each GWAS signal-conditional eQTL pair using coloc2[115], [158].

3.3 **Results**

3.3.1 **Sample characteristics**

I performed *cis*-eQTL meta-analysis in subcutaneous adipose tissue, using genotype and RNA-seq based gene expression level data from four studies: TwinsUK, METSIM, GTEx v8 release (hereafter referred to as GTEx), and FUSION. METSIM had the first batch of 426 samples collected using needle biopsy (METSIM-1), and the second batch of 420 samples collected using surgical biopsy (METSIM-2). As METSIM-1 and METSIM-2 samples had different tissue biopsy procedures and were sequenced in different sequencing centers, they were included as two separate groups of samples in the meta-analysis. The biopsy and experimental characteristics are shown in Table3.3.1. GTEx and FUSION had both males and female samples, while TwinsUK only had female samples and METSIM only had male samples. All of the samples in TwinsUK, METSIM and FUSION studies were of European ancestry. Of the 581 GTEx samples that had both genotype and gene expression data, 479 (82.4%) were of European ancestry, 71 (12.2%) were of African American ancestry, and 31 (5.3%) were of Asian ancestry. As most of our samples were of European ancestry, I performed one meta-analysis including only samples of European ancestry (TwinsUK, METSIM, FUSION, and GTEx European) and a separate meta-analysis including samples of mixed ancestry individuals (TwinsUK, METISIM, FUSION, and all GTEx). The sample demographic characteristics are summarized in Table3.3.2.

Study	Biopsy site	Technique	Genotyping	Imputation
TwinsUK	Sub-umbilical area	Punch biopsy	Array	HRC
METSIM-1	Near-umbilicus area	Needle biopsy	Array	HRC
METSIM-2		Surgical scalpel	Array	HRC
GTEx	Lower leg	Surgical incision	WGS	No imputation
FUSION	Lateral of the umbilicus	Surgical scalpel	Array	HRC

Table 3.3.1. Biopsy and experimental characteristics of participating studies. WGS: Whole-genome sequencing. HRC: the Haplotype Reference Consortium panel.

Study	Sample size	Ancestry	Sex	Age(years, 1st-3rd quantiles)
TwinsUK	722	European	Female	59[52-65]
METSIM-1	426	Finnish	Male	54 [51-59]
METSIM-2	420	Finnish	Male	
GTEx-Euro	407	European	Both	53.4[21-70]
GTEx-all	495	European,African, Asian	Both	
FUSION	280	Finnish	both	60[55-65]

Table 3.3.2. Demographic characteristics of participating studies

3.3.2 Comparison of the conditional eQTL meta-analysis results obtained using the conditional eQTL meta-analysis function of APEX and GCTA-COJO to those obtained using the individual-level data approach

We aimed to combine the conditional eQTL associations from individual studies via meta-analysis to improve power for eQTL discovery. Software that can be used for conditional eQTL meta-analysis include the conditional eQTL meta-analysis function of APEX[150] and GCTA-COJO[125]. APEX uses score statistic vectors of marginal eQTL associations and variance-covariance matrices between variants from individual studies to perform conditional eQTL meta-analysis. GCTA-COJO uses the summary statistics of combined marginal variant-GWAS trait associations and an external reference panel to perform approximate conditional analysis for phenotypic GWAS study. GCTA-COJO has the potential to be used for approximate conditional analysis of eQTL associations by treating the expression level of a gene as a quantitative GWAS trait.

To evaluate the conditional eQTL associations obtained from APEX and those obtained from GCTA-COJO, I separately compared their results with those obtained using a gold-standard approach. The gold-standard approach to perform conditional eQTL detection in a meta-analysis without access to individual-level data is to perform sequential conditional eQTL detection within each study using the lead variants selected from the previous meta-analysis rounds (denoted as individual-level data approach). This is a laborious process and has to be conducted up to several times for each gene to complete the sequential conditional analysis. Therefore, I detected conditional eQTLs with the individual-level data approach for ten genes on chromosome 22, using FUSION and GTEx v7 release data that I had access to. The ten genes were randomly selected genes with $\geq$ 2 eQTLs

from a preliminary conditional-eQTL meta-analysis of TwinsUK, METSIM-1, and FUSION studies. Using the individual-level data approach, I detected 1 eQTL for one gene, 3 eQTLs for 5 genes, 4 eQTLs for 2 genes, and 5 eQTLs for 2 genes in the meta-analysis of FUSION and GTEx v7.

For the same ten genes, I performed conditional eQTL meta-analysis using APEX and GCTA-COJO separately. Comparing the APEX results to those from the individual-level data approach, for nine of the ten genes, the two approaches agreed in both the number of eQTL signals and the lead eQTL variants (the same variant or in high LD (min R^2 = 0.93)) (Table3.3.3). For one (*PI4KAP2*) gene, APEX detected one more signal (rs138649538) than using the individual-level data, which was slightly more significant than the p-value threshold of 2.5×10^{-6}. The p-values from the two approaches were consistent overall, with very significant (small) p-values displayed a larger discrepancy. Of the 34 variant-gene pairs, 28 (82.4%) had p-values in the same order of magnitude, three differed in one order of magnitude, two differed in two orders of magnitude, and one differed in three orders of magnitude.

As 96% of our samples were of European ancestry, I used 10k randomly-selected unrelated samples from UKB as a reference panel for GCTA-COJO. Comparing the GCTA-COJO results to those from individual-level data approach, for the ten genes, the number of conditionally independent eQTLs detected by GCTA-COJO (6 to 21 eQTLs) were much larger than the individual-level data approach (1 to 5 eQTLs); although the lead variants of the primary eQTLs were the same, none of the secondary eQTL was the same or in high LD (R^2 threshold = 0.8) with those detected by the individual-level data approach.

Gene	Lead eQTL varaint			Conditional p-values		Ratio between p-values
	Individual-level data approach	APEX	LD R^2	Individual-level data approach	APEX	
TUBGCP6	rs112983849	rs112983849	Same	3.24E-09	3.18E-09	1.02
TBC1D22A	rs801640	rs801640	Same	1.38E-46	1.38E-46	1.00
	rs15646	rs15646	Same	7.01E-07	1.67E-07	4.19
	rs12389	rs2295441	0.93	3.49E-08	3.06E-08	1.14
CDC42EP1	rs9610795	rs9610795	Same	6.04E-17	1.19E-17	5.09
	rs7291467	rs7291467	Same	3.58E-07	1.71E-07	2.09
	rs4821677	rs7290515	1	2.03E-06	1.74E-06	1.16
MMP11	rs9624318	rs9624318	Same	2.27E-11	7.73E-11	0.29
	rs5751789	rs5751789	Same	2.19E-06	1.97E-06	1.11
	rs62239011	rs62239011	Same	2.42E-06	2.50E-06	0.97
AP000347.4	rs7289879	rs10222270	1	5.27E-40	2.76E-39	0.19
	rs5759963	rs11090280	1	1.26E-16	8.31E-16	0.15
	rs61479247	rs61479247	Same	8.52E-07	2.10E-06	0.41
C1QTNF6	rs739040	rs739040	Same	2.22E-13	3.16E-13	0.70
	rs62235065	rs62235065	Same	2.03E-06	2.13E-06	0.95
	rs10854698	rs10854698	Same	1.59E-06	2.49E-06	0.64
PRODH	rs367766	rs367766	Same	5.40E-36	3.51E-34	0.02
	rs8137125	rs8137125	Same	4.66E-15	3.31E-15	1.41
	rs759404	rs759404	Same	8.41E-08	1.19E-07	0.70
	rs111404325	rs11913840	1	4.60E-07	3.11E-07	1.48
FAM118A	rs104664	rs104664	Same	7.15E-136	2.24E-133	0.003
	rs738176	rs738176	Same	3.92E-33	2.60E-33	1.51
	rs58667	rs58667	Same	3.36E-06	2.76E-06	1.22
	rs2294202	rs2294202	Same	7.28E-06	1.31E-06	5.56
SERHL	rs5751306	rs5751306	Same	1.85E-26	8.42E-26	0.22
	rs137055	rs137055	Same	1.03E-09	4.04E-09	0.25
	rs5758768	rs5758768	Same	4.07E-08	6.77E-08	0.60
	rs143108908	rs143108908	Same	1.36E-06	2.36E-06	0.58
	rs8139383	rs8139383	Same	3.61E-07	6.22E-07	0.58
PI4KAP2	rs861787	rs861787	Same	4.30E-18	4.92E-18	0.87
	rs861848	rs861848	Same	9.03E-14	1.08E-14	8.37
	rs178047	rs2072516	1	5.34E-11	5.41E-11	0.99
	rs464694	rs464694	Same	2.23E-07	8.79E-07	0.25
	rs465500	rs458361	0.99	3.65E-36	4.25E-34	0.01

Table 3.3.3. Conditional eQTLs identified by the individual-level data approach and the conditional eQTL meta-analysis function of APEX. LD: linkage disequilibrium.

3.3.3 Comparison of the conditional eQTL meta-analysis results obtained using the conditional eQTL meta-analysis function of APEX and GCTA-COJO

We further compared conditional eQTL meta-analysis results obtained using the conditional eQTL meta-analysis function of APEX or GCTA-COJO for a larger set of genes (all 538 genes on chromosome 20) and a meta-analysis of three studies (TwinsUK, METSIM-1, and FUSION). GCTA-COJO uses an external reference panel to estimate the LD correlation between genetic variants across samples in meta-analysis participating studies. Therefore, the sample size of the reference panel and the genetic similarity of the reference panel to the participant studies were critical for GCTA-COJO. Of the 1428 biopsy donors across the three studies (TwinsUK, METSIM-1, and FUSION), half were of Finnish ancestry, half were of a broader European ancestry. We ran GCTA-COJO with three reference panels: 10K array-genotyped METSIM samples, 10K and 20K randomly selected UK Biobank (UKB) samples. The 10K METSIM samples are a representation of Finnish ancestry, and the 10K or 20K UKB samples are a broader representation of European ancestry.

Of the 538 genes on chromosome 20, APEX identified 0 eQTL for 107 (19.8%) genes, 1 eQTL for 295 (54.8%) genes, 2 eQTLs for 90 (16.7%) genes, and 3 to 5 eQTLs for 46 (8.5%) genes (Table3.3.4). Using a reference panel of 10K HRC-imputed METSIM samples, GCTA-COJO identified 0 eQTL for 114 (21.1%) genes, 1 eQTL for 2 (0.37%) genes, 2 eQTLs for 1 (0.19%) gene, 3 to 5 eQTLs 27 (5.02%) for genes and 6 to 37 eQTLs for 394 (73.2%) genes (Figure3.3.1). Similarly, using HRC-imputed genotypes of 10K or 20K UKB, GCTA-COJO identified 6 to 47 eQTLs for 402 (74.7%) and 400 (74.3%) genes respectively (Figure3.3.1). The results from the 538 gene analysis suggest that GCTA-COJO may have provided many spurious conditional eQTLs.

Number of independent eQTLs	0	1	2	3	4	5
Number of genes	107	295	90	39	6	1
Percent of tested genes (%)	19.8	54.8	16.7	7.2	1.1	0.2

Table 3.3.4. Number and percent of genes with n ($0 \leq n \leq 5$) eQTLs detected by using the conditional eQTL meta-analysis function of APEX. Variants were considered to be an eQTL if it had p $\leq 0.05/20,000 = 2.5 \times 10^{-6}$ with a given gene.

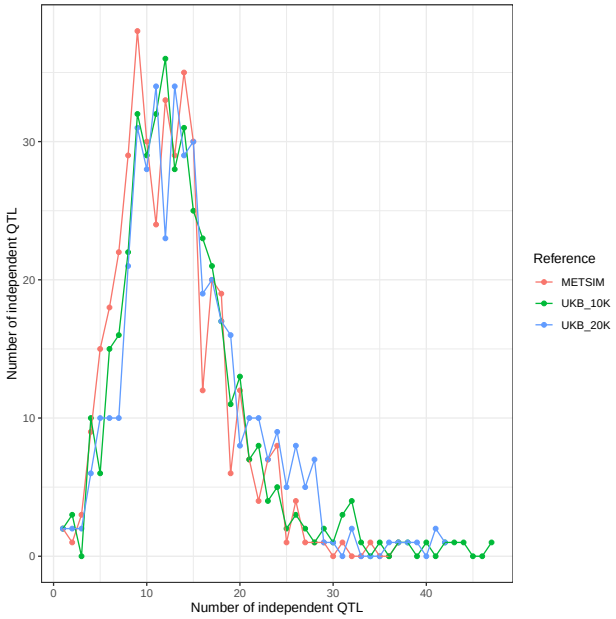


Figure 3.3.1. Number of genes with n (1≤n≤47) eQTL detected by GCTA-COJO using different reference panels for LD estimation. I identified conditional eQTLs for all 538 chromosome 20 genes, by providing GCTA-COJO the summary statistics of marginal eQTL associations and an external reference panel. I conducted the same analysis with three reference panels, 10k samples from the larger METSIM cohort, 10k and 20K randomly selected unrelated samples from the UK Biobank (UKB). Variants were considered to be an eQTL if it had p $\leq 0.05/20,000 = 2.5 \times 10^{-6}$ with a given gene.

3.3.4 **Genome-wide conditional eQTL analysis in individual studies and meta-analysis**

I performed genome-wide conditional eQTL analysis for each individual study using the single-study conditional eQTL analysis function of APEX and for meta-analysis using the conditional eQTL meta-analysis function of APEX. In each individual study, I tested genes that had $\geq$ 5 CPMs in $\geq$ 25% samples (22.3K-28.7K per study). A total of 30,604 genes were present in at least one study, and 19,108 of the genes were present in the intersection of all studies, 2165 genes were present only in TwinsUK, METSIM-1, METSIM-2, and FUSION, and 1812 genes were present only in METSIM-1, METSIM-2, GTEx, and FUSION (Figure 3.3.2).

The proportions of genes with zero to ten eQTLs were almost identical with and without the non-European individuals in the meta-analysis (Figure3.3.3). Of the 30,604 genes present in at least one study (Figure3.3.3A), I identified $\geq$ 1 eQTL for 39.3% – 49% of genes in individual studies, and for 63.9% of genes in the meta-analysis of European individuals. I

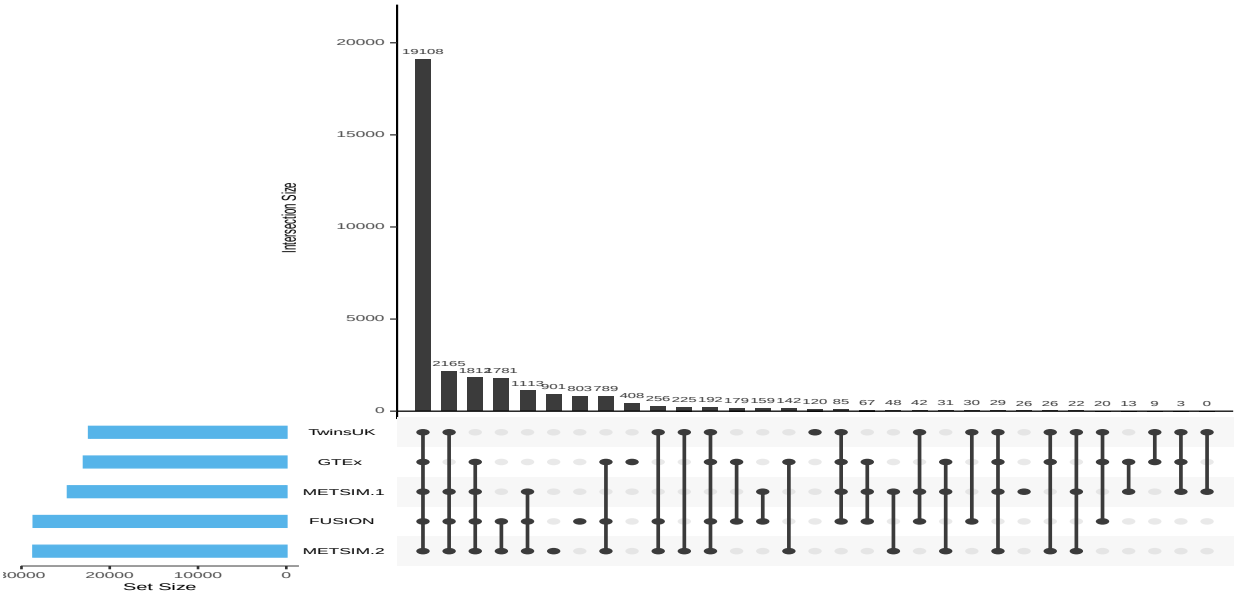


Figure 3.3.2. Number of genes present in different combination of studies. RNA-seq based gene expression data from TwinsUK, METSIM-1, METSIM-2, GTEx v8, FUSION were used in the meta-analysis. The blue bars on the left show the number of genes tested in a study. The black bars show the number of genes present in a given combination of studies, denoted by the dots under the x-axis.

identified exactly one eQTL for 28.8% – 39.1% of genes in individual studies, and 30.4% of genes in the meta-analysis of European individuals. I identified $\geq$ 2 eQTLs for 8.0% – 13.8% of genes in individual studies, and 33.6% of genes in the meta-analysis of European individuals. Compared to TwinsUK, the individual study with the largest sample size, the meta-analysis of European individuals increased the proportion of genes with $\geq$ 1 eQTL from 49% to 63.9%, and increased the proportion of genes with $\geq$ 2 eQTLs from 13.5% to 33.6%.

I stratified the results by whether the genes were in the intersection of all studies or in a subset of the studies. Of the 19108 genes present in all studies (Figure3.3.3B), I identified $\geq$ 1 eQTL for 43.2% – 54.4% of genes in individual studies, and $\geq$ 1 eQTL for 80.3% of genes in the meta-analysis. I identified exactly one eQTL for 34.4% – 39.1% of genes in individual studies, and 33.7% of genes in the meta-analysis. I identified $\geq$ 2 eQTLs for 8.8% – 15.3% of genes in individual studies and 46.6% of genes in the meta-analysis. Compared to TwinsUK, the meta-analysis increased the proportion of genes with $\geq$ 1 from 52.8% to 80.3%, and increased the proportion of genes with $\geq$ 2 eQTLs from 15.0% to

46.6%. Of the 19108 genes, meta-analysis identified eQTLs for 1794 genes for which individual studies did not identify any eQTL, and identified more eQTLs for 6192 of the 19108 genes. However, meta-analysis identified fewer eQTLs for 487 of the 19108 genes.

Of the genes not present in all studies (Figure3.3.3C), I identified $\geq$ 1 eQTL for 23.3% – 31.1% of genes in individual studies, and $\geq$ 1 eQTL for 33.7% of genes in the meta-analysis. I identified exactly one eQTL for 16.3% – 24.6% of genes in individual studies, and 22.8% of genes in the meta-analysis. I identified $\geq$ 2 eQTLs for 3.8% – 6.7% of genes in individual studies and 10.9% of genes in the meta-analysis.

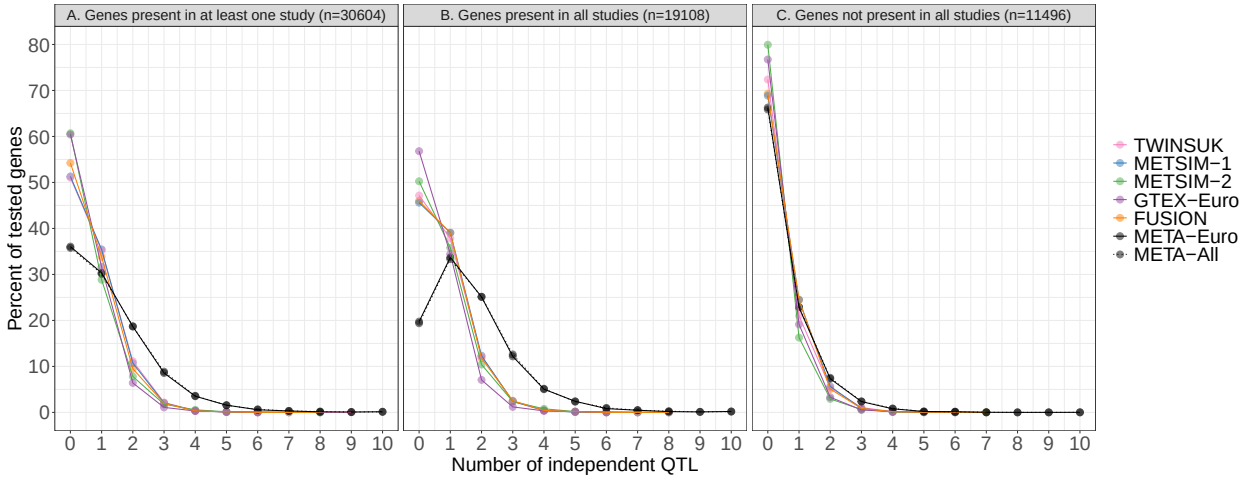


Figure 3.3.3. Proportion of tested genes with $(1 \leq N \leq 10)$ eQTLs. eQTL detection was performed for genes in the union set of studies (A), for genes in the intersection of studies (B), and for genes in a subset of studies (C). The x-axis denotes the number of conditional eQTLs, and the y-axis denotes the percent of tested genes with a given number of eQTLs.

3.3.5 Colocalization between adipose eQTLs identified in the meta-analysis and GWAS loci for cardiometabolic traits

Excessive accumulation and malfunction of subcutaneous adipose tissue may play a crucial role in the development of cardiometabolic diseases. Therefore, we sought to identify subcutaneous adipose tissue *cis*-eQTLs that shared potential causal variants with genetic associations for seven cardiometabolic diseases and related traits: T2D, BMI, WHR, WHRadjBMI, CAD, fasting glucose, and fasting insulin by performing colocalization analysis. Colocalization analysis requires that the GWAS study and eQTL study are from the same underlying population and have same LD patterns across samples. As the majority of participants of the GWAS studies used for the colocalization analysis and of the present

eQTL meta-analysis were of European ancestry, I used the results from individuals of European ancestry for the remaining analysis.

I accessed the evidence for colocalization between the GWAS variants of cardiometabolic disease and related traits and subcutaneous adipose tissue eQTLs identified in the meta-analysis using coloc2[115], [158]. Coloc2 is a Baysian method that estimates the posterior probability for five hypotheses. If the hypotheses H4 (both GWAS and eQTL had association signals and they share the same causal variant) had a posterior probability (PP4) $\geq$ 0.8, an eQTL was considered to be colocalized with a GWAS variant.

I performed colocalization analysis for 19,569 genes present in at least one study and had $\geq$ 1 eQTL. I first tested for colocalization between GWAS variants and eQTLs from 9,289 genes with exactly one eQTL, using the summary statistics of marginal variant-gene associations. eQTLs of 162 genes were colocalized with at least one GWAS locus, and eQTLs of 44 genes were colocalized with GWAS loci of more than one trait. Of the 162 colocalized genes, 20 genes were only identified by meta-analysis.

I next tested for colocalization between GWAS variants and the eQTLs for 10,280 genes with $\geq$ 2 eQTL detected in the meta-analysis, using conditional eQTL associations. eQTLs of 355 genes were colocalized with at least one GWAS locus, eQTLs of 100 genes were colocalized with GWAS loci of more than one trait. Of the 355 genes, eQTLs of eight genes were identified only in the meta-analysis. Stratifying the colocalization results from genes with $\geq$ 2 eQTL by primary eQTLs and secondary eQTLs, I observed that 171 primary eQTLs were colocalized with 159 GWAS loci; 221 secondary eQTLs were colocalized with 193 GWAS loci.

Taking the colocalization results from genes with exactly one eQTL and with $\geq$ 2 eQTLs together (Supplementary Table3.7.1), I identified colocalization for 61 T2D loci, 115 BMI loci, 110 WHR loci, 132 WHRadjBMI loci, four CAD loci, and four fasting glucose loci. A total of 517 genes had eQTLs colocalized with GWAS signals. Of the 517 genes, 334 genes had primary eQTL colocalized with at least one GWAS signal, 202 genes had secondary eQTLs colocalized with at least one GWAS signal. The number of genes with primary or

	Primary eQTL signals		Secondary eQTL signals	
	Colocalized GWAS loci	Colocalized genes	Colocalized GWAS loci	Colocalized genes
T2D	47	66	25	31
BMI	75	111	62	71
WHR	73	114	62	84
WHRadjBMI	94	140	63	83
CAD	2	6	2	2
Fasting glucose	3	4	1	1
Fasting insulin	0	0	0	0

Table 3.3.5. Number of genes with primary or secondary eQTLs colocalized with each of the seven cardiometabolic disease and traits.

secondary eQTLs colocalized with GWAS traits are shown in Table3.3.5.

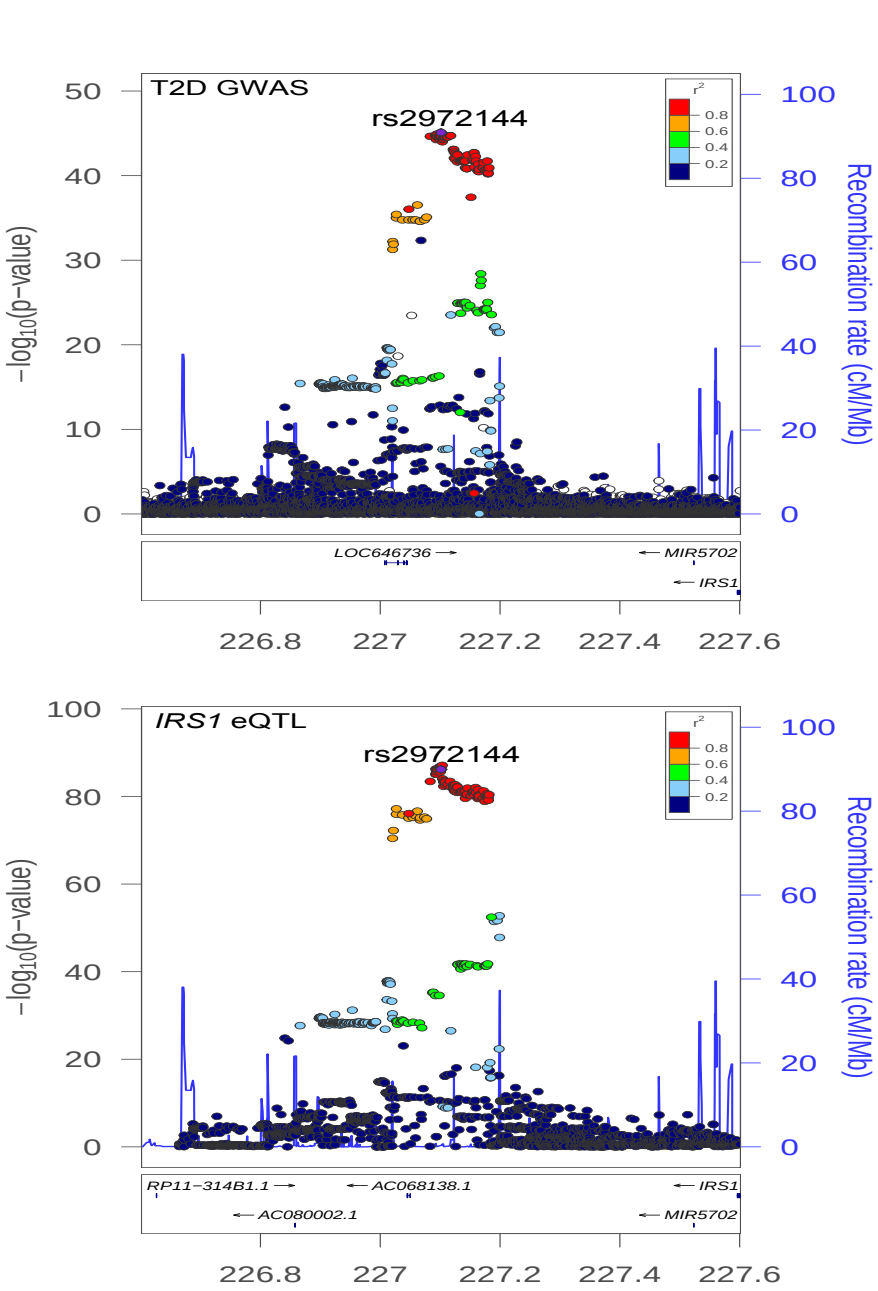


Figure 3.3.4. T2D GWAS locus rs2972144 is colocalized with the primary eQTL for *IRS1*, colored by 1000G Phase 3 European LD. (A) Marginal association plot for T2D meta-analysis from Mahajan et al. at the rs2972144 locus (p-value= 2.80×10^{-45}); (B) Marginal association plot for *IRS1* expression level in the adipose eQTL meta-analysis. The GWAS variant rs2972144, in high LD (R^2=0.97 and D'=0.99) with the lead eQTL variant for *IRS1* rs2138157, had strong association with *IRS1* expression level in the meta-analysis (p-value= 1.04×10^{-83}).

The colocalization analysis identified genes that have clear links to the disease mechanisms. As an example, the primary eQTL of *IRS1* (lead variant rs2138157, conditional p-value= 1.04×10^{-83}, PP4=0.96) was colocalized with the T2D GWAS signal (lead variant rs2972144 GWAS p-value = 2.80×10^{-45}) (Figure3.3.4), consistent with the previous colocalization evidence between the same GWAS locus and same *IRS1* eQTL[29]. rs2972144 and rs2138157 were in high LD (1000G EURO R^2=0.97,D'=0.99). rs2138157 had directionally consistent effects on *IRS1* across studies. There was only a single GWAS signal near *IRS1* and a second eQTL for *IRS1* revealed in the meta-analysis was not colocalized with the GWAS signal. The T2D risk allele rs2972144-G was associated with a lower expression level of *IRS1*. The IRS1 protein is a substrate for the insulin receptor and other tyrosine kinases. Reduced IRS1, caused by chronic insulin exposure, contributed to chronic insulin resistance in neonatal rat ventricular cardiomyocytes[159] and IRS1 protein degradation impaired glucose uptake in the adipose tissue of T2D mice[160].

We observed one instance where distinct GWAS signals at the same locus colocalized with independent eQTLs for the same gene. The GWAS locus for WHRadjBMI with lead variant rs863750 has a complex LD structure and encompasses many genes. Ying et al. dissected the variant-WHRadjBMI associations in this region into two GWAS signals, represented by rs863750 and rs7133378, respectively[157] (Figure 3.3.5). rs863750 and rs7133378 are located 95 kb away and are in low linkage disequilibrium (R^2=0.016,D'=0.15 in 1000G Euro). Ying et al.[157] and Raulerson et al.[29] found that the rs863750 WHRadjBMI signal was colocalized with the eQTL of *ZNF664*. From the initial colocalization analysis with the marginal GWAS associations (unseparated signals), I found many eQTL signals colocalized with the marginal GWAS associations at this locus. These colocalized eQTLs appeared by visual inspection to colocalize with the one of the two WHRadjBMI GWAS signals (rs863750 and rs7133378). The rs863750 WHRadjBMI signal appeared to be colocalized with the primary eQTL for *ZNF664* (rs10773049, conditional p-value= 2.87×10^{-90}, PP4=1) as well as the secondary eQTLs for *FAM101A* (rs10773049, conditional p-value= 8.22×10^{-10}, PP4=1) and *CCDC92* (rs863750, conditional p-value= 2.38×10^{-59}, PP4=1) (Table3.3.6). The rs7133378 WHRadjBMI signal appeared to be colo-

calized with the primary eQTL for *FAM101A* (rs7133378, conditional p-value= 2.82×10^{-12}, PP4=0.83). Although rs7133378 was also the secondary eQTLs for *ZNF664* (conditional p-value= 4.75×10^{-67}), no significant colocalization was detected between them.

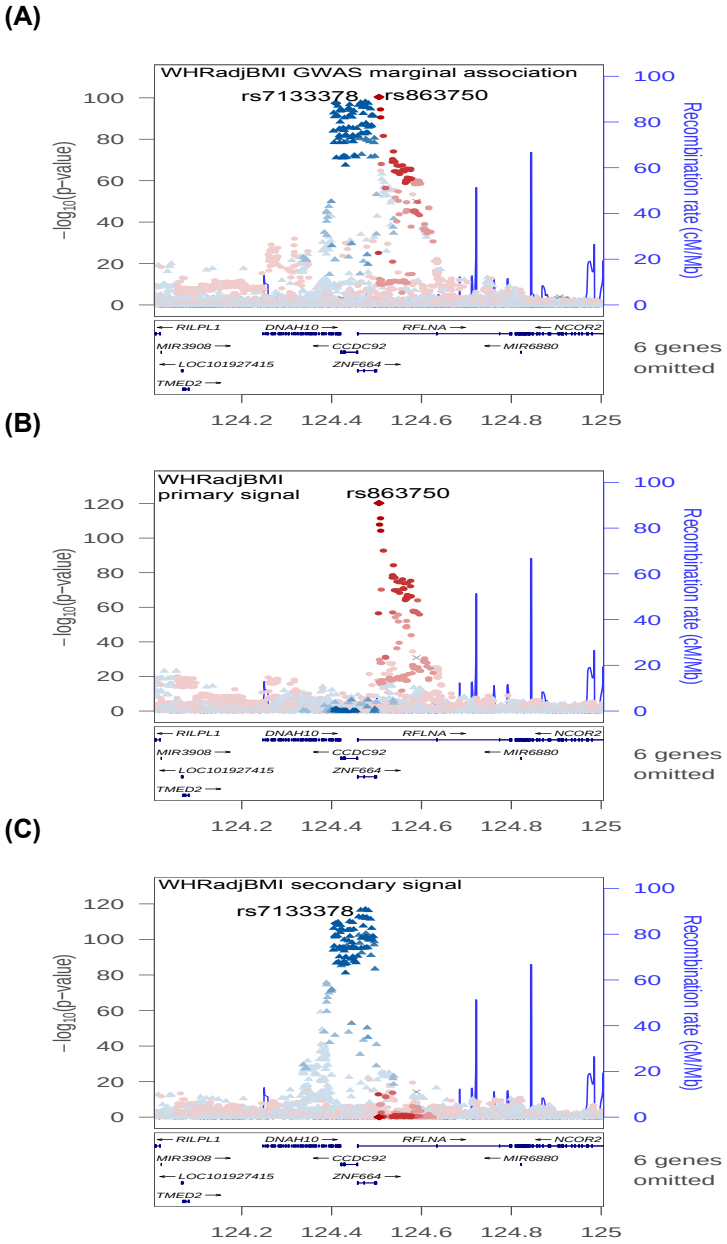
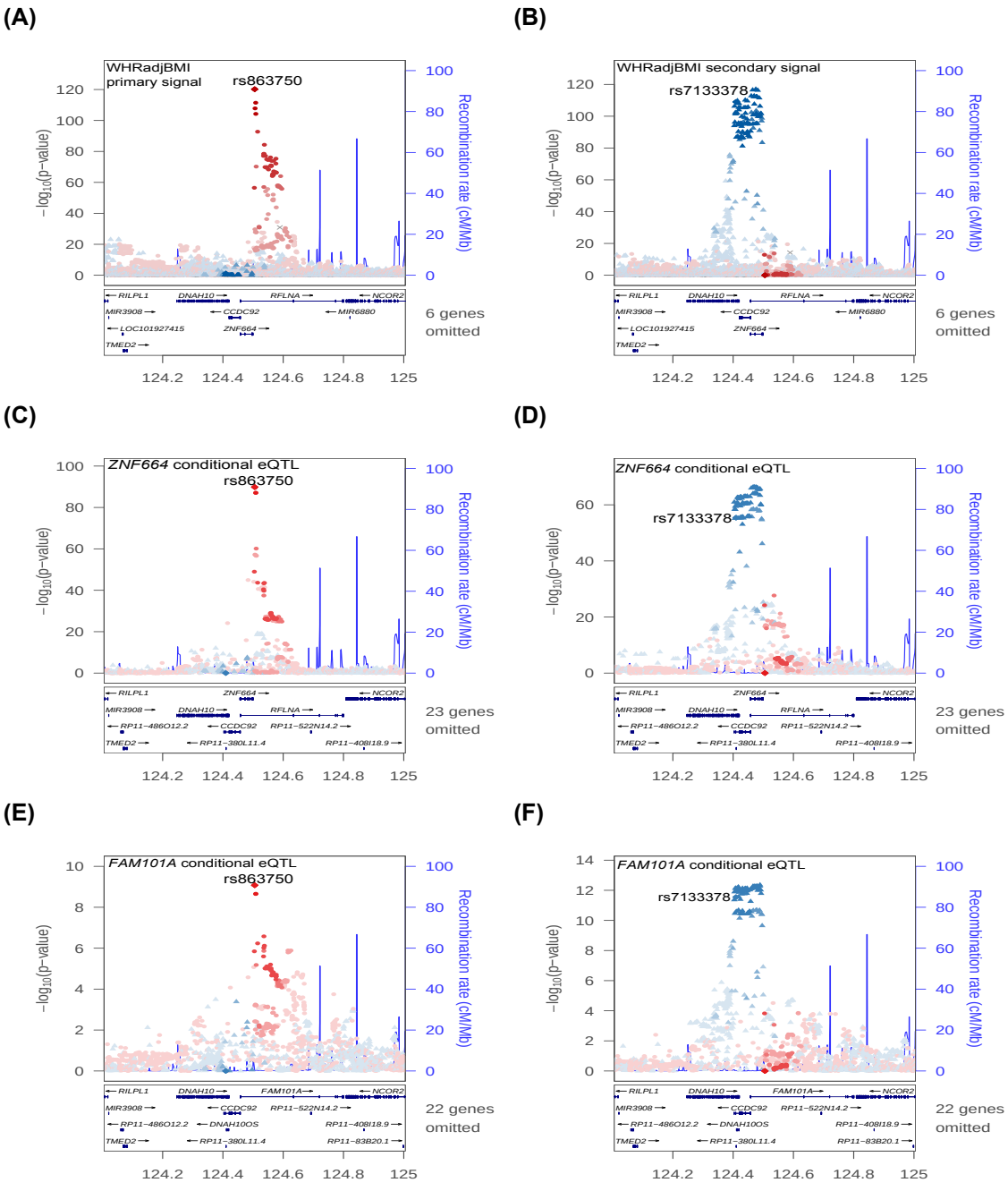


Figure 3.3.5. Two distinct WHRadjBMI singals (rs863750 and rs7133378) at the locus near *ZNF664*, colored by 1000 Genomes (1000G) Phase 3 Euro LD. (A) Marginal association plot for single-variant WHRadjBMI meta-analysis from Pulit et al. near *ZNF664*. rs863750 p-value= 4.17×10^{-101}, rs7133378 p-value= 4.0×10^{-98}; (B) Conditional association plot for WHRadjBMI conditioning on rs7133378 from approximate conditional analysis using GCTA-COJO. Conditioning on rs7133378, rs863750-WHRadjBMI association p-value = 6.40×10^{-121}; (C) Conditional association plot for WHRadjBMI conditioning on rs863750 from approximate conditional analysis using GCTA-COJO. Conditioning on rs863750, 7133378-WHRadjBMI association p-value =9.99×10^{-110}.

As the existence of multiple signals in the GWAS locus may reduce the power for coloc2[115], [158] to detect colocalization, we explored whether using the conditional summary statistics for the GWAS associations would mitigate such impact. For the two GWAS signals at this locus (rs863750 and rs7133378), we separately computed the approximate conditional summary statistics for each signal, conditioning on the other (using GCTA-COJO). The colocalization results using the separated WHRadjBMI signals found evidence for colocalization between the rs863750 WHRadjBMI signal and the primary eQTL for *ZNF664* (Figure 3.3.6C) and secondary eQTLs for *FAM101A* (Figure 3.3.6E) and *CCDC92* (Figure 3.3.6g), and between the rs7133378 WHRadjBMI signal and the secondary eQTL of *ZNF664* (PP4=0.99) (Figure 3.3.6D), and the primary eQTLs of *FAM101A* (Figure 3.3.6F), *DNAH10OS* (PP4=0.83) (Figure 3.3.6h) and *RP11-214K3.24* (PP4=0.96) (Figure 3.3.6j). Compared to the initial colocalization results with marginal GWAS associations, the colocalization results found additional colocalization between the rs7133378 WHRadjBMI signal and the eQTLs for *ZNF664*, *DNAH10OS* and *RP11-214K3.24* (Table3.3.6).

GWAS variant	Lead eQTL variant	gene	eQTL rank*	total number of eQTLs	eQTL associations in the joint model†		R^2	D'	PP.H4	
					beta	p			Marginal GWAS	Conditional GWAS
rs863750	rs10773049	*ZNF664*	1	8	0.28	2.87E-90	1.00	1.00	1.00	1.00
rs863750	rs10773049	*FAM101A*	2	2	0.11	8.22E-10	1.00	1.00	1.00	1.00
rs863750	rs863750	*CCDC92*	3	7	-0.27	2.38E-59	1.00	1.00	1.00	1.00
rs7133378	rs952632	*ZNF664*	2	8	0.22	4.75E-67	0.88	0.97	0.51	0.99
rs7133378	rs7133378	*FAM101A*	1	2	0.14	2.82E-12	1.00	1.00	0.83	0.95
rs7133378	rs10846580	*DNAH10OS*	1	3	0.26	1.19E-36	0.90	0.99	0.00	0.83
rs7133378	rs4765562	*RP11-214K3.24*	1	1	0.29	2.83E-16	0.89	0.99	0.00	0.96

Table 3.3.6. Primary and secondary eQTLs colocalized with the two-signal (rs863750 and rs7133378) WHRadjBMI GWAS locus near*ZNF664*. *eQTL_rank: the order of the variants being selected as an eQTL. †Joint model: all selected QTL variants are included in the model. For example, the model for *ZNF664* had eight eQTL variants. R^2 and D': Between the GWAS variant and the eQTL variant, estimated using 1000 Genomes Phase 3 European LD.


(A)
WHRadjBMI primary signal
rs863750
(B)
WHRadjBMI secondary signal
rs7133378
(C)
ZNF664 conditional eQTL
rs863750
(D)
ZNF664 conditional eQTL
rs7133378
(E)
FAM101A conditional eQTL
rs863750
(F)
FAM101A conditional eQTL
rs7133378
−log10(p−value)
Recombination rate (cM/Mb)
6 genes omitted
23 genes omitted
22 genes omitted
124.2
124.4
124.6
124.8
125

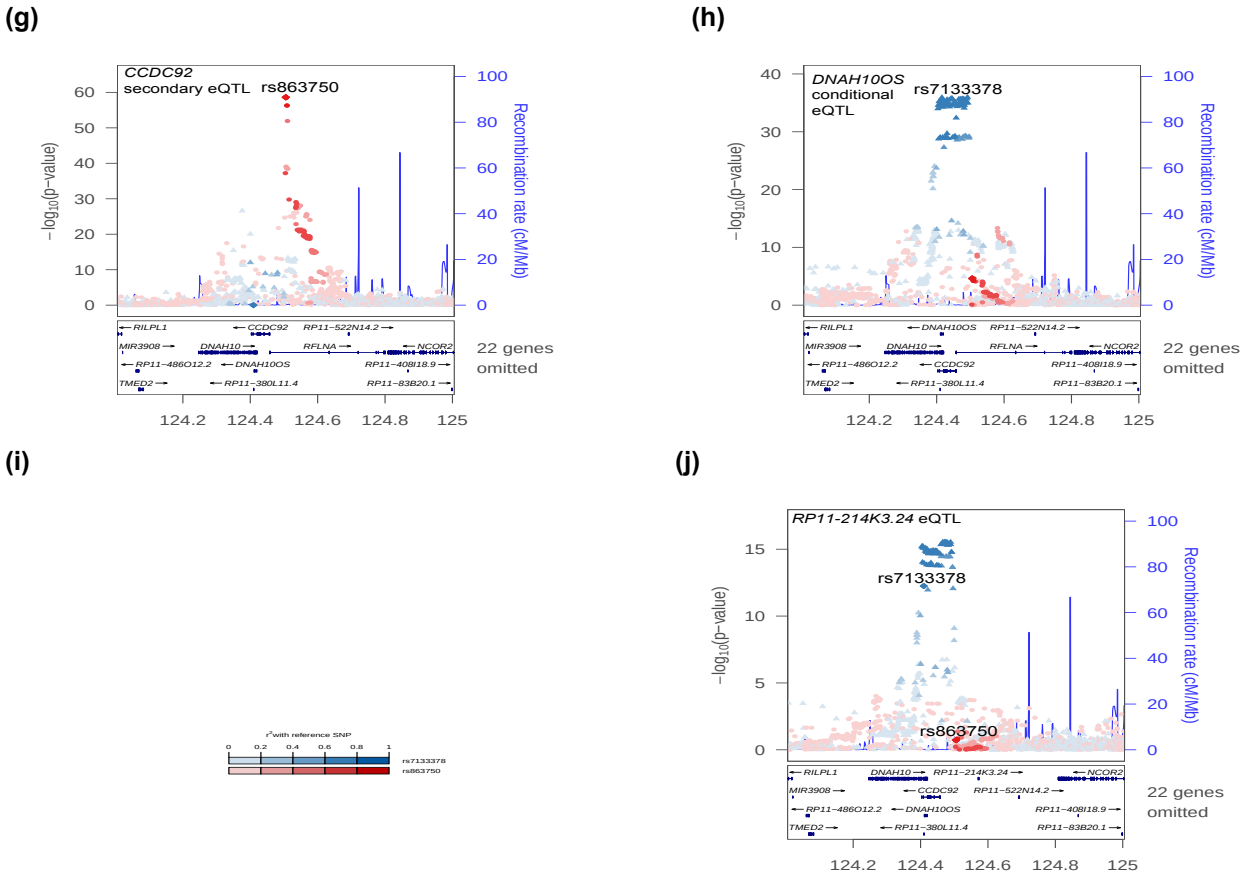


Figure 3.3.6. A two-signal WHRadjBMI locus (rs863750 and rs7133378) is colocalized with eQTLs of three and four genes correspondingly, colored by 1000G Phase 3 European LD.(A) Regional association plot for WHRadjBMI conditioning on rs7133378 from approximate conditional analysis using GCTA-COJO; (B) Regional association plot for WHRadjBMI conditioning on rs863750; (C) Residual eQTL associations for *ZNF664* after accounting for the other seven eQTLs except for the eQTL (rs10773049) that colocalized with the GWAS signal rs863750. Lead eQTL variant rs10773049 is in high LD with rs863750 (R^2=1.00 and D'=1); (D) Residual eQTL associations with *ZNF664* after accounting for the other seven eQTLs, representing the conditional association of eQTL rs952632. rs952632 colocalized with the GWAS signal rs7133378 (R^2=0.88 and D'=0.97); (E) Residual eQTL associations with *FAM101A* after accounting for the primary eQTL of *FAM101A* (rs7133378), representing the conditional association of the secondary eQTL (rs10773049) that colocalized with the GWAS signal rs863750; (F) Residual eQTL associations with *FAM101A* after accounting for the secondary eQTL of *FAM101A* (rs10773049), representing the conditional association of the primary eQTL (rs7133378); (G) Residual eQTL associations with *CCDC92* after accounting for the other six eQTLs except for the eQTL rs863750; (H) Residual eQTL associations with *DNAH10OS* after accounting for the other two eQTLs, representing the conditional association of eQTL rs10846580 that colocalized with the GWAS signal rs86375 (R^2=0.90 and D'=0.99); (I) eQTL associations with *RP11-214K3.24*. *RP11-214K3.24* eQTL rs4765562 colocalized with the GWAS signal rs7133378 (R^2=0.89 and D'=0.99).

Both the rs863750 and rs7133378 WHRadjBMI signals are associated with multiple other GWAS traits. rs863750-C allele is associated with higher BMI[155] and HDL[161], [162] as well as lower WHR[155] and triglyceride[163], [164]. rs7133378-A allele is associated with higher BMI[155], lower WHR[155], and higher reticulocyte count[165]. rs863750-C was associated with higher expression of both *ZNF644*, *CCDC92* and *FAM101A*. rs7133378-A allele was associated with the higher expression of *FAM101A*, *ZNF664*, *DNAH10OS* and *RP11-214K3.24*. rs863750 is an intronic variant within *FAM101A*, and rs7133378 is an intronic variant within *DNAH10OS* and *CCDC92*. Our knowledge is limited about the functions of these four genes. *ZNF664* is predicted to be a transcription factor. *CCDC92* is a coiled-coil domain protein interacting with proteins at the centriole–ciliary interface[166].

Our colocalization results revealed significant colocalization for 517 of 19,569 genes that had $\geq$ 1 eQTL. Of the 1794 genes that I identified eQTLs only in meta-analysis (not in single-study analyses), 28 genes were found to have eQTLs colocalized with a GWAS variant. One example was that WHRadjBMI loci rs3892816 (GWAS meta-analysis p-value= 4.0×10^{-13})[155] was colocalized the eQTL of *EXOC3L1* rs11552322 (eQTL meta-analysis p-value= 8.1×10^{-8}, PP4=0.90). rs3892816 and rs11552322 were in high LD (R^2=0.861 and D'=0.96) (Figure3.3.7). We did not identify any eQTLs for *EXOC3L1* within each individual study, but identified one eQTL (rs11552322) in the meta-analysis. rs3892816-T, the only GWAS signal in this locus, was associated with higher WHRadjBMI and a lower expression level of *EXOC3L1*. *EXOC3L1* encodes a protein of the exocyst complex, which functions as a tether of secretory vesicles to the plasma membrane to facilitate molecular trafficking[167], [168].

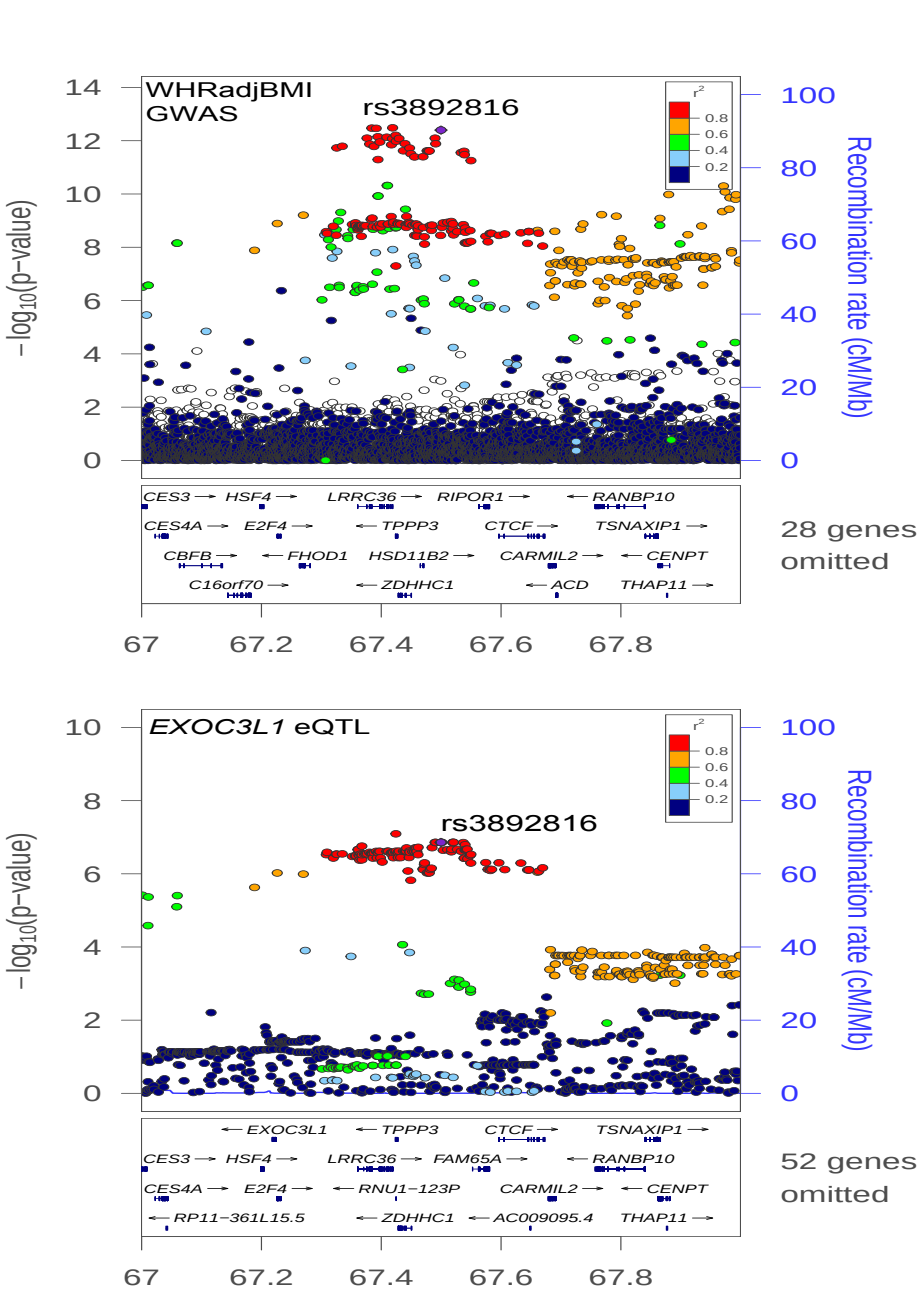
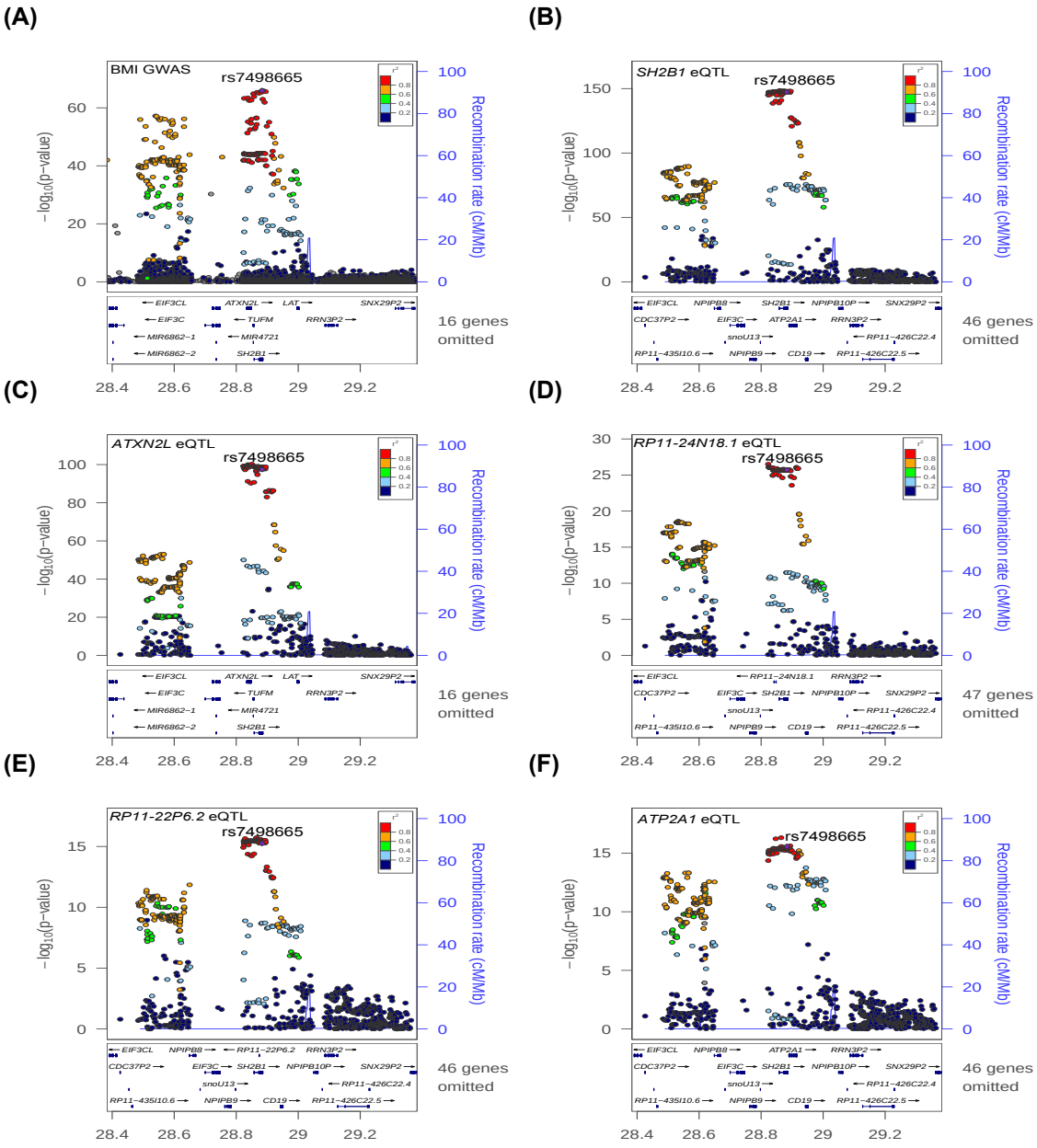


Figure 3.3.7. WHRadjBMI GWAS locus rs3892816 was colocalized with the eQTL for *EXOC3L1*, colored by 1000 Genomes (1000G) Phase 3 European LD. (A) Marginal association plot for WHRadjBMI meta-analysis from Pulit et al. at the rs3892816 locus (p-value= 4.0×10^{-13}); (B) Marginal association plot for *EXOC3L1* expression level in the adipose eQTL meta-analysis. The GWAS variant rs3892816, in high LD (R^2=0.861 and D'=0.96) with the lead eQTL variant for *EXOC3L1* rs11552322, were found to be associated with *EXOC3L1* expression level in the meta-analysis p-value= 8.1×10^{-8}.

We observed that a GWAS locus was colocalized with eQTLs of more than one gene for 130 GWAS loci. The GWAS variant that colocalized with the largest number of genes (seven) was the BMI GWAS locus rs7498665. rs7498665 colocalized with eQTL for seven genes, four protein-coding genes (*SH2B1*,*ATP2A1*, *ATXN2L*, *EIF3C*), and three antisense RNA *RP11-24N18.1*, *RP11-22P6.2*, *RP11-1348G14.5* (Figure3.3.8). The BMI increasing allele rs7498665-G was associated with lower expression of *SH2B1*, *RP11-24N18.1*, *RP11-22P6.2*, *RP11-1348G14.5* and *ATXN2L*, and higher expression of *ATP2A1* and *EIF3C*. In the meta-analysis, I observed ten eQTLs for *EIF3C*, three eQTLs for *ATXN2L*, and one eQTL for the other five genes. rs7498665 has genome-wide significant associations (p-value < 5×10^{-8}) with obesity[169], waist circumference[170], waist-hip ratio[155], height[171], weight[172], visceral adipose tissue measure[173] and T2D [174], [175]. rs7498665 is a missense variant (Thr484Ala, 1000G EUR MAF=0.26) of *SH2B1*[176], [177], which encodes a member of the SH2-domain containing mediators family. SH2B1 protein mediates activation of various kinases and mediates leptin, enhances insulin, and TrkA, TrkB and TrkC signaling[178].


(A)
BMI GWAS
rs7498665
16 genes omitted
(B)
SH2B1 eQTL
rs7498665
46 genes omitted
(C)
ATXN2L eQTL
rs7498665
16 genes omitted
(D)
RP11-24N18.1 eQTL
rs7498665
47 genes omitted
(E)
RP11-22P6.2 eQTL
rs7498665
46 genes omitted
(F)
ATP2A1 eQTL
rs7498665
46 genes omitted
−log10(p-value)
Recombination rate (cM/Mb)

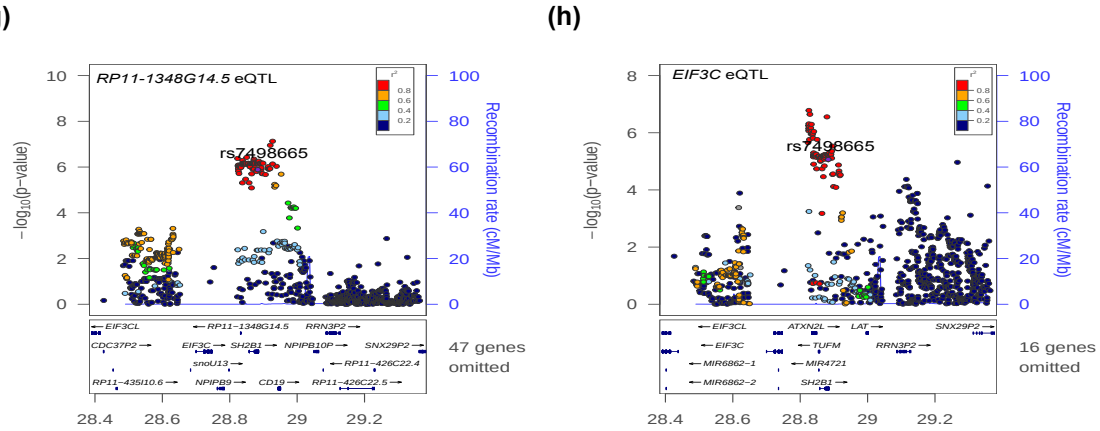


Figure 3.3.8. BMI GWAS locus rs7498665 is colocalized with eQTLs for seven genes, colored by 1000G Phase 3 European LD. (A). Marginal association plot for BMI meta-analysis from Pulit et al. at the rs7498665 locus. p-value= 1.14×10^{-66}; (B) Marginal association plot for *SH2B1* expression level in the adipose eQTL meta-analysis; (C) Residual eQTL associations with *ATXN2L* after accounting for the other two eQTLs, representing the conditional association of eQTL rs62036658 that colocalized with the GWAS signal rs7498665 (R^2=0.95 and D'=1.00); (D) Marginal association plot for *RP11-24N18.1* expression level in the adipose eQTL meta-analysis; (E) Regional association plot for *RP11-22P6.2* expression level in the adipose eQTL meta-analysis; (F) Regional association plot for *ATP2A1* expression level in the adipose eQTL meta-analysis; (G) Marginal association plot for *RP11-1348G14.5* expression level in the adipose eQTL meta-analysis; (H) Residual eQTL associations with *EIF3C* after accounting for the other nine eQTLs, representing the conditional association of eQTL rs7189927 that colocalized with the GWAS signal rs7498665 (R^2=0.85 and D'=0.99).

3.4 Discussion

I performed a *cis*-eQTL meta-anaylsis in subcutaneous adipose tissue of 2256 individuals of European ancestry from TwinsUK, METSIM, GTEx, and FUSION studies and identified conditionally independent eQTLs on a per gene basis. Integrating the conditional eQTL results with the GWAS signals for seven cardiometabolic traits: T2D, Body mass index, Waist-hip ratio, BMI adjusted waist-hip ratio, Coronary artery disease, fasting glucose and fasting insulin, I identified 517 genes that had eQTLs colocalized with GWAS signals.

To select a software for conditional eQTL meta-analysis across the genome, I first compared the conditional eQTL results obtained using the conditional eQTL meta-analysis function of APEX[150] and GCTA-COJO[125] to those obtained using the individual-level data approach (gold standard) for ten genes. Conditional eQTLs identified by APEX were consistent with those identified by the individual-level data approach in both the number of eQTLs and lead variants except that APEX detected one more eQTL with p-value slightly more significant than the p-value threshold for *PI4KAP2*. The p-values generated by APEX and the individual-level data approach were overall consistent, with individual-level data approach tend to result in smaller p-values for very significant (small) p-values. This discrepancy may in part be due to the probability distributions used to calculate p-values. APEX uses the t distribution to calculate meta-analysis p-values. The individual-level data uses the inverse-variance based approach in METAL to calculate the meta-analysis p-values, where the z distribution is used. As the t distribution has thicker tails than the the z-distribution and takes into account of the degree of freedom, the small p-values from the tails of t distribution are larger than those from the tails of the z-distribution. For these ten genes, GCTA-COJO identified six to 21 eQTLs and none of the secondary eQTLs was consistent with the individual-level data approach. The results from the comparison with the individual-level data approach suggested that APEX provided results closer to the individual-level data approach, compared to GCTA-COJO. To be able to further evaluate the consistency of the eQTLs identified by APEX and the individual-level data approach, we need to do comparisons on more genes. Second, I compared the conditional eQTL results from APEX and from GCTA-COJO on all genes on chromosome 20.

APEX detected one to five eQTL signals per gene. Regardless of the reference panel used, GCTA detected a large number (up to 47) of conditionally independent eQTLs per gene, suggesting potentially spurious eQTLs were identified. Given our results and the fact that GCTA-COJO has not been used in dissecting marginal eQTL associations into independent eQTLs, although it has been widely used in dissecting marginal GWAS associations into independent signals[33], [100], [179], [180], we recommend caution using GCTA-COJO for conditional eQTL detection.

One possible explanation for the large numbers of eQTLs detected by GCTA-COJO was that GCTA-COJO estimated LD within the study sample using an external reference panel. GCTA-COJO has almost exclusively been used to estimate LD for phenotypic GWA studies where the sample sizes are at the scale of tens of thousands of people. A better estimation is achieved when the reference panel is large ($\geq$ 5K)[125] and genetically similar to GWAS participants. Here, we need to estimate LD between genetic variants across 2K people from which the eQTL statistics were derived. It is likely that it is less accurate to estimate LD between variants for a small sample set using the reference panel, due to the randomness intrinsic to the small sample set.

TwinsUK, METSIM, GTEx, and FUSION studies collected genotype data and measured gene expression levels using RNA-seq in subcutaneous adipose tissue samples from participants, enabling us to perform the *cis*-eQTL detection in each study. For the 19.1K genes in the intersection of studies, we compared the proportion of genes with $\geq$ 1 eQTL (eGene discovery) in each study. METSIM-1 had the largest proportion (54.4%) of genes with$\geq$1 eQTL, GTEx-Euro had the smallest proportion (43.2%). Several factors can impact the proportion of genes identified to have eQTL. For example, the discovery of genes with $\geq$ 1 eQTL increases with sample sizes[31]. TwinsUK had the largest sample size(722), while FUSION had the smallest sample size (280). METSIM-1, METSIM-2 and GTEx had sample sizes of more than 400. In addition, the cell-type heterogeneity may impact the eGene discovery. On one hand, biopsies that are more homogeneous in cell-type compositions may result in less variation in gene expression levels and thereby increase the eGene discovery. On the other hand, biopsies that are more heterogeneous in cell-type

compositions may allow for the discovery of eQTLs existing in a larger set of cell types. In our previous work[141], we estimated the adipocyte proportions of TwinsUK, METSIM-1 and FUSION samples and found that METSIM-1 had the lowest adipocyte proportions on average, which was likely due to blood contamination of needle biopsy.

We used genotype and RNA-seq based gene expression data from the five studies (TwinsUK, METSIM-1, METSIM-2, GTEx, and FUSION) and performed the largest (n=2256) subcutaneous adipose tissue *cis*-eQTL meta-analysis to our knowledge. For the 19.1K genes in the intersection of all studies, compared to TwinsUK which was the individual study with the largest sample size, the proportion of genes with $\geq$ 2 eQTL was increased from 15.0% to 46.6%, while the proportion of genes with exactly one eQTL was slightly decreased from 37.8% to 33.7%. The decrease in the proportion of genes with exactly one eQTL was expected because more genes were found to have $\geq$ 2 eQTLs in the meta-analysis compared to individual-study analysis. These results showed that the meta-analysis greatly increased eQTL discovery, especially secondary eQTL discovery. Furthermore, we identified 221 secondary eQTLs colocalized with GWAS loci, which we would not have identified if we had only identified primary eQTLs through the meta-analysis. GWAS variants may affect disease susceptibility by acting as either primary or secondary eQTLs. Compared to primary eQTLs, secondary eQTLs tend to reside more distally from the gene they are associated with, and more frequently found to be tissue/cell type specific eQTLs[115]. Knowing whether a GWAS variant is the primary or secondary eQTL of a gene help us understand the biological mechanisms underlying it.

In the second chapter (FUSION study analysis chapter), we used fastEnloc[83], [84] for the colocalization analysis because we had access to genotype and gene expression data from the FUSION study, which allowed us to use DAP[82] to compute the posterior probability for each independent eQTL signal. FastEnloc uses the posterior probability of each independent eQTL signal generated by DAP to test for colocalization between a GWAS signal and each independent eQTL. In addition to fastEnloc, coloc2[115] is another widely used software for colocalization analysis[29], [115], [157]. Coloc2 uses summary statistics (regression coefficients and their variance) of eQTL associations and internally converts

eQTL association p-values to posterior probabilities assuming one causal variant exists for an eQTL signal. In this chapter (adipose eQTL meta-analysis chapter), we identified primary and secondary eQTLs through conditional eQTL meta-analysis and aimed to use these eQTLs for colocalization analysis. We could not use DAP to compute posterior probabilities of meta-analysis eQTLs because we did not have access to the individual-level gene expression and genotype data from each study. Therefore, we could not use fastEnloc for colocalization analysis. However, we had summary statistics (regression coefficients and their variance) of eQTL associations from the conditional eQTL meta-analysis. We also isolated each conditional eQTL signal and obtained the summary statistics of each independent eQTL signal through the "all-but-one" analysis, which were sufficient for coloc2 to compute posterior probabilities of each independent eQTL for the subsequent colocalization analysis. For genes with $\geq$ 2 eQTLs, we were still able to identify colocalization with multiple eQTLs while using coloc2 by applying the colocalization analysis to each GWAS locus-conditional eQTL pair.

In the colocalization results, we observed 130 GWAS loci, each of which colocalized with eQTLs of more than one gene. The BMI GWAS variant rs7498665 was colocalized with eQTLs for the largest number (seven) of genes. rs7498665 is a missense variant located in one of the seven genes, *SH2B1*[176], [177]. Several other coding variants in *SH2B1* are linked to obesity[176], insulin resistance, and T2D[174], [175] in humans. A multitude of evidence suggests that *SH2B1* is essential for regulating energy balance, body weight, peripheral insulin sensitivity, and glucose homeostasis[176]. In addition, rs7498665 was associated with the expression level of *ATP2A1*, *ATXN2L*, *EIF3C*, *RP11-24N18.1*, *RP11-22P6.2*, and *RP11-1348G14.5*. Missense variants can also work as eQTLs[31], besides leading to changes in amino acid makeup and the function of the encoded protein. It is possible that the BMI GWAS variant rs7498665 not only changes the amino acid of the SH2B1 protein but also affects the expression level of *SH2B1* and the other six genes. Although no other eQTL was identified for *SH2B1*, the possibility that another variant in high LD with rs7498665 is the causal variant responsible for the changes in gene expression levels could not be excluded. *ATP2A1* encodes one of the SERCA Ca(2+)-ATPases,

which catalyzes the hydrolysis of ATP coupled with the translocation of calcium from the cytosol to the sarcoplasmic reticulum lumen, and thus plays a fundamental role in muscular excitation and contraction[181]. *ATXN2* encodes a protein involved in endocytosis, mTOR signaling, ribosomal translation, and mitochondrial function[182]. Coding and non-coding genetic variants in the *ATXN2* were found to cause severe early-onset obesity in children[183]. *EIF3C* encodes a protein, which is part of EIF3. EIF3 is one of 12 known Eukaryotic translation initiation factors (EIFs) and is closely connected to cell growth cell cycle, and tumorigenesis. EIF3C knockdown inhibited cell proliferation and promoted apoptosis in a pancreatic cancer cell[184] and EIF3C-enhanced exosome secretion increased angiogenesis and accelerated tumor progression for human hepatocellular carcinoma[185].

I performed colocalization analysis using the marginal GWAS associations and conditional eQTLs identified in the meta-analysis. Of all the colocalization results based on marginal GWAS signals, I observed only one instance, where the marginal WHRadjBMI signal colocalized with two conditionally independent eQTLs of a single gene *FAM101A*. This observation exemplified the situation where the allelic heterogeneity in the genetic control at the gene level can propagate to the phenotypic level. Two distinct WHRadjBMI signals (rs863750 and rs7133378) exist in the GWAS locus near *ZNF664*. The rs863750 WHRadjBM signal was found to be colocalized with the primary eQTL of ZNF664[157]. Our colocalization results suggested that the primary WHRadjBMI GWAS signal rs863750 was colocalized with the primary eQTL for *ZNF664* and the secondary eQTLs for *FAM101A* and *CCDC92*. The secondary WHRadjBMI GWAS signal rs7133378 was colocalized with the primary eQTL for *FAM101A*. Although the secondary WHRadjBMI GWAS signal rs7133378 appeared to share the same causal variants with the eQTLs for other nearby genes in the regional association plots, the colocalization probabilities were not significant. For this two-signal (rs863750 and rs7133378) WHRadjBMI locus alone, I estimated the conditional summary statistics for the GWAS associations using the approximate conditional association analysis. Compared to the initial colocalization results with the marginal GWAS associations, the colocalization results using conditional GWAS associations iden-

tified additional colocalization between the secondary GWAS signal rs7133378 and the eQTLs for *ZNF664*, *DNAH10OS*, and *RP11-214K3.24*. This observation suggests that although I have identified the multiple eQTLs for the genes, if the multiple signals were not separated at multi-signal GWAS loci, it may still diminish the power to detect significant colocalizations.

For the scenario that a GWAS locus is colocalized with eQTLs of multiple genes, it is possible that these genes might work alone or in combination to confer disease risk. However, it is also possible that not all or none of them are involved in the biological mechanisms contributing to the predisposition to diseases. Colocalization analysis cannot distinguish the causal effect scenario (the same causal variant affects a gene and disease risk) from the pleiotropic effect scenario (the same causal variant affects a gene and disease risk independently) or the different causal variant in high LD scenario (different causal variants affect a gene and disease risk separately). Therefore, some or even all of the genes may not mediate the disease risk conferred by the GWAS locus. Integrating these colocalization results with other data and with functional follow-up are necessary to further elucidate the possible mechanisms driven by these genes.

The growing sample sizes in both phenotypic traits GWAS and eQTL studies will place the field in an increasingly better position to investigate the multiple regulatory effects that contribute to the variations at the gene and phenotype level. Compared to colocalization analysis based on marginal associations for both GWAS and eQTL associations or dissected signals for either one of them, future colocalization analysis based on clearly dissected signals for both GWAS and eQTL associations will have stronger power and will likely lead to a more complete characterization of the functional impact of GWAS variants. This improved knowledge will help translate the associations into causative mechanisms and thereby develop therapeutic drugs and approaches.

3.5 **My contributions**

The work is a joint project with TwinsUK and METSIM groups to combine RNA-seq based eQTL associations from TwinsUK, METSIM, GTEx and FUSION via meta-analysis. Dr.

Julia Moustafa performed PEER factor analysis, generated score statistics and variance-covariance matrices for conditional eQTL meta-analysis using APEX for TwinsUK. Sarah Brotman performed PEER factor analysis, generated score statistics and variance-covariance matrices for METSIM and GTEx. I performed PEER factor analysis, generated score statistics and variance-covariance matrices for FUSION. I compared the conditional eQTLs detected for ten genes in the meta-analysis of FUSION and GTEx v7 data by using APEX or GCTA-COJO to those by using individual-level data. I also compared the conditional eQTL results obtained using APEX and GCTA-COJO for 538 chromosome 20 gene in the meta-analysis of three studies (TwinsUK, METSIM-1, and FUSION). I performed genome-wide conditional analyses in individual studies and in meta-analysis using APEX. I performed colocalization analysis for the seven cardiometabolic diseases and traits. I wrote all the texts with the guidance of Dr. Laura Scott and created all figures.

3.6 **Supplementary figures**

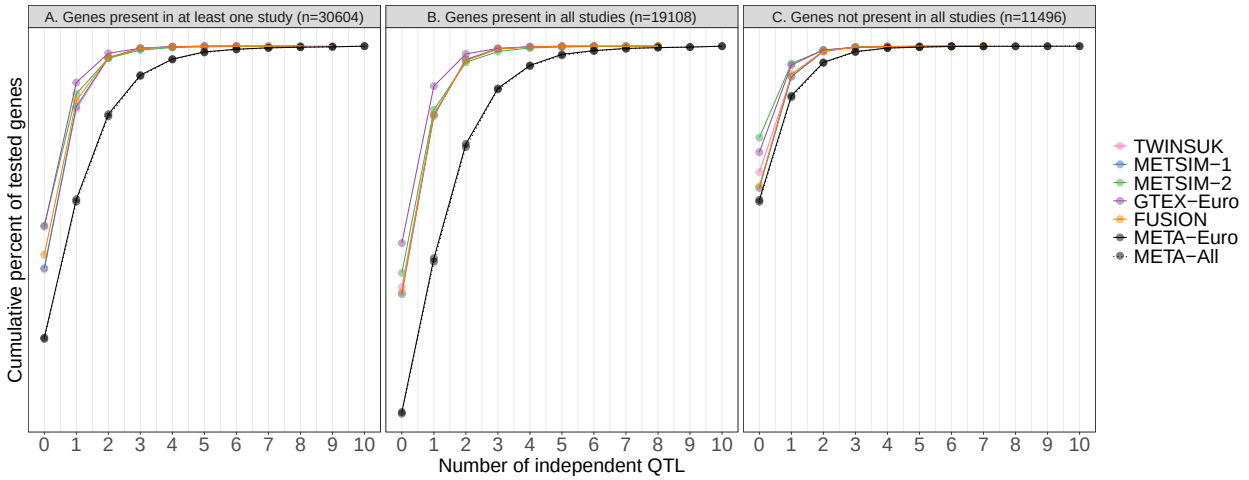


Figure 3.6.1. Cumulative proportion of tested genes with $(1 \leq N \leq 10)$ eQTLs. eQTL detection was performed for genes in the union set of studies (A), for genes in the intersection of studies (B), and for genes in a subset of studies (C). The x-axis denotes the number of conditional eQTLs, and the y-axis denotes the percent of tested genes with a given number of eQTLs.

3.7 **Supplementary Tables**

Table 3.7.1. Significant colocalization between cardiometabolic disease trait GWAS loci and eQTLs identified in meta-analysis (PP4 > 0.95)

Trait	GWAS SNP	PP.H4.abf	GWAS Beta	GWAS p	eQTL beta	eQTL p	Symbol	eQTL_rank	Total eQTL
T2D	rs2581782	0.97	-0.035	4.60E-08	-0.58	2.65E-187	*RFT1*	1	1
T2D	rs508419	0.98	-0.079	1.20E-25	0.46	1.23E-70	*ANK1*	1	1
T2D	rs1049481	0.99	0.045	6.70E-12	0.16	3.15E-43	*CALR*	1	1
T2D	rs2943647	0.97	-0.094	2.80E-45	-0.21	6.29E-33	*RP11-395N3.2*	1	1
T2D	rs2421016	0.98	-0.046	2.50E-13	-0.10	1.60E-28	*PLEKHA1*	1	1
T2D	rs76895963	1.00	0.48	5.30E-70	0.64	9.12E-23	*RP11-264F23.3*	1	1
T2D	rs3768321	0.99	0.085	1.30E-26	-0.17	6.56E-20	*RP11-69E11.8*	1	1
T2D	rs2242517	0.99	0.045	4.40E-12	-0.07	4.49E-13	*CTC-425F1.4*	1	1
T2D	rs76895963	1.00	0.48	5.30E-70	0.56	3.24E-12	*RP11-264F23.4*	1	1
T2D	rs2581787	0.95	0.036	3.00E-08	-0.18	2.07E-11	*RP11-894J14.5*	1	1
T2D	rs3742305	0.98	0.041	1.70E-08	-0.08	7.27E-11	*HMGB1*	1	1
T2D	rs62271373	1.00	0.088	1.00E-09	-0.17	2.26E-10	*TSC22D2*	1	1
T2D	rs2972144	0.97	-0.094	7.90E-46	-0.13	3.28E-09	*RP11-395N3.1*	1	1
T2D	rs7124681	0.97	0.037	6.40E-09	0.04	3.35E-07	*NDUFS3*	1	1
T2D	rs516946	0.98	-0.08	4.70E-26	0.13	4.34E-07	*RP11-930P14.1*	1	1
T2D	rs4977213	0.98	-0.051	4.40E-14	0.05	3.02E-06	*DGAT1*	1	1
T2D	rs1061810	0.96	0.05	8.50E-13	-0.72	0	*HSD17B12*	1	3
T2D	rs764729	0.95	-0.047	1.10E-10	-0.89	4.99E-193	*ITGB6*	1	2
T2D	rs2723064	0.96	0.051	3.90E-15	-0.42	5.10E-112	*CEP68*	1	2
T2D	rs2972146	0.96	0.094	2.80E-45	-0.30	1.04E-83	*IRS1*	1	2
T2D	rs11926707	0.98	-0.038	1.50E-08	-0.23	1.62E-66	*PTH1R*	1	2
T2D	rs1061810	0.96	0.05	8.50E-13	-0.36	3.28E-65	*RP11-613D13.5*	1	2
T2D	rs2925979	1.00	0.053	2.10E-14	-0.22	1.81E-51	*CMIP*	1	3
T2D	rs76895963	1.00	0.48	5.30E-70	0.88	2.96E-48	*CCND2*	1	2
T2D	rs7679066	0.99	0.038	4.60E-08	-0.17	3.91E-48	*DCAF16*	1	2
T2D	rs11688682	1.00	-0.058	1.40E-14	-0.27	1.19E-41	*INHBB*	1	2
T2D	rs849135	1.00	-0.092	6.70E-48	0.20	1.25E-40	*JAZF1*	1	3
T2D	rs4729854	0.98	0.037	3.30E-08	-0.32	1.00E-29	*RASA4B*	3	7
T2D	rs3776706	0.96	-0.053	6.00E-13	-0.20	1.57E-24	*FST*	1	2
T2D	rs649961	0.99	0.038	1.30E-09	0.22	2.02E-23	*SLC12A8*	1	2
T2D	rs2453056	0.97	0.073	3.10E-14	-0.59	3.59E-23	*RP5-1042I8.7*	2	2
T2D	rs3768321	1.00	0.085	1.30E-26	-0.25	2.45E-21	*PABPC4*	1	4
T2D	rs1493694	0.97	0.084	2.10E-16	0.28	6.37E-21	*RP5-1042I8.7*	1	2
T2D	rs650558	0.98	0.051	9.10E-12	0.24	3.12E-16	*HSD17B1P1*	2	2
T2D	rs11688682	1.00	-0.058	1.40E-14	-0.26	4.16E-15	*AC073257.2*	2	2
T2D	rs9860221	0.99	0.055	9.20E-15	0.13	1.07E-14	*ADAMTS9*	1	2
T2D	rs11072553	0.98	-0.046	1.00E-09	-0.09	7.45E-13	*IMP3*	2	2
T2D	rs11709077	0.96	-0.11	1.60E-27	-0.16	1.70E-12	*TIMP4*	3	3
T2D	rs3747207	0.98	0.049	2.10E-10	0.13	9.32E-12	*PNPLA3*	1	2
T2D	rs35186585	1.00	-0.046	1.50E-09	-0.17	1.69E-11	*HSD17B12*	3	3
T2D	rs4729854	0.97	0.037	3.30E-08	-0.24	2.23E-11	*POLR2J2*	10	10
T2D	rs10228495	0.98	0.036	2.50E-08	-0.18	5.54E-11	*RP11-514P8.8*	6	10
T2D	rs10157145	0.97	-0.036	1.10E-08	-0.10	2.03E-10	*NUAK2*	2	2
T2D	rs55966194	0.99	0.047	7.00E-11	-0.17	2.45E-09	*EYA2*	3	4
T2D	rs13389219	1.00	-0.061	2.00E-20	-0.14	2.51E-09	*GRB14*	2	4
T2D	rs4102217	0.96	0.061	8.00E-14	0.09	7.90E-08	*LTBP3*	4	4
T2D	rs55966194	0.99	0.047	7.00E-11	-0.29	5.00E-07	*RP5-1050K3.3*	1	2
T2D	rs3797580	0.95	0.053	2.10E-15	-0.12	1.38E-06	*POC5*	2	2
T2D	rs7640294	0.95	0.036	3.00E-08	-0.08	4.30E-06	*TMEM110*	4	5
T2D	rs11709077	0.95	-0.11	1.60E-27	-0.12	2.27E-05	*SYN2*	4	4
BMI	rs3888190	0.96	0.0284	1.57E-66	-0.28	1.38E-148	*SH2B1*	1	1
BMI	rs2232015	0.98	-0.0123	9.02E-12	-0.46	2.07E-144	*PRMT6*	1	1
BMI	rs17207196	1.00	-0.022	1.58E-36	0.37	1.70E-133	*POM121C*	1	1
BMI	rs17207196	1.00	-0.022	1.58E-36	-0.44	9.79E-79	*AC018720.10*	1	1
BMI	rs6463489	0.96	0.0167	2.50E-10	-0.44	4.51E-73	*FBXL18*	1	1
BMI	rs6079138	0.98	-0.0107	4.74E-10	-0.19	9.01E-73	*NSFL1C*	1	1

Table 3.7.1 continued from previous page

Trait	GWAS SNP	PP.H4.abf	GWAS Beta	GWAS p	eQTL beta	eQTL p	Symbol	eQTL_rank	Total eQTL
BMI	rs34517439	1.00	0.0391	3.40E-39	-0.24	3.16E-40	*FUBP1*	1	1
BMI	rs984222	0.98	0.0123	8.76E-14	0.17	7.51E-31	*TBX15*	1	1
BMI	rs7133378	1.00	0.0127	2.93E-13	0.27	9.19E-31	*RP11-380L11.4*	1	1
BMI	rs13240600	0.98	0.018	7.90E-16	-0.12	2.00E-27	*CPSF4*	1	1
BMI	rs1784460	0.96	0.0138	2.46E-15	-0.48	5.65E-26	*RP11-110I1.14*	1	1
BMI	rs7928810	0.98	0.0117	1.96E-12	0.18	6.14E-25	*NCR3LG1*	1	1
BMI	rs34720381	1.00	0.0229	7.89E-12	0.14	2.88E-22	*PRRC2C*	1	1
BMI	rs10788797	0.96	-0.0194	5.25E-14	-0.31	1.57E-18	*BNIPL*	1	1
BMI	rs12022461	0.95	-0.0167	9.78E-14	-0.12	4.47E-15	*S100PBP*	1	1
BMI	rs3811125	0.99	-0.0152	9.42E-13	0.31	2.10E-14	*RP11-440G5.2*	1	1
BMI	rs7871866	0.97	0.0179	7.56E-14	0.20	1.43E-13	*RP11-339B21.13*	1	1
BMI	rs9379827	0.98	-0.0134	4.44E-12	-0.20	1.75E-13	*HIST1H2BD*	1	1
BMI	rs2260051	0.97	-0.0184	6.50E-22	0.06	8.17E-13	*AIF1*	1	1
BMI	rs879620	1.00	0.0226	8.45E-39	-0.29	1.02E-11	*RP11-462G12.2*	1	1
BMI	rs1045411	0.98	-0.0139	6.66E-14	-0.08	9.65E-11	*HMGB1*	1	1
BMI	rs12714199	0.97	-0.0141	3.22E-16	-0.05	1.21E-10	*KDM3A*	1	1
BMI	rs2246012	0.99	-0.0161	1.15E-13	-0.20	3.54E-10	*ARG1*	1	1
BMI	rs10668	0.97	-0.0117	1.40E-11	0.11	1.55E-08	*MYLK-AS1*	1	1
BMI	rs2681781	0.96	-0.0235	8.47E-48	0.23	2.74E-08	*CTD-2330K9.3*	1	1
BMI	rs12692596	1.00	0.012	1.03E-12	0.06	3.45E-08	*RBMS1*	1	1
BMI	rs979012	0.99	-0.0186	1.03E-27	0.11	1.72E-07	*BMP2*	1	1
BMI	rs4911382	0.96	0.0127	1.69E-13	0.04	1.72E-07	*RALY*	1	1
BMI	rs6265	0.98	-0.0413	7.40E-89	-0.15	1.89E-07	*BDNF*	1	1
BMI	rs7124681	0.99	0.0257	3.96E-55	0.04	3.35E-07	*NDUFS3*	1	1
BMI	rs3957285	0.98	0.015	2.46E-18	-0.09	5.76E-06	*CTB-31O20.2*	1	1
BMI	rs17207196	1.00	-0.022	1.58E-36	0.63	4.16E-164	*STAG3L1*	1	2
BMI	rs7127212	0.95	0.0135	4.16E-15	0.27	2.65E-128	*HMBS*	1	3
BMI	rs983583	0.99	-0.012	6.39E-12	-0.27	7.61E-107	*YWHAZ*	1	3
BMI	rs863750	1.00	-0.0112	1.12E-11	-0.28	1.44E-90	*ZNF664*	1	8
BMI	rs1700138	0.96	0.0108	3.99E-09	0.33	2.76E-72	*MMP16*	1	3
BMI	rs12597511	0.99	-0.0209	2.85E-34	-0.25	1.07E-59	*KAT8*	1	2
BMI	rs863750	1.00	-0.0112	1.12E-11	-0.27	2.38E-59	*CCDC92*	3	7
BMI	rs7523023	0.96	-0.0115	1.29E-09	-0.12	3.47E-59	*EDEM3*	1	2
BMI	rs62491456	0.96	0.0197	1.29E-09	0.21	1.45E-43	*JHDM1D*	1	3
BMI	rs849135	1.00	0.0106	4.74E-11	0.20	1.25E-40	*JAZF1*	1	3
BMI	rs17175624	0.99	0.0118	1.37E-09	0.16	3.18E-38	*FGFR1*	1	3
BMI	rs17207196	0.96	-0.022	1.58E-36	0.33	3.52E-36	*PMS2P3*	1	3
BMI	rs2245368	0.98	-0.0238	1.72E-25	-0.36	3.70E-29	*UPK3BP1*	1	3
BMI	rs2966431	0.96	0.0161	1.27E-09	0.26	4.64E-22	*AC092171.4*	2	4
BMI	rs10497870	0.97	0.0121	1.97E-13	-0.10	1.20E-20	*CARF*	1	2
BMI	rs12282785	0.98	-0.0157	1.43E-11	0.21	2.61E-20	*RP11-21L23.2*	1	3
BMI	rs2516739	1.00	-0.0153	4.38E-14	0.10	4.20E-18	*TSC2*	2	3
BMI	rs3738476	0.97	-0.0195	3.35E-14	-0.17	4.43E-17	*PRUNE*	1	2
BMI	rs799451	0.99	0.0129	1.03E-13	0.08	4.61E-17	*ZMIZ2*	2	2
BMI	rs6050446	0.97	-0.0346	3.17E-13	0.22	8.42E-17	*ENTPD6*	3	9
BMI	rs4889606	0.95	0.0209	2.90E-36	-0.13	6.05E-16	*RP11-196G11.1*	3	4
BMI	rs11781016	0.96	0.0104	7.82E-10	0.06	1.50E-15	*CPNE3*	2	3
BMI	rs7133378	1.00	0.0127	2.93E-13	0.28	2.42E-15	*DNAH10*	1	4
BMI	rs7133378	1.00	0.0127	2.93E-13	0.21	3.68E-15	*RP11-380L11.3*	2	3
BMI	rs4616635	0.99	-0.0106	2.20E-09	0.14	5.90E-15	*ADAMTS9*	1	2
BMI	rs6512302	1.00	0.0134	1.52E-11	0.05	2.19E-13	*TCEA2*	2	2
BMI	rs905938	0.97	-0.0143	1.71E-14	-0.10	1.56E-12	*ZBTB7B*	2	2
BMI	rs17207196	0.97	-0.022	1.58E-36	0.19	2.79E-11	*TRIM73*	1	4
BMI	rs5771118	1.00	-0.013	2.56E-09	-0.09	3.66E-11	*PLXNB2*	2	4
BMI	rs3923783	0.99	-0.0222	4.12E-23	0.14	4.56E-11	*RTN4RL1*	2	4
BMI	rs73683491	0.98	0.0214	1.59E-10	0.13	3.02E-10	*SKAP2*	2	2
BMI	rs2710323	0.98	-0.0141	3.65E-18	-0.07	4.75E-10	*STAB1*	2	2
BMI	rs10773049	1.00	-0.0113	9.36E-12	0.11	8.22E-10	*FAM101A*	2	2
BMI	rs11236924	0.98	-0.0157	1.70E-11	0.13	2.23E-09	*TSKU*	3	3
BMI	rs2297674	0.96	0.0123	4.93E-12	0.08	2.98E-09	*COL16A1*	2	4

Table 3.7.1 continued from previous page

Trait	GWAS SNP	PP.H4.abf	GWAS Beta	GWAS p	eQTL beta	eQTL p	Symbol	eQTL_rank	Total eQTL
BMI	rs3808477	0.98	-0.0182	8.73E-22	0.09	4.80E-09	*TRPS1*	2	2
BMI	rs998584	1.00	-0.013	8.10E-15	-0.06	6.07E-09	*VEGFA*	3	3
BMI	rs9818870	0.98	-0.0137	1.14E-09	-0.08	2.22E-08	*MRAS*	2	3
BMI	rs3750944	1.00	-0.0103	4.74E-09	0.08	4.39E-08	*FAM160A2*	2	2
BMI	rs11227313	0.99	0.0188	4.45E-21	0.08	7.58E-08	*AP001266.1*	2	3
BMI	rs10773049	0.95	-0.0113	9.36E-12	0.14	8.65E-08	*RP11-380L11.3*	3	3
BMI	rs1652376	1.00	-0.0203	6.27E-33	0.12	1.11E-07	*NPC1*	3	3
BMI	rs12282785	0.95	-0.0157	1.43E-11	0.14	1.16E-07	*RP11-21L23.3*	2	2
BMI	rs3782224	0.99	-0.0115	4.24E-10	0.16	1.55E-07	*CHFR*	2	2
BMI	rs7138803	0.95	0.0297	3.10E-71	-0.10	5.53E-07	*FAIM2*	3	3
BMI	rs905938	0.97	-0.0143	1.71E-14	-0.13	7.94E-07	*DCST1*	2	2
BMI	rs761423	0.96	0.0105	6.51E-10	-0.14	1.45E-06	*MST1L*	1	9
BMI	rs879620	0.98	0.0226	8.45E-39	-0.12	4.08E-06	*RP11-462G12.1*	2	3
BMI	rs1000940	0.97	-0.0152	7.80E-18	0.08	5.22E-06	*SCIMP*	3	3
BMI	rs6451675	0.98	-0.014	2.54E-14	-0.12	6.04E-06	*NIM1*	2	2
BMI	rs6720868	0.95	0.0154	1.82E-17	-0.04	1.79E-05	*TRIP12*	2	2
BMI	rs17207196	0.97	-0.022	1.58E-36	-0.06	3.60E-05	*NCF1*	3	3
WHR	rs34696	0.96	-0.0112	3.97E-10	0.68	5.72E-159	*MAST4-AS1*	1	1
WHR	rs2232015	0.97	-0.0115	9.59E-10	-0.46	2.07E-144	*PRMT6*	1	1
WHR	rs1051684	0.96	0.0136	1.04E-12	0.23	4.50E-61	*CD79B*	1	1
WHR	rs761391	0.95	-0.0114	4.69E-10	-0.29	2.94E-36	*RP11-132M7.3*	1	1
WHR	rs2280600	0.96	0.0152	4.32E-11	-0.12	2.71E-27	*CPSF4*	1	1
WHR	rs1784460	0.97	0.014	1.34E-14	-0.48	5.65E-26	*RP11-110I1.14*	1	1
WHR	rs17264866	0.99	-0.0145	7.78E-12	-0.15	1.45E-19	*RP11-69E11.8*	1	1
WHR	rs6983481	1.00	-0.0187	2.54E-19	-0.19	7.65E-16	*STC1*	1	1
WHR	rs12001634	0.98	-0.0121	3.33E-09	0.28	5.73E-14	*RP11-440G5.2*	1	1
WHR	rs36232	0.98	-0.0162	4.60E-13	-0.23	9.01E-14	*RAB26*	1	1
WHR	rs2236744	0.97	0.0111	4.46E-10	0.11	1.15E-11	*PCK1*	1	1
WHR	rs1534696	1.00	-0.0226	3.67E-39	0.06	4.91E-11	*HNRNPA2B1*	1	1
WHR	rs1360485	0.99	0.0152	2.01E-16	0.08	7.70E-11	*HMGB1*	1	1
WHR	rs34322	1.00	0.0104	4.60E-09	-0.06	1.31E-10	*CDKN1B*	1	1
WHR	rs62271373	1.00	0.032	7.35E-14	-0.17	2.26E-10	*TSC22D2*	1	1
WHR	rs2509967	0.96	-0.0154	1.02E-16	-0.08	1.63E-09	*MTA2*	1	1
WHR	rs2301453	0.97	-0.0224	2.21E-40	0.08	2.68E-08	*DNM3OS*	1	1
WHR	rs979012	0.99	0.0116	7.86E-11	0.11	1.72E-07	*BMP2*	1	1
WHR	rs11150580	0.96	0.0172	6.46E-21	-0.03	2.53E-07	*TMEM219*	1	1
WHR	rs9905140	0.98	0.0114	3.87E-11	0.08	1.47E-05	*C17orf82*	1	1
WHR	rs1534696	1.00	-0.0226	3.67E-39	0.86	2.43E-256	*SNX10*	2	2
WHR	rs56271783	1.00	0.0441	8.85E-20	-0.71	5.27E-228	*VEGFB*	1	2
WHR	rs72868727	0.95	0.025	6.29E-12	-0.76	6.68E-130	*RNF157*	1	4
WHR	rs8103017	1.00	-0.0165	2.24E-14	0.36	6.71E-128	*SSC5D*	1	2
WHR	rs1789882	1.00	0.016	9.97E-12	0.51	2.00E-127	*ADH1A*	1	3
WHR	rs11051005	0.98	0.013	2.61E-11	-0.31	2.70E-105	*IPO8*	1	2
WHR	rs72801474	1.00	-0.0262	3.03E-14	-0.58	3.87E-97	*HSPA4*	1	2
WHR	rs55747707	1.00	-0.0154	3.40E-11	0.42	9.30E-92	*MLXIPL*	1	2
WHR	rs863750	1.00	0.0259	6.20E-51	-0.28	1.44E-90	*ZNF664*	1	8
WHR	rs13333747	0.97	0.018	1.42E-12	-0.25	3.08E-84	*PKD1*	1	2
WHR	rs805770	1.00	0.0178	3.42E-22	0.23	9.94E-66	*GPCPD1*	1	3
WHR	rs4444402	0.96	-0.0131	5.19E-10	0.24	1.17E-61	*LRRC45*	1	2
WHR	rs11150580	0.97	0.0172	6.46E-21	-0.15	3.83E-61	*DOC2A*	1	2
WHR	rs863750	1.00	0.0259	6.20E-51	-0.27	2.38E-59	*CCDC92*	3	7
WHR	rs2167750	0.98	0.0199	6.47E-29	0.20	6.91E-58	*FAM13A*	2	3
WHR	rs1534696	1.00	-0.0226	3.67E-39	0.21	3.01E-57	*CBX3*	1	3
WHR	rs72959041	1.00	0.126	4.56E-183	0.71	2.10E-54	*RSPO3*	1	2
WHR	rs1482852	0.99	0.0269	2.07E-53	0.27	7.51E-54	*LINC00886*	4	4
WHR	rs2925979	1.00	0.0215	1.62E-31	-0.22	1.81E-51	*CMIP*	1	3
WHR	rs13316065	0.99	0.017	1.34E-20	-0.29	7.78E-47	*RNF123*	1	2
WHR	rs28451064	1.00	0.0177	4.23E-09	-0.38	1.28E-41	*LINC00310*	1	2
WHR	rs4883198	0.99	-0.0136	1.80E-11	-0.33	1.68E-41	*PHC1*	1	3
WHR	rs4738141	0.97	-0.02	7.34E-25	0.45	7.38E-41	*EYA1*	1	3

Table 3.7.1 continued from previous page

Trait	GWAS SNP	PP.H4.abf	GWAS Beta	GWAS p	eQTL beta	eQTL p	Symbol	eQTL_rank	Total eQTL
WHR	rs17175624	0.98	0.0143	9.12E-13	0.16	3.18E-38	*FGFR1*	1	3
WHR	rs1534696	1.00	-0.0226	3.67E-39	0.34	2.32E-37	*AC004540.4*	1	3
WHR	rs11664106	1.00	0.02	8.69E-22	0.13	3.81E-33	*EMILIN2*	1	2
WHR	rs15285	0.96	-0.0117	1.62E-09	0.15	3.55E-32	*LPL*	1	6
WHR	rs10891539	0.96	-0.0111	4.44E-10	0.25	1.28E-27	*ANKK1*	2	2
WHR	rs2167750	0.99	0.0199	6.47E-29	0.13	3.87E-25	*FAM13A-AS1*	2	3
WHR	rs3776706	0.95	0.0155	3.27E-14	-0.20	1.57E-24	*FST*	1	2
WHR	rs4727695	0.97	0.0215	1.85E-13	0.28	1.72E-24	*LAMB1*	2	2
WHR	rs28451064	1.00	0.0177	4.23E-09	-0.24	2.36E-21	*MRPS6*	1	3
WHR	rs3768321	1.00	0.0163	5.53E-11	-0.25	2.45E-21	*PABPC4*	1	4
WHR	rs3128759	0.98	0.0223	1.39E-28	0.52	2.47E-21	*AL645922.1*	1	3
WHR	rs7350438	0.98	-0.0111	1.66E-09	0.17	2.06E-20	*NET1*	1	3
WHR	rs2745353	0.99	0.0349	2.84E-95	0.19	2.28E-19	*RSPO3*	2	2
WHR	rs501351	0.97	-0.0152	4.87E-14	0.21	6.52E-18	*SLC37A4*	3	5
WHR	rs4738141	0.99	-0.02	7.34E-25	0.36	6.75E-17	*RP11-1102P16.1*	1	2
WHR	rs1534696	1.00	-0.0226	3.67E-39	0.20	2.68E-16	*AC004540.5*	5	5
WHR	rs650558	1.00	0.0177	9.91E-15	0.24	3.12E-16	*HSD17B1P1*	2	2
WHR	rs28451064	0.99	0.0177	4.23E-09	-0.23	3.22E-16	*AP000320.7*	1	2
WHR	rs1752169	0.98	0.0123	6.82E-10	0.13	4.92E-16	*DENND1A*	1	2
WHR	rs2294239	0.99	0.02	3.17E-31	0.12	9.68E-16	*ZNRF3*	1	2
WHR	rs11051005	0.95	0.013	2.61E-11	0.27	2.07E-15	*RP11-77I22.2*	1	2
WHR	rs4077074	0.97	0.0132	5.00E-11	0.08	3.57E-15	*RAC3*	2	3
WHR	rs56113850	0.99	-0.0119	1.68E-09	0.23	4.82E-15	*CYP2A6*	1	4
WHR	rs4616635	0.99	0.0279	6.68E-52	0.14	5.90E-15	*ADAMTS9*	1	2
WHR	rs11208660	1.00	0.023	2.64E-14	-0.20	4.33E-14	*LEPR*	4	6
WHR	rs1482852	0.97	0.0269	2.07E-53	0.10	7.07E-14	*SSR3*	2	3
WHR	rs905938	1.00	0.0126	7.83E-11	-0.17	1.62E-13	*RP11-307C12.11*	3	3
WHR	rs905938	1.00	0.0126	7.83E-11	-0.10	1.56E-12	*ZBTB7B*	2	2
WHR	rs12437696	0.96	0.0118	2.07E-09	0.20	9.71E-12	*C15orf57*	3	3
WHR	rs805770	1.00	0.0178	3.42E-22	0.14	4.98E-11	*C20orf196*	3	4
WHR	rs11187537	0.99	0.013	1.73E-10	0.13	7.53E-11	*FFAR4*	2	2
WHR	rs59043281	1.00	-0.019	5.88E-15	-0.13	2.33E-10	*MIR4435-1HG*	2	2
WHR	rs3094621	0.98	0.0194	4.80E-13	-0.31	3.46E-10	*HLA-G*	5	5
WHR	rs10773049	1.00	0.0259	7.72E-51	0.11	8.22E-10	*FAM101A*	2	2
WHR	rs4894803	1.00	0.0149	3.74E-16	0.05	1.28E-09	*FNDC3B*	3	3
WHR	rs10891290	0.99	0.0199	2.05E-27	-0.18	1.30E-09	*PPP2R1B*	2	5
WHR	rs1572993	0.98	0.0128	7.19E-13	-0.09	1.71E-09	*NUAK2*	2	2
WHR	rs9897538	0.97	0.0161	1.53E-19	0.11	1.92E-09	*KCNJ2*	1	2
WHR	rs3814614	0.99	0.0108	9.93E-10	-0.11	2.07E-09	*GRID1*	2	2
WHR	rs11133377	0.99	0.0143	8.92E-13	0.07	2.48E-09	*CLOCK*	2	2
WHR	rs10184004	1.00	-0.0228	1.61E-40	-0.14	2.87E-09	*GRB14*	2	4
WHR	rs1345203	1.00	0.0187	3.07E-15	-0.13	3.54E-09	*BCL2L11*	2	2
WHR	rs7213608	1.00	-0.0171	6.95E-19	0.13	5.70E-09	*KCNJ12*	1	2
WHR	rs998584	1.00	0.0351	7.42E-92	-0.06	6.07E-09	*VEGFA*	3	3
WHR	rs1482852	0.99	0.0269	2.07E-53	0.15	1.00E-08	*LEKR1*	2	3
WHR	rs9400239	0.98	-0.0134	1.50E-13	0.18	1.64E-08	*LINC00222*	1	2
WHR	rs10096191	0.98	-0.0328	9.16E-23	0.33	1.92E-08	*EYA1*	3	3
WHR	rs2298632	1.00	-0.0144	3.49E-17	-0.07	2.53E-08	*ZNF436*	2	2
WHR	rs3775061	0.97	0.0128	2.04E-11	0.07	3.06E-08	*HTT*	2	3
WHR	rs1482852	0.99	0.0269	2.07E-53	0.07	4.31E-08	*RP11-305K5.1*	3	4
WHR	rs863750	0.97	0.0259	6.20E-51	-0.14	8.42E-08	*RP11-380L11.3*	3	3
WHR	rs4471313	0.99	0.0266	6.60E-43	-0.10	1.83E-07	*PRRX1*	3	3
WHR	rs6448429	0.99	0.0236	4.79E-19	0.08	2.33E-07	*RBPJ*	2	2
WHR	rs6800707	0.98	-0.0187	6.11E-16	-0.10	2.64E-07	*NISCH*	2	2
WHR	rs905938	0.99	0.0126	7.83E-11	-0.06	3.08E-07	*ADAM15*	2	2
WHR	rs2008514	0.96	0.017	8.56E-23	0.22	4.90E-07	*EIF3C*	1	10
WHR	rs2112347	0.99	0.015	8.41E-18	-0.12	5.32E-07	*POC5*	2	2
WHR	rs7138803	0.95	0.0125	1.09E-12	-0.10	5.53E-07	*FAIM2*	3	3
WHR	rs905938	1.00	0.0126	7.83E-11	-0.13	7.94E-07	*DCST1*	2	2
WHR	rs55747707	0.97	-0.0154	3.40E-11	0.11	5.81E-06	*STX1A*	2	2

Table 3.7.1 continued from previous page

Trait	GWAS SNP	PP.H4.abf	GWAS Beta	GWAS p	eQTL beta	eQTL p	Symbol	eQTL_rank	Total eQTL
WHR	rs62063287	0.96	-0.0242	2.79E-23	-0.09	6.26E-06	*NSF*	2	2
WHR	rs2301453	0.96	-0.0224	2.21E-40	-0.10	1.06E-05	*PIGC*	2	4
WHR	rs9370243	0.97	0.0198	1.13E-09	0.27	1.10E-05	*MLIP-IT1*	2	2
WHR	rs1482852	0.95	0.0269	2.07E-53	0.04	2.81E-05	*CCNL1*	2	2
WHRadjBMI	rs7740107	0.95	0.0177	3.78E-18	-0.42	2.14E-146	*L3MBTL3*	1	1
WHRadjBMI	rs1061093	0.99	0.02	7.79E-26	-0.24	9.36E-139	*EEF1G*	1	1
WHRadjBMI	rs1061093	0.99	0.02	7.79E-26	-0.25	3.19E-129	*MIR3654*	1	1
WHRadjBMI	rs1051684	0.99	0.0132	4.39E-12	0.23	4.50E-61	*CD79B*	1	1
WHRadjBMI	rs6446204	0.97	-0.0207	1.16E-24	-0.20	1.47E-57	*PRKAR2A*	1	1
WHRadjBMI	rs1547149	1.00	-0.0123	6.55E-11	-0.28	1.26E-32	*FGF9*	1	1
WHRadjBMI	rs4362930	0.98	-0.0137	1.36E-13	0.17	1.32E-27	*AC012613.2*	1	1
WHRadjBMI	rs17766692	0.99	0.0243	1.14E-09	-0.29	1.71E-26	*C19orf80*	1	1
WHRadjBMI	rs8887	1.00	-0.0109	3.22E-09	0.09	2.22E-26	*HDGFRP2*	1	1
WHRadjBMI	rs7975576	0.97	-0.0149	2.87E-13	-0.15	4.48E-26	*RIC8B*	1	1
WHRadjBMI	rs7928810	0.99	-0.0113	1.41E-10	0.18	6.14E-25	*NCR3LG1*	1	1
WHRadjBMI	rs2277339	1.00	-0.0224	2.35E-14	-0.30	3.85E-22	*PRIM1*	1	1
WHRadjBMI	rs889129	0.99	-0.0198	9.32E-12	0.18	8.54E-20	*CTD-2553C6.1*	1	1
WHRadjBMI	rs9644033	0.99	0.0222	3.40E-26	-0.19	2.04E-15	*STC1*	1	1
WHRadjBMI	rs6861681	0.95	0.0275	4.18E-49	0.28	1.34E-14	*NSG2*	1	1
WHRadjBMI	rs6861681	0.95	0.0275	4.18E-49	0.07	4.24E-14	*CPEB4*	1	1
WHRadjBMI	rs28610092	0.97	0.0158	1.62E-09	0.22	4.38E-13	*RAB26*	1	1
WHRadjBMI	rs12774134	0.95	-0.0188	1.99E-12	0.42	5.77E-13	*U8*	1	1
WHRadjBMI	rs12450700	0.99	0.0131	1.37E-12	-0.13	1.09E-11	*PITPNC1*	1	1
WHRadjBMI	rs9579574	0.99	-0.0115	3.64E-09	0.08	6.36E-11	*HMGB1*	1	1
WHRadjBMI	rs62271373	1.00	0.0408	3.44E-21	-0.17	2.26E-10	*TSC22D2*	1	1
WHRadjBMI	rs3989103	0.95	-0.0183	3.71E-18	-0.22	4.54E-10	*RP11-148O21.4*	1	1
WHRadjBMI	rs2509903	0.96	-0.0178	3.41E-13	0.10	5.09E-10	*ACVR1C*	1	1
WHRadjBMI	rs714515	0.97	-0.0276	6.49E-59	0.08	1.62E-08	*DNM3OS*	1	1
WHRadjBMI	rs668459	0.96	-0.0228	1.77E-39	0.10	4.46E-08	*CITED2*	1	1
WHRadjBMI	rs2165241	0.96	-0.0128	7.82E-13	0.07	1.56E-07	*LOXL1*	1	1
WHRadjBMI	rs2145272	0.99	-0.0247	3.95E-43	0.11	2.35E-07	*BMP2*	1	1
WHRadjBMI	rs757608	0.98	0.0194	1.10E-26	0.09	1.15E-05	*C17orf82*	1	1
WHRadjBMI	rs35565646	1.00	0.0181	7.17E-12	1.39	0	*DISP2*	1	3
WHRadjBMI	rs56271783	1.00	0.059	1.21E-33	-0.71	5.27E-228	*VEGFB*	1	2
WHRadjBMI	rs2250127	1.00	-0.0161	1.58E-14	-0.58	1.48E-147	*CTC-228N24.3*	1	2
WHRadjBMI	rs357438	1.00	-0.0114	3.79E-09	0.42	3.98E-147	*TRIM24*	1	2
WHRadjBMI	rs8103017	1.00	-0.0196	2.98E-19	0.36	6.71E-128	*SSC5D*	1	2
WHRadjBMI	rs11051005	0.98	0.0159	6.41E-16	-0.31	2.70E-105	*IPO8*	1	2
WHRadjBMI	rs12913300	0.99	0.0181	7.17E-12	1.03	3.16E-99	*RP11-64K12.4*	1	3
WHRadjBMI	rs72801474	1.00	-0.0295	2.37E-17	-0.58	3.87E-97	*HSPA4*	1	2
WHRadjBMI	rs55747707	0.99	-0.0244	3.05E-25	0.42	9.30E-92	*MLXIPL*	1	2
WHRadjBMI	rs863750	1.00	0.0373	4.17E-101	-0.28	1.44E-90	*ZNF664*	1	8
WHRadjBMI	rs4362930	0.98	-0.0137	1.36E-13	0.31	1.57E-87	*ABLIM3*	1	2
WHRadjBMI	rs2293413	0.97	0.0105	4.98E-09	-0.26	9.50E-84	*ITGA7*	1	2
WHRadjBMI	rs805770	1.00	0.0222	1.33E-33	0.23	9.94E-66	*GPCPD1*	1	3
WHRadjBMI	rs1045241	1.00	-0.0185	3.51E-22	0.24	1.24E-65	*TNFAIP8*	1	2
WHRadjBMI	rs4704389	0.99	0.0123	1.74E-12	-0.30	2.82E-61	*PDE8B*	1	5
WHRadjBMI	rs863750	1.00	0.0373	4.17E-101	-0.27	2.38E-59	*CCDC92*	3	7
WHRadjBMI	rs2167750	0.96	0.0274	5.53E-53	0.20	6.91E-58	*FAM13A*	2	3
WHRadjBMI	rs4974081	0.98	0.0206	1.33E-24	-0.23	3.06E-57	*QRICH1*	1	2
WHRadjBMI	rs900400	0.99	0.029	9.06E-57	0.28	1.89E-54	*LINC00886*	4	4
WHRadjBMI	rs72959041	1.00	0.1624	2.08E-293	0.71	2.10E-54	*RSPO3*	1	2
WHRadjBMI	rs2925979	1.00	0.0265	7.33E-46	-0.22	1.81E-51	*CMIP*	1	3
WHRadjBMI	rs28451064	1.00	0.018	2.87E-09	-0.38	1.28E-41	*LINC00310*	1	2
WHRadjBMI	rs4883198	0.99	-0.0182	4.32E-19	-0.33	1.68E-41	*PHC1*	1	3
WHRadjBMI	rs7102	1.00	-0.0115	1.38E-10	-0.27	3.39E-41	*CTD-3088G3.8*	1	4
WHRadjBMI	rs917191	1.00	0.0138	5.11E-14	0.15	9.78E-41	*SEMA3C*	1	2
WHRadjBMI	rs1057119	1.00	0.0134	9.17E-10	0.21	1.08E-36	*HOMEZ*	1	2
WHRadjBMI	rs8054299	0.96	0.0146	5.79E-14	-0.18	5.16E-36	*AKTIP*	1	2
WHRadjBMI	rs4285804	0.99	0.0109	3.29E-10	0.15	6.73E-36	*TRIM8*	1	2

Table 3.7.1 continued from previous page

Trait	GWAS SNP	PP.H4.abf	GWAS Beta	GWAS p	eQTL beta	eQTL p	Symbol	eQTL_rank	Total eQTL
WHRadjBMI	rs4704389	0.99	0.0123	1.74E-12	-0.20	9.96E-36	*WDR41*	2	8
WHRadjBMI	rs12913300	0.99	0.0181	7.17E-12	0.83	1.28E-35	*RP11-64K12.10*	1	2
WHRadjBMI	rs11664106	1.00	0.0282	5.90E-41	0.13	3.81E-33	*EMILIN2*	1	2
WHRadjBMI	rs13198178	1.00	0.0308	1.16E-16	-0.21	4.31E-29	*TFEB*	2	3
WHRadjBMI	rs78058190	1.00	0.0357	4.27E-12	0.77	1.31E-28	*WNT6*	1	2
WHRadjBMI	rs146182298	0.97	0.0314	1.45E-09	-0.35	2.16E-28	*NID2*	1	2
WHRadjBMI	rs7235010	0.98	0.0186	5.65E-19	-0.22	1.32E-27	*CABLES1*	1	3
WHRadjBMI	rs2167750	0.97	0.0274	5.53E-53	0.13	3.87E-25	*FAM13A-AS1*	2	3
WHRadjBMI	rs11766345	0.98	-0.0309	1.49E-23	0.29	6.45E-25	*LAMB1*	2	2
WHRadjBMI	rs7726234	0.98	0.0107	3.16E-09	-0.18	3.29E-24	*REEP2*	2	2
WHRadjBMI	rs11636147	0.95	-0.0149	1.59E-10	-0.24	4.91E-23	*IVD*	4	6
WHRadjBMI	rs28451064	1.00	0.018	2.87E-09	-0.24	2.36E-21	*MRPS6*	1	3
WHRadjBMI	rs9435732	0.99	-0.0128	2.57E-11	-0.19	3.96E-21	*MFAP2*	1	2
WHRadjBMI	rs351385	0.99	0.0133	2.40E-13	0.13	4.67E-20	*PPP2R5A*	1	2
WHRadjBMI	rs1936807	0.99	-0.0408	5.39E-126	0.19	3.89E-19	*RSPO3*	2	2
WHRadjBMI	rs12325636	0.97	-0.0123	4.56E-09	0.17	4.44E-17	*ZNF200*	1	2
WHRadjBMI	rs12679556	0.99	-0.0254	6.06E-39	0.36	8.04E-17	*RP11-1102P16.1*	1	2
WHRadjBMI	rs28451064	0.99	0.018	2.87E-09	-0.23	3.22E-16	*AP000320.7*	1	2
WHRadjBMI	rs2294239	0.99	0.0243	4.04E-44	0.12	9.68E-16	*ZNRF3*	1	2
WHRadjBMI	rs11051005	0.95	0.0159	6.41E-16	0.27	2.07E-15	*RP11-77I22.2*	1	2
WHRadjBMI	rs4450871	1.00	0.0166	3.23E-18	0.13	1.05E-14	*MSX1*	1	4
WHRadjBMI	rs10049088	0.98	-0.029	1.45E-59	0.11	1.05E-14	*SSR3*	2	3
WHRadjBMI	rs2371767	0.98	-0.0402	1.00E-100	0.14	1.20E-14	*ADAMTS9*	1	2
WHRadjBMI	rs905938	1.00	0.0243	1.35E-35	-0.17	1.62E-13	*RP11-307C12.11*	3	3
WHRadjBMI	rs4704389	0.95	0.0123	1.74E-12	-0.15	2.99E-13	*ZBED3-AS1*	5	7
WHRadjBMI	rs2373078	1.00	0.0215	1.11E-13	0.18	4.81E-13	*AC007563.5*	1	4
WHRadjBMI	rs2373078	1.00	0.0215	1.11E-13	0.20	7.45E-13	*IGFBP5*	2	3
WHRadjBMI	rs905938	1.00	0.0243	1.35E-35	-0.10	1.56E-12	*ZBTB7B*	2	2
WHRadjBMI	rs13324341	1.00	0.0176	4.08E-14	0.17	4.60E-11	*MRAS*	3	3
WHRadjBMI	rs11187533	0.99	0.016	5.34E-15	0.13	7.42E-11	*FFAR4*	2	2
WHRadjBMI	rs7246865	0.96	0.0142	2.86E-13	-0.06	9.27E-11	*MYO9B*	2	2
WHRadjBMI	rs10418336	0.97	-0.0327	7.45E-13	-0.19	1.03E-10	*NFIX*	1	2
WHRadjBMI	rs3094621	0.99	0.0292	6.35E-27	-0.31	3.46E-10	*HLA-G*	5	5
WHRadjBMI	rs863750	1.00	0.0373	4.17E-101	-0.11	8.40E-10	*FAM101A*	2	2
WHRadjBMI	rs2747398	0.99	-0.0187	2.78E-25	-0.09	9.33E-10	*TSHZ2*	3	4
WHRadjBMI	rs1345203	1.00	0.0264	2.23E-28	-0.12	1.08E-09	*MIR4435-1HG*	2	2
WHRadjBMI	rs10502148	1.00	-0.023	5.87E-36	-0.18	1.19E-09	*PPP2R1B*	2	5
WHRadjBMI	rs4894803	1.00	0.0149	6.17E-16	0.05	1.28E-09	*FNDC3B*	3	3
WHRadjBMI	rs2276824	0.99	0.0222	8.90E-38	-0.07	1.33E-09	*STAB1*	2	2
WHRadjBMI	rs1396514	0.97	-0.0226	1.01E-39	0.11	2.16E-09	*KCNJ2*	1	2
WHRadjBMI	rs1128249	1.00	-0.0324	2.45E-77	-0.14	2.62E-09	*GRB14*	2	4
WHRadjBMI	rs2343813	0.98	0.0192	8.62E-10	-0.09	3.08E-09	*NDST1*	3	3
WHRadjBMI	rs1345203	1.00	0.0264	2.23E-28	-0.13	3.54E-09	*BCL2L11*	2	2
WHRadjBMI	rs7213608	1.00	-0.0115	3.15E-09	0.13	5.70E-09	*KCNJ12*	1	2
WHRadjBMI	rs55779591	0.99	-0.014	3.97E-11	0.12	5.76E-09	*NBPF1*	4	10
WHRadjBMI	rs998584	1.00	0.0487	1.22E-170	-0.06	6.07E-09	*VEGFA*	3	3
WHRadjBMI	rs10049088	0.98	-0.029	1.45E-59	0.16	7.63E-09	*LEKR1*	2	3
WHRadjBMI	rs910382	0.99	-0.0187	2.76E-25	-0.11	9.99E-09	*RP4-678D15.1*	2	3
WHRadjBMI	rs10100423	0.99	-0.0418	5.91E-34	0.32	1.28E-08	*EYA1*	3	3
WHRadjBMI	rs7678138	0.95	0.0181	1.74E-09	0.08	1.59E-08	*USP53*	3	3
WHRadjBMI	rs9848655	0.98	0.0187	1.61E-15	-0.08	1.89E-08	*MRAS*	2	3
WHRadjBMI	rs2298632	1.00	-0.0155	4.96E-19	-0.07	2.53E-08	*ZNF436*	2	2
WHRadjBMI	rs863750	0.97	0.0373	4.17E-101	-0.14	8.42E-08	*RP11-380L11.3*	3	3
WHRadjBMI	rs1883711	0.96	0.0493	1.31E-17	-0.18	8.89E-08	*MAFB*	2	2
WHRadjBMI	rs12774134	0.95	-0.0188	1.99E-12	-0.09	9.37E-08	*AKR1C1*	2	3
WHRadjBMI	rs1138714	0.97	0.0148	3.41E-16	0.06	1.01E-07	*TALDO1*	2	2
WHRadjBMI	rs8003238	0.99	0.0127	2.61E-12	-0.12	2.08E-07	*C14orf64*	2	3
WHRadjBMI	rs10919388	0.99	-0.0333	9.71E-66	-0.10	2.89E-07	*PRRX1*	3	3
WHRadjBMI	rs905938	0.99	0.0243	1.35E-35	-0.06	3.08E-07	*ADAM15*	2	2
WHRadjBMI	rs3851294	0.97	-0.0257	9.03E-17	0.19	3.70E-07	*TMCC2*	2	2

Table 3.7.1 continued from previous page

Trait	GWAS SNP	PP.H4.abf	GWAS Beta	GWAS p	eQTL beta	eQTL p	Symbol	eQTL_rank	Total eQTL
WHRadjBMI	rs910382	0.99	-0.0187	2.76E-25	-0.10	4.37E-07	*AL354993.1*	3	3
WHRadjBMI	rs1482853	0.95	-0.0291	9.25E-57	0.07	4.83E-07	*RP11-305K5.1*	3	4
WHRadjBMI	rs905938	1.00	0.0243	1.35E-35	-0.13	7.94E-07	*DCST1*	2	2
WHRadjBMI	rs10074193	0.99	-0.0146	2.33E-09	-0.10	8.57E-07	*C1QTNF3*	2	2
WHRadjBMI	rs2915407	0.97	0.0169	4.07E-13	0.09	1.65E-06	*RRAS2*	2	2
WHRadjBMI	rs62466318	0.98	-0.0236	7.98E-24	0.11	5.06E-06	*STX1A*	2	2
WHRadjBMI	rs9678859	0.98	0.0188	7.97E-16	-0.12	5.34E-06	*AFF3*	5	5
WHRadjBMI	rs750460	0.96	-0.013	4.81E-13	0.09	7.95E-06	*LOXL1-AS1*	3	3
WHRadjBMI	rs714515	0.96	-0.0276	6.49E-59	-0.10	8.28E-06	*PIGC*	2	4
WHRadjBMI	rs11231144	0.95	-0.02	4.25E-26	0.04	2.08E-05	*AHNAK*	3	3
WHRadjBMI	rs1482853	0.96	-0.0291	9.25E-57	0.04	2.67E-05	*CCNL1*	2	2
CAD	rs3918226	1.00	-0.133315	1.69E-09	-0.18	2.18E-12	*NOS3*	1	1
CAD	rs3918226	1.00	-0.133315	1.69E-09	-0.19	2.43E-09	*ATG9B*	1	1
CAD	rs28451064	1.00	-0.127571	1.33E-15	-0.38	1.28E-41	*LINC00310*	1	2
CAD	rs7528419	1.00	0.11453	1.97E-23	0.23	6.58E-27	*PSRC1*	2	2
CAD	rs28451064	1.00	-0.127571	1.33E-15	-0.24	2.36E-21	*MRPS6*	1	3
CAD	rs28451064	0.99	-0.127571	1.33E-15	-0.23	3.22E-16	*AP000320.7*	1	2
CAD	rs11065979	0.97	-0.068556	1.93E-10	-0.07	2.21E-05	*ACAD10*	2	2
f_glucose	NA	0.98	-0.025	3.08E-09	0.34	7.60E-42	*PACSIN3*	1	2
f_glucose	NA	0.97	-0.079	1.26E-68	0.13	4.07E-11	*SMCO4*	2	2

Chapter 4

Discussion and Future Directions

GWAS has led to the discovery of numerous genetic regions that contribute to the genetic predisposition to diseases. As of 2020 July, genetic variants associated with 4466 diseases or traits from 4054 research papers[18] have been collected in the NHGRI-EBI GWAS catalog. However, the growing list of disease-associated genetic variants is only the essential first step toward curing complex diseases by developing drugs and therapeutic treatments that target their genetic causes. More than 90% of the discovered GWAS variants lie in non-coding regions, making it challenging to identify the underlying genes and the mechanisms of action[186].

Type 2 diabetes (T2D) is a complex disease with high and increasing prevalence worldwide, presenting a tremendous clinical, economic, and social burden[56]. In chapter 2, I presented my work as part of the FUSION tissue biopsy study group, where I used multiple types of molecular profiling data to identify genes and DNA methylation (DNAme) sites that were potentially involved in T2D pathophysiology. Here, I review key findings from this project, and suggest directions for future research on T2D genetics and related molecular traits.

In chapter 2, I detected QTLs for mRNA and miRNA expression and DNA methylation levels and observed that the lead variants for mRNA and miRNA QTLs had lower minor allele frequencies than the total set of tested variants; this trend was more pronounced

for miRNAs. This observation indicated that purifying selection may act more strongly on the genetic variants influencing miRNA levels than mRNA levels. I also identified eight mRNAs and 116 DNAme sites in skeletal muscle tissue, 14 mRNAs and 105 DNAme sites in subcutaneous adipose tissue with QTLs that colocalized with T2D GWAS signals, suggesting which mRNAs and DNAme sites these GWAS signals may work through to affect susceptibility to T2D. For BMI, relative fat mass, waist, fasting serum insulin, HOMA, fasting serum C peptide, fasting serum C peptide 30min, significantly more mRNAs were associated in both muscle and adipose than expected by chance (Fisher's exact test $p \leq 1.04 \times 10^{-3}$). These mRNAs with significant associations in both tissues may be part of the biological processes causing or responding to physiological trait levels in the same or different cell types in both tissues.

We found evidence for *INHBB* underlying the T2D GWAS loci rs11688682 from both colocalization and physiological trait-mRNA association analyses. T2D risk allele rs11688682-G was associated with a higher expression level of *INHBB* and lower methylation levels of cg14231073 and cg15344192 in skeletal muscle and subcutaneous adipose tissues. *INHBB* expression level was positively correlated with insulin resistance indices in subcutaneous adipose tissue. The luciferase assay conducted by our collaborators also confirmed that T2D risk allele rs11688682-G increased transcriptional activity in preadipocytes and adipocytes. The multiple lines of evidence suggested that T2D risk allele rs11688682-G may confer disease risks by upregulating *INHBB* expression level.

4.0.1 Use of single-cell molecular profiling technology to study molecular and cellular mechanisms

Although leveraging bulk tissue molecular profiling data provided valuable insights into the molecular traits potentially involved in T2D pathophysiology, it did not allow us to distinguish between tissue-level and cell type-level regulatory mechanisms. Tissue biopsies contain a mixture of cell types, including constituent cell types of the tissue of interest and cells from the blood supply and adjacent tissues. The molecular profile of a bulk tissue biopsy is a snapshot of the aggregated effects over all cell types present. This cell-type heterogeneity involved in bulk data can obscure cell-type specific mechanisms

and confound bulk-level analyses. Therefore, while molecular profiling of skeletal muscle and subcutaneous adipose tissue biopsies provided valuable insights into molecular sites potentially involved in T2D pathophysiology, it did not allow us to distinguish between tissue-level and cell type-level regulatory mechanisms.

For example, the conditionally independent eQTLs we identified for a given gene may reflect multiple genetic loci regulating this gene in one cell type or reflect different genetic loci regulating this gene in different cell types in one tissue. To adjust for variability likely caused by cell-type heterogeneity and other technical confounding factors, a typical approach in the *cis*-QTL analysis is to adjust for a large number of inferred hidden covariates, such as PEER factors[77]. Hidden covariates capture sources of shared variation across many molecular traits, and therefore generally do not capture *cis*-QTLs, which typically affect only a small number of genes. Unlike *cis*-QTLs, physiological traits often have a widespread impact on molecular traits, and are likely to be captured by hidden factor covariates. Therefore, addressing the cell-type heterogeneity is more challenging for the physiological-molecular trait association analysis than *cis*-QTL analysis.

Another approach is to use cell-type deconvolution methods. Adjusting for the cell-type heterogeneity leveraging cell-type deconvolution methods helps ameliorate its impact on analyses . However, as we and others observed, the deconvolution performance depends heavily on the reference panel and is biased toward known cell types[187], and therefore provides a limited ability in distinguishing regulatory effects manifesting in different cell types.

The rapidly evolving single-cell transcriptomic and epigenomic sequencing technologies[188] make it possible to simultaneously estimate cellular composition and molecular trait level in a cell-type specific way. Analysis based on single-cell technology is better positioned to investigate genetic regulatory effects in specific cell types and compare between cell types. Single-cell molecular profiling technology is poised to become a mainstream approach to understand cell-type-specific molecular mechanisms in the next few years[189].

4.0.2 Experimental follow-ups are necessary to fully understand the biological mechanisms behind GWAS loci

Previous studies[29], [31], [157] and the present study showed that using statistical analysis to integrate QTL with GWAS findings was valuable in narrowing the search for effector molecular traits to a few likely candidates. However, the integration approach does not guarantee the nominated molecular traits to be part of the causal disease pathway. Experimental follow-ups to validate and detail the roles of the nominated molecular traits are necessary to fully understand the links from genotypes to phenotypes. Rapid technological advancements in this area provide many types of experiments to test certain hypotheses of interest. For example, allelic differences in transcriptional activity can be validated using a reporter assay. In chapter 2, our collaborator used reporter assay and confirmed the T2D risk allele rs11688682-G increased transcriptional activity in preadipocytes and adipocytes, providing additional evidence supporting that rs11688682-G may confer disease risks by upregulating *INHBB* expression level. Massive parallel reporter assays, which allow for evaluating the effects of thousands of variants in a single experiment, have been widely used in in-vitro expression systems[190], cell cultures[191], and live animals[192]. Allele-specific protein-DNA interactions can be validated using electrophoretic mobility shift assay, which also has the capability to be conducted in a high-throughput manner[193]. Genome editing technology such as CRISPR/Cas9 enables the precise manipulation of specific mutations. A particularly informative study design is to apply environmental perturbations to model cell lines or induced pluripotent stem cells (iPSCs) that have been edited with the CRISPR/Cas9 technology, followed by collecting multi-omic molecular profiling and measuring cell physiology at different time points[194], [195]. Such study design generates a dynamic landscape of the changes at the molecular and physiological levels[194]. In-depth knowledge acquired from biological experiments will greatly expedite translating the genetic associations into causative mechanisms for complex diseases and eventually inform therapeutic strategies.

In chapter 3, I generated a catalog of conditionally independent eQTLs associated with subcutaneous adipose tissue gene expression levels, using RNA-seq based gene expres-

sion profiling and genotype data from TwinsUK, METSIM, GTEx, and FUSION studies. Of the 19,108 genes present in all studies, individual studies identified exactly one eQTL for 34.4% – 39.1% of the genes and $\geq$ 2 eQTLs for 8.0% – 15.3% of the genes; meta-analysis identified exactly one eQTL for 33.7% of the genes and $\geq$ 2 eQTLs for 46.6% of the genes. The more powerful meta-analysis enabled the detection of many conditionally independent eQTLs, leading to a larger increase in the proportion of genes with $\geq$ 1 eQTL compared to the proportion of genes with one eQTL. Using this eQTL resource with conditional independent eQTLs per gene, we identified colocalization for 61 T2D loci, 115 BMI loci, 110 WHR loci, 132 WHRadjBMI loci, four CAD loci, and four fasting glucose loci, providing hypotheses for molecular mechanisms underlying these GWAS loci.

4.0.3 Dissecting both GWAS and eQTL associations to independent signals, instead of eQTL associations alone, may lead to a more powerful colocalization analysis

In the analysis of identifying GWAS loci that colocalized with secondary eQTLs of subcutaneous adipose tissue gene expression levels, we observed one instance suggesting that more colocalization may be identified if the GWAS signals had been dissected into conditionally independent signals. A WHRadjBMI locus near *ZNF664* had two distinct GWAS loci, the primary (rs863750) and secondary (rs7133378) GWAS signals[157]. The initial colocalization results using the marginal GWAS associations showed that rs863750 was colocalized with the primary eQTL of *ZNF664* and the secondary eQTLs of *CCDC92* and *FAM101A*; rs7133378 was colocalized with the eQTLs of *FAM101A*. Although the initial marginal GWAS colocalization results suggested insignificant probabilities for colocalization between the secondary GWAS signal (rs7133378) and other adjacent genes except for *FAM101A*, rs7133378 appeared to overlap with eQTLs of a few adjacent genes in the regional association plots. It was suggested that the power of colocalization tests with coloc2 would be improved for multi-signal GWAS loci if conditional summary statistics were supplied[157]. We experimented with this approach and observed high colocalization probabilities between rs7133378 and eQTLs of *ZNF664*, *DNAH10OS* and *RP11-214K3.24*. This result suggested that besides dissecting eQTLs into multiple signals, colo-

calization analysis would be more powerful if studies could dissect GWAS signals into multiple signals and provide conditional summary statistics to the scientific community.

4.0.4 **Lack of ethnic diversity in existing large-scale QTL studies**

We improved the power to identify eQTL in subcutaneous adipose tissue through the use of meta-analysis, but we did not improve the ancestry diversity of the eQTL resources. This lack of ancestry diversity issue was noticed earlier in phenotypic GWAS discoveries. For many years, non-European populations were heavily underrepresented in GWAS discoveries[196], [197], which contributed to the low performance of the predictive value of polygenic risk score (PRS)[198] and reduced the applicability of commercial genetic tests in non-European individuals[199]. Increased attention to this issue and international efforts to address it have led to the publication of more multi-ancestry GWAS studies[200]–[203]. Human tissue QTL studies, which emerged later than GWAS, however, have received less attention and efforts to improve population diversity. Most published QTL studies, especially the large-scale ones, are also heavily concentrated in the European population, leaving other ancestries underrepresented in the eQTL catalogs[204], [205]. For example, 85% of samples in GTEx were from European individuals. This lack of diversity impedes our ability to fully understand the genetic regulatory architecture of molecular traits. The limited population diversity of QTL discoveries also has direct consequences for colocalization analysis. Often, the only QTL studies available are those using samples of European ancestry, even when the GWAS discoveries were from a mixed ancestry or non-European population. In practice, the mismatch in LD structure between samples used for GWAS studies and QTL studies usually results in a diminished power to infer molecular traits that likely influence disease susceptibility[206], which in turn will hamper our ability to discover the biological underpinning for GWAS loci.

Appendix

A Versatile Toolkit for Molecular QTL Mapping and Meta-analysis at Scale

Corbin Quick[1]*, Li Guan[2i], Zilin Li[1], Xihao Li[1], Rounak Dey[1], Yaowu Liu[3], Laura Scott[4], Xihong Lin[1]*

[1]Harvard T. H. Chan School of Public Health, Department of Biostatistics

[2]University of Michigan, Department of Computational Medicine and Bioinformatics

[3]Southwestern University of Finance and Economics

[4]University of Michigan, Department of Biostatistics and Center for Statistical Genetics

*Contact: qcorbin@hsph.harvard.edu, xlin@hsph.harvard.edu

[i]This project is in revision for publication[150]. This is joint work with Dr. Corbin Quick, who conceived the software package. Dr. Xihong Lin, Dr. Laura Scott, and Dr. Corbin Quick and I helped conceptualize the framework and analysis. Dr. Corbin Quick, Dr. Zilin Li, Dr. Xihao Li, Dr. Rounak Dey, Dr. Yaowu Liu, and I contributed to software development, statistical methods, and/or data analysis. My specific contributions toward this project are developing statistical methods for conditional QTL meta-analysis with heterogeneous effects in collaboration with Dr. Corbin Quick, and implementing the algorithms for homogeneous-effect and heterogeneous-effect conditional QTL meta-analysis. I also helped with data processing and analysis: 1) I generated PEER factors and performed PEER factor optimization analysis for the Hapmap and Geuvadis datasets; 2) I performed the analysis of comparing the genetic constraint (LOEUF) between different groups of genes (Main figure 4C and 4D); 3) I performed the analysis of evaluating enrichment of GWAS signals in eQTLs. (Supplementary figures 11 and 12). I also contributed to drafting the sections relevant to the data analyses that I contributed.

Abstract

Molecular QTLs (xQTLs) are widely studied to identify functional variation and possible mechanisms underlying genetic associations with diseases. Larger xQTL sample sizes are critical to help identify causal variants, improve predictive models, and increase power to detect rare associations. This will require scalable and accurate methods for analysis of tens of thousands of molecular traits in large cohorts, and/or from summary statistics in meta-analysis, both of which are currently lacking. We developed APEX (All-in-one Package for Efficient Xqtl analysis), an efficient toolkit for xQTL mapping and meta-analysis that provides (a) highly optimized linear mixed models to account for relatedness and shared variation across molecular traits; (b) rapid factor analysis to infer latent technical and biological variables from molecular trait data; (c) fast and accurate trait-level omnibus tests that incorporate prior functional weights to increase statistical power; and (d) compact summary data files for flexible and accurate joint analysis of multiple variants (e.g., joint/conditional regression or Bayesian fine-mapping) without individual-level data in meta-analysis. We applied the methods to data from three LCL eQTL studies and the UK Biobank. APEX is open source: `https://corbinq.github.io/apex`.

Introduction

Human genetics studies have identified tens of thousands of molecular QTLs- genetic loci associated with differences in molecular quantitative traits- including mRNA (eQTL), microRNA (miQTL), or protein (pQTL) expression, metabolite (metQTL), methylation (mQTL) levels[207], [208]. By mapping DNA sequence variation to heritable differences in the transcriptome and epigenome, xQTL studies have provided important insights into genome function and gene regulation ([209]–[211]). xQTLs are also of interest in genome-wide association studies (GWAS) as possible biological antecedents of genetic associations with complex traits and diseases[212]–[215]. Integrative analyses of xQTL and GWAS data have provided insight into the biological mechanisms underlying GWAS associations, and helped identify causal disease genes and therapeutic targets[216]–[218].

Larger xQTL studies are crucial to identify causal variants driving xQTL association signals, detect low-frequency and rare xQTL variants, and more accurately predict expression or methylation levels from genotype data. The next generation of xQTL studies will require scalable methods for association analysis in large multi-ethnic cohorts, accurate methods for downstream statistical analysis (e.g., Bayesian finemapping and colocalization analysis) from summary statistics in meta-analysis, and integrative methods to utilize prior knowledge of genome function. We developed APEX, a toolkit for scalable xQTL association analysis and meta-analysis, to address these challenges.

Molecular trait data suffers from a high degree of technical and biological variation, which can both mask and confound *cis* and *trans* genetic associations[77], [219]–[221]. Latent variable models such as PEER[220] and dimension reduction techniques such as principal component analysis (PCA)[79], [222] are often used to infer unobserved common sources of technical and biological variation in xQTL studies. PEER is particularly effective in xQTL analysis, but computationally demanding. In APEX, we implemented simple, efficient algorithms for high-dimensional factor analysis using early stopping for regularization[223]. We found that this approach is nearly as fast as PCA and far faster than PEER, while yielding equal or greater numbers of *cis* discoveries than either method.

Linear mixed models (LMM) are widely used to account for population structure and cryptic familial relatedness in genome-wide association studies (GWAS), and can additionally account for shared technical and biological variation across molecular traits in xQTL studies[222]. However, despite multiple existing LMM methods for xQTL analysis[222], [224], ordinary least squares (OLS) is often used in practice for its greater computational efficiency. Even family-based eQTL studies often use a two-stage approach in which LMM residuals are used as response variables in OLS[45], [146], which may reduce statistical power[225]. In APEX, we developed efficient algorithms for LMM association analysis to account for population structure, relatedness, and technical variation with tens of thousands of traits, which are accurate for small samples and scale linearly in sample size.

Permutation tests are the current standard to calculate trait-level xQTL omnibus tests and account for correlations between tests statistics across variants and traits in xQTL discovery[31], [79], [151]. This approach is burdensome for large sample sizes, and does not readily capitalize prior knowledge of variant functionality. The aggregated Cauchy association test (ACAT) is a recently-developed method to combine p-values under arbitrary dependence structures[153], [154]. We applied ACAT to aggregate xQTL test statistics for each molecular trait, which scales linearly in the number of variants and independent of sample size. Unlike permutation tests, which implicitly assign equal prior weight to all variants, ACAT can incorporate functional prior weights between variants and molecular traits. We found that simply weighting by the chromosomal distance between each variant and transcription start site (TSS)[226] substantially increased xQTL discoveries.

While dozens of xQTL studies have been conducted[208], meta-analysis is hampered by difficulties sharing human genomic data. Marginal variant-trait associations can be meta-analyzed using regression slopes and standard errors or z-scores alone. However, these statistics are not sufficient for analyses that involve the joint effects of multiple variants, such as joint and conditional analysis[125], [227], Bayesian fine-mapping[82], [228]–[230], aggregation tests[227], [231], [232], and colocalization analysis[83]. Multiple-variant analysis further requires variance-covariance or linkage disequilibrium (LD) matrices, which characterize the joint distribution of single-variant xQTL association statistics. In GWAS,

proxy LD from a genotype reference panel is often used for multiple-variant analysis from summary statistics, but this is problematic for small or ancestrally heterogenous samples[125], [229], both of which are common in omics studies[31], [208], [209], [221]. Indeed, previous xQTL meta-analyses have generally analyzed only marginal variant-trait associations[54], [55], [233]. In APEX, we developed compact xQTL summary association data formats for accurate multiple-variant analysis in meta-analysis without individual-level data.

Results

Software development

We developed APEX (All-in-one Package for Efficient Xqtl analysis), a software toolkit for scalable xQTL mapping and meta-analysis. Core running modes for molecular trait preprocessing, *cis* and *trans* association analysis, and xQTL meta-analysis are summarized in Figure 1 (Figure5.0.1) (see Methods and Supplementary Materials for further details). APEX is a command-line tool implemented in C++, supports multi-threading to expedite linear algebra, and provides flexible sub-setting options to facilitate parallelization across genomic regions. It uses the Eigen[234] and Spectra[235] C++ libraries for linear algebra, and HTSlib to process indexed BED, BCF, and VCF files[236]. Precompiled Linux binaries and source code are available online.

Application to 3 lymphoblastoid cell line (LCL) eQTL data sets

We analyzed LCL eQTLs using genotype, expression, and technical covariate data from the GTEx project v8[31], Geuvadis project[211], and HapMap project[209], [237], [238](Table 1 (Table5.0.1)). GTEx (n = 147) and Geuvadis (n = 454) have RNA-seq LCL expression measurements and whole genome sequencing (WGS) based genotype calls. HapMap (n = 518) has array-based LCL expression measurements and array-based genotype calls, from which we imputed genotypes using the 1000 Genomes Project reference panel[239]. Data and processing procedures for each study are further described in Methods.

Rapid factor analysis of molecular traits for xQTL analysis

We inferred hidden covariates from gene expression measurements in each study using PEER[220], expression principal component (ePC) analysis[79], and expression factor analysis (eFA)[223]. For each method, we varied the number of hidden covariates from 1 to 100. eFA and PEER explicitly model shared and unique variances for each trait, whereas ePCs capture maximal variance across all traits[240]. Conceptually, ePC can be viewed as a special case of eFA in which all traits are assumed to have equal unique variance (unexplained by common factors). Further details are given in Methods and Supplementary Materials.

We used APEX to perform *cis*-eQTL analysis in each study modeling the hidden factor covariates as either fixed effects using ordinary least squares (OLS) or random effects using restricted maximum likelihood (REML)[219] (Figure 2 (Figure5.0.2)). ePC and eFA covariates were calculated directly in APEX, and PEER factors were calculated using the PEER R package[220]. For each method and data set, we varied the number of inferred covariates between 1 and 100. Across the studies, APEX eFA was 86 to 5033 times faster than PEER for models with >50 common factors (and 30 to 779 times faster for 20 to 50 factors), and provided equal or greater numbers of *cis* discoveries in each of the 3 data sets (Figure 2, panel A). Random-effect eFA provided the greatest number of discoveries in each of the 3 data sets, and fixed-effect or random-effect ePCs generally yielded the smallest numbers of discoveries.

To assess Type I error rates for fixed-effect and random-effect models with ePC or eFA covariates, we simulated 100 expression data sets under the null hypothesis in the Geuvadis study. We used the empirical covariance between expression and observed covariates (not inferred from expression) and empirical variance matrix of expression residuals (projecting out observed covariates) to simulate expression under the null hypothesis matching the observed covariance structure (Supplementary Figures 1-2). With each simulated expression matrix, we re-calculated the inferred covariates (eFA or ePC) and performed *cis*-eQTL analysis modeling the inferred covariates as either fixed or random effects. As-

sociation tests from all configurations (fixed-effect or random-effect models with between 1 and 100 inferred covariates) showed well-calibrated Type I error rates (Supplementary Figure 3 (Figure5.0.3)).

Fast, scalable linear mixed models with tens of thousands of molecular traits

We assessed the computational performance and numerical concordance of APEX and standard tools for linear mixed model (LMM) association analysis: FastGWA[241], BOLT-LMM[242], GMMAT[243], and GENESIS[244]. APEX uses a 3-stage approach to efficiently estimate LMM null models and association statistics with tens of thousands of traits (Supplementary Figure 4), whereas the other tools are intended for single-trait analysis. We note that each of these tools supports a variety of features not considered in our analysis here – for example, GMMAT and GENESIS support flexible generalized LMM (GLMM) for binary and other non-normal traits, and BOLT-LMM supports flexible variance partitioning. For LMM association analysis, FastGWA and BOLT-LMM use approximations for efficient analysis in large cohorts, which may be less accurate with smaller sample sizes (e.g., < 5000[245]. GENESIS, GMMAT, and APEX do not use such approximations, and APEX further uses small-sample LMM association tests (Supplementary Materials). To assess computational performance for LMM association analysis in large cohorts, we used genotype data and a sparse GRM for 10,000 individuals from the UK Biobank study, and simulated expression data for 16,329 traits with heritability drawn from a uniform distribution (Methods). Variant component estimates and single-variant association test statistics were nearly numerically equivalent between APEX, GMMAT, and GENESIS, as expected; FastGWA test statistics showed lower concordance with other methods (Supplementary Figure 5). LMM association analysis using APEX was >200-fold faster than GENESIS and GMMAT, 51.4-fold faster than BOLT-LMM, and 2.5-fold faster than FastGWA (Supplementary Table 1).

Powerful and efficient *cis*-xQTL omnibus tests

We performed single-variant and gene-level *cis*-eQTL analysis in each study using APEX, FastQTL, and QTLtools (Figure 3). APEX and FastQTL use multiple linear regression

(MLR) to adjust for covariates by default, whereas QTLtools uses simple linear regression with expression residuals (SLR-resid). We note that QTLtools can also perform MLR by regressing out covariates from genotype files prior to association analysis. Gene-level p-values from QTLtools and FastQTL use a Beta-approximated permutation test (Beta), whereas APEX uses either ACAT with constant weights (ACAT) or ACAT with distance-to-TSS weights between each variant and gene (ACAT-dTSS). FastQTL was run using adaptive p-values with 100 to 1000 permutations; QTLtools was run with 1000 permutations.

We compared the numbers of *cis*-eQTL discoveries at 1% false discover rate (FDR) in each study from Beta permutation tests using FastQTL[151] or QTLtools[79], and from ACAT[153], [154] using APEX (Figure 3 panel A). Each method calculates gene-level omnibus *cis*-eQTL p-values (*cis*-eGene p-values) based on single-variant association test statistics within a 1 megabase (Mbp) window of the transcription start site (TSS). QTLtools and FastQTL use permutation tests of the minimum p-value across variants, and expedite computation by modeling the null distribution as a beta density using a fixed number of permutations[151]. In each of the three studies, ACAT and permutation-based p-values were generally concordant (Supplementary Figure 6), but ACAT yielded more *cis*-eGene discoveries overall and was >30x faster (Figure 3, panels A and D). We also calculated weighted ACAT test statistics, in which each variant received a weight proportional to $e^{-\gamma|d|}$ where d is the number of base pairs between the variant and TSS and γ = 1e-5 (*30*). dTSS weighting further increased the number of *cis*-eGene discoveries by 14 to 30% across single studies (Figure 3, panel A).

We assessed p-value calibration for ACAT (implemented in APEX) and permutation-based p-values (implemented in FastQTL and QTLtools) by simulating expression data under the null hypothesis using genotype and expression data from the Geuvadis study (Figure 3 panel B). We used the sample covariance matrices of expression and observed covariates to simulate expression traits under the observed covariance structure (Methods). Empirical Type I error rates were well-controlled for both ACAT and Beta p-values, and SLR-resid p-values were conservative (shown previously in[246]. Permutation test p-values from

SLR-resid were also notably conservative, which is expected because while trait residuals and genotype residuals are orthogonal to covariates, permuted trait residuals and unadjusted genotypes are not.

Accurate multiple-variant xQTL meta-analysis from summary statistics

We assessed CPU time, memory, and storage required to create summary files for xQTL meta-analysis using APEX. We generated single-variant association summary statistics (sumstat files) and adjusted LD matrices (vcov files, which store the variance-covariance of association test statistics) for each of the 3 studies using APEX (Supplementary Figures 7-8). Summary statistics files were generated across all autosomes in 0.17 to 0.33 CPU hours and required 0.42 to 0.49 Gb storage per study (Supplementary Table 2). Adjusted LD files, which included LD for all pairs of variants within sliding 2 Mbp windows, were generated across all autosomes in 32.1 to 75.3 CPU hours and required 21.5, 34.3, 119.7 GB storage for GTEx, Geuvadis, and HapMap respectively (Supplementary Table 2). HapMap, which used imputed genotype dosages, required notably more time and storage than the other studies, which used WGS-based hard-call genotypes. We also compared adjusted LD storage using RareMetalWorker (RMW)[227], a tool for rare-variant association meta-analysis, across the 3 studies. APEX was 1.5 to 2.2-fold faster and required 4.5 to 21.5-fold less storage than RMW (Supplementary Table 3).

Score statistics and adjusted LD (stored in APEX sumstat and vcov files) are sufficient for a wide range of analyses involving the joint effects of multiple variants, including joint and conditional analysis, Bayesian finemapping, and penalized linear regression. We used APEX sumstat and vcov files from each LCL study to perform stepwise regression analysis using APEX-meta (Figure 4 (Figure5.0.4) and Supplementary Figure 9) and Bayesian finemapping using the *susieR* R package[228] (Figure 5 (Figure5.0.5)) in individual studies and meta-analysis. To assess the accuracy of summary-based analyses, we also performed these analyses from individual-level data. Stepwise regression slopes and p-values and fine-mapping posterior inclusion probabilities (PIPs) were nearly numerically equivalent between individual-level vs sumstat data (Pearson Rsq > 0.999; Figure 5 panel C).

To assess the accuracy of joint analysis from association summary statistics using proxy LD or unadjusted LD rather than APEX vcov files (which store adjusted LD), we performed finemapping with association summary statistics from HapMap and either (a) unadjusted LD (the correlation matrix of genotypes in HapMap, not adjusted for PCs and other covariates), or (b) proxy LD (adjusted LD from Geuvadis as a proxy for adjusted LD from HapMap). Unadjusted LD is often used for multiple-variant analysis from GWAS summary statistics (e.g., [125]), and differs from adjusted LD when genotypes are correlated with covariates (e.g., genotype PCs in multi-ethnic studies). This approach is closely related to simple linear regression with trait residuals (not adjusting genotypes for technical covariates in individual-level analysis). PIPs using proxy LD or unadjusted LD yielded substantially lower concordance with the exact PIPs that adjusted LD (Figure 5 panel C), which is expected due to the relatively small sample sizes and differences in ancestry composition between HapMap and Geuvadis. Notably, many other xQTL studies have relatively small sample size and heterogeneous ancestry composition (Supplementary Figure 10).

Functional characterization of LCL eQTL variants and genes

We hypothesized that mRNA expression heritability is lower for genes that are more evolutionarily constrained, and that therefore eGenes detected only in meta-analysis are more constrained on average than those detected in single studies. To assess this hypothesis, we compared the loss-of-function observed/expected upper bound fraction (LOEUF), a recently developed metric of genetic constraint (smaller LOEUF suggests greater constraint)[247], across genes that were tested in all 3 studies (11,750 genes). Novel LCL eGenes (eQTL associations detected by meta-analysis, but not by individual studies) and genes with no significant signal had significantly lower LOEUF than previously-identified eGenes (Mann–Whitney p = 2.1e-7 and 2.2e-16 respectively), while the difference in LOEUF was less pronounced for novel eGenes vs genes with no detected eQTLs (p = 0.0096) (Figure 4 panel C). Moreover, genes with larger numbers of significant *cis*-eQTL signals (identified in stepwise regression; Methods) tend to have larger LOEUF values (p < 2.2e-16) (Figure 4 panel D). While gene length is associated with LOEUF, we observed no significant trends between gene length and eQTL signals. These results sug-

gest that larger samples sizes will be required to detect xQTLs for more biologically important genes, highlighting the utility of meta-analysis.

We assessed functional enrichment of primary and secondary LCL eQTL variants identified in meta-analysis across the 3 studies. We used binomial logistic regression to identify features associated with LCL eQTL variants controlling for distance to nearest TSS and minor allele frequency (MAF) (Methods). First, we assessed enrichment of LCL eQTL variants in tissue-specific DNase I hypersensitive sites (DHSs) across 16 tissue groups[248]. LCL eQTLs showed striking enrichment in lymphoid-specific DHS compared to other tissue groups (Supplementary Figure 11 (5.0.6)). Next, we assessed overlap enrichment of LCL eQTL variants overlapping GWAS variants identified using the NHGRI-EBI GWAS Catalog[18]. Among 15 categories of GWAS traits, LCL eQTL variants showed strongest enrichment with GWAS variants for immune diseases (Supplementary Figure 12 (5.0.7)). These results suggest that LCL eQTL variants capture cell-type specific functionality, and highlight the utility of xQTL analysis in diverse tissues and cell types.

Discussion

Future xQTL studies will be conducted in increasingly large and diverse cohorts, and are poised to capitalize on growing knowledge of functional elements in the human genome. We developed APEX to empower these studies by providing a flexible, scalable framework for *cis* and *trans* xQTL analysis and meta-analysis. APEX provides rapid high-dimensional factor analysis to infer latent technical and biological factors, efficient linear mixed model (LMM) association analysis for *cis* and *trans* xQTL mapping, procedures to incorporate prior weights in primary and secondary xQTL signal discovery, and a framework for accurate joint analysis of multiple variant effects from xQTL summary data.

Our LMM framework for molecular traits extends upon previous work[219], [224] by optimizing association analysis with high-dimensional traits and structured random-effect covariance matrices. In particular, we precompute and recycle computationally expensive terms for each molecular trait and each variant, and exploit the structure of random-effect covariance matrices (low-rank or block-diagonal) to expedite linear algebra. With these

optimizations, LMM association analysis scales linearly in sample size and the number of traits, enabling rapid analysis with large xQTL cohorts. APEX also uses small-sample adjustment and avoids large-sample approximations to provide accurate p-values for smaller cohorts.

In GWAS, random effects are typically used to account for infinitesimal genetic effects or familial relatedness in LMM association analysis. In xQTL studies, random effects can also be used to model shared technical and biological variation across traits[219], [224]. Our results suggest that this strategy outperforms ordinary least squares (OLS) when using expression factor analysis covariates, but underperforms OLS when using expression PC covariates. A variety of other methods can be applied to infer hidden covariates from molecular trait data, and various other strategies (e.g., penalized regression) can be used to include these covariates in xQTL analysis. We believe this is a worthy area for further research. Here, our work provides rapid inference of latent technical and biological covariates from molecular trait data, and a flexible LMM framework to include these covariates as fixed or random effects in xQTL association analysis.

Our meta-analysis framework extends from previous eQTL meta-analysis tools[249] by enabling accurate multiple-variant analysis, including joint/conditional analysis (using APEX mode meta), Bayesian fine-mapping (using *susieR*[228] or DAP[250]), and colocalization analysis (using external software), from xQTL summary statistics. These methods are fundamental in a variety of applications, including predictive weight estimation (e.g., for TWAS) and integrative analysis of GWAS loci. Methods that use LD from a reference panel as a proxy for meta-analysis LD may be inaccurate when reference or meta-analysis sample size is limited (e.g., < 5000), ancestry composition differs between reference vs meta-analysis samples, or genotypes are correlated with covariates in meta-analysis. In APEX, we provide exact study-specific adjusted LD matrices (vcov files); similar strategies have been used for rare-variant association meta-analysis[227], [231], but not to our knowledge for genome-wide xQTL or fine-mapping meta-analysis. The proposed xQTL meta-analysis framework enables flexible and highly accurate multiple-variant modeling with arbitrary sample sizes, ancestry compositions, and sets of covariates.

While our applications focused on eQTL studies, APEX sumstat and vcov formats are also well-suited for GWAS of quantitative traits, which can be used, for example, in colocalization analysis of GWAS and xQTL signals. More broadly, we encourage GWAS and xQTL studies to publicly release adjusted LD data in addition to single-variant association summary statistics when possible. With streamlined tools for the analysis of such data, greater availability of sufficient statistics including LD would increase reproducibility, enhance meta-analysis, and accelerate discovery.

The statistical methods in APEX can be extended in a variety of ways, such as by (a) leveraging correlations between molecular traits across multiple tissues or cell-types, (b) modeling genetic correlations between traits of the same tissue or cell-type, or (c) supporting generalized linear models for non-normal traits. Multivariate LMMs can be used to account for the correlation structure of genetic and environmental components of molecular traits across and within tissues or cell-types. Also, zero-inflated Poisson or negative binomial generalized linear mixed models (GLMMs) may be desirable for some types of molecular trait data.

Our data applications have several limitations, including (a) analysis of only LCL eQTLs, (b) relatively small eQTL sample sizes, and (c) limited *trans*-eQTL analysis. Our LCL eQTL analysis revealed striking enrichment with relevant tissue-specific DHS, highlighting the utility of xQTL analysis across diverse tissues and cell types. Moreover, APEX is well suited for analysis of mRNA expression and other molecular traits across broader sets of tissues or cell types due to its computational efficiency. While the three LCL eQTL had limited sample sizes, our simulation studies using UK Biobank genotype data demonstrated that APEX is scalable to larger cohorts, with > 100-fold improvement in CPU time relative to standard tools. Finally, we note that APEX fully supports *trans*-eQTL analysis, as illustrated in simulation studies.

In summary, APEX provides an efficient and comprehensive framework for *cis* and *trans* xQTL mapping and meta-analysis. For xQTL studies of a single cohort, APEX provides efficient inference of latent technical and biological factors from molecular trait data[223],

which performs competitively with state-of-the-art methods in *cis*-eQTL analysis and orders of magnitude faster; rapid LMM association analysis with tens of thousands of molecular traits; powerful, efficient trait-level xQTL omnibus tests; and accurate multiple-variant analysis. For xQTL meta-analysis, APEX provides accurate single-variant and joint multiple-variant regression analysis, and compact summary data formats for flexible and accurate multiple-variant modeling (e.g., Bayesian finemapping) without individual-level data across multiple studies.

Online Methods

Statistical methods implemented in APEX

Principal components and factor analysis of molecular traits

APEX provides efficient algorithms to calculate principal components (PCA) and factor analysis (FA) of molecular traits. For PCA, we calculate k PC covariates as the first k left singular vectors of the truncated singular value decomposition (SVD) of the $n \times p$ normalized expression matrix $\mathbf{Y}$, which is scaled and centered so that each column (trait) has mean 0 and variance 1. The SVD is $\mathbf{Y} = \mathbf{UDV}^\top$, and $\mathbf{U}_{(k)} = (\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_k)$ are the PC covariates. When the number of traits is larger than the number of samples, we calculate $\mathbf{U}_{(k)}$ from the truncated SVD (or eigendecomposition) of $\mathbf{YY}^\top$, as $\mathbf{YY}^\top = \mathbf{UD}^2\mathbf{U}^\top$. Otherwise, we calculate $\mathbf{U}_{(k)} = \mathbf{YV}_{(k)}\mathbf{D}_{(k)}^{-1}$, where the right singular vectors $\mathbf{V}_{(k)}$ are calculated from $\mathbf{Y}^\top\mathbf{Y} = \mathbf{VD}^2\mathbf{V}^\top$.

The FA model is $\mathbf{Y}=\mathbf{ZB}+\mathbf{E}$ where $\mathbf{Z}$ is the $n \times k$ matrix of common factors, $\mathbf{B}$ is the $k \times p$ matrix of factor loadings, and $\mathbf{E}$ is the $n \times p$ matrix of unique factors. The rows of $\mathbf{E}$ are independent, and each row vector is multivariate normal with covariance matrix $\Sigma = diag(\sigma_1^2, \ldots, \sigma_p^2)$. In APEX, we estimate the common factors $\mathbf{Z}$ using an SVD of $\mathbf{Y}\hat{\Sigma}^{-1/2}$, which we initialize with constant variances $\hat{\sigma}_j^2 = 1$ for all $j = 1, 2, \ldots, p$. Given the first k left singular vectors $\tilde{\mathbf{U}}_{(k)}$ of $\mathbf{Y}\hat{\sigma}^{-1/2}$, we update the estimates as $\hat{\sigma}_j^2 = \frac{1}{n-1}\left\|(I - \tilde{\mathbf{U}}_{(k)}\tilde{\mathbf{U}}_{(k)}^\top)\mathbf{Y}_j\right\|_2^2$ for each trait $j = 1, 2, \ldots, p$, and repeat. A similar algorithm was suggested by (*62*), but the underlying likelihood is unbounded if $\hat{\sigma}_j^{-1} \to 0$, and convergence generally fails in practice. As proposed by[223], we perform regularization by halting after a fixed number of

iterations. If the number of samples is greater than the number of traits ($n > p$), we modify this approach using the $p \times k$ right singular vectors rather than the $n \times k$ left singular vectors of $\mathbf{Y}\hat{\Sigma}^{-1/2}$. The time complexity of this procedure is $O(\min(n,p)^2 k + pnk)$, where n is the sample size, p is the number of traits, and k is the number of factors. Further details are given in Supplementary Materials.

Statistical methods for *cis* and *trans* LMM association analysis

APEX provides a scalable linear mixed model (LMM) framework to account for familial relatedness[219], [251] or technical variation[222], [224] (Supplementary Figure 4). For traits t = 1, 2, …, p, we assume the model

$$\mathbf{Y}_t = \mathbf{C}\alpha_t + \mathbf{G}\beta_t + \mathbf{Z}\mathbf{b}_t + \varepsilon_t$$

where $\mathbf{Y}_t$ is the observed trait, **C** is the matrix of fixed-effect covariates, **G** is the matrix of genotypes, and **Z** is the matrix of random-effect covariates. To account for relatedness, $\mathbf{Z}\mathbf{Z}^\top = \mathbf{K}$ where **K** is a genetic relatedness matrix (GRM); and to account for technical and biological variation, **Z** is comprised of inferred factor covariates. We assume the residual ε_t is multivariate normal distributed with mean **0** and variance $\mathbf{I}\sigma_t^2$, and the random effects are multivariate normally distributed with mean **0** and variance $\mathbf{I}\tau_t^2$.

By default, variance components are estimated by restricted maximum likelihood (REML) under the null hypothesis of no single-variant associations. APEX supports sparse[66], [252] and low-rank[253] covariance matrices for random effects, and uses specialized optimizations for each structure. We expedite computation by precomputing and saving variance component estimates and LMM residuals for each trait, and residual genotypic variance terms for each variant. While APEX precomputes LMM residuals, we note that it does not use the GRAMMAR-gamma[254] or related approximations. For *trans*-xQTL analysis (considering all variant-trait pairs), the time complexity of LMM estimation and association testing in APEX is $O(pm^2n + npq + nmq)$ where n is the sample size, p the number of traits, m the number of covariates, and q the number of variants. Further details are provided in Supplementary Materials.

Omnibus p-values for cis-xQTL signals

We used the aggregated Cauchy association test (ACAT)[153], [154] to calculate omnibus *cis* region p-values for primary and secondary signals. ACAT omnibus p-values are calculated as $p^O = F\{\sum_i w_i F^{-1}(p_i)\}$ where F is the cumulative density function (CDF) of the standard Cauchy distribution, w_i are non-negative weights with $\sum_i w_i = 1$, and p_i are p-values. ACAT provides valid p-values under arbitrary dependence structures, provided that p_i are valid p-values (calibrated under the null hypothesis). When p_i are single-variant p-values in the *cis* region, we find that ACAT p-values with constant weights are highly concordant with permutation-based p-values (Supplementary Figure 6), but much faster (Figure 3, Panel B).

Data formats for flexible and accurate xQTL meta-analysis

APEX provides genetic association summary statistics (sumstat) and variance-covariance (vcov) data in an indexed, compressed binary format (Supplementary Figures 7-8). For fixed effects models, APEX sumstat files store the vector of score statistics $\mathbf{U}_t = \mathbf{G}^\top \mathbf{P}\mathbf{Y_t}$ and residual sum of squares ${\mathbf{Y}_t}^\top \mathbf{P}\mathbf{Y}_t$ for each trait t, where **G** is the genotype matrix, **P** is a projection matrix, and **Y** is the matrix of molecular traits; APEX vcov files store the variance-covariance matrix of score statistics $\mathbf{V} = \mathbf{G}^\top \mathbf{P}\mathbf{G}$ (also called adjusted LD matrix). For *cis* analysis, we store only score statistics for variants within a window of each molecular trait (1 Mbp by default), and adjusted LD for variants within twice the specified window size. These statistics are sufficient for a wide variety of downstream statistical analyses (for example, multiple-variant joint and conditional regression modeling, aggregation tests, Bayesian fine-mapping, and colocalization analysis), and preserve the genetic privacy of xQTL study participants. Similar strategies have been used to aggregate variants for gene-based tests in rare-variant (RV) GWAS meta-analysis[227], [231], but to our knowledge no existing methods exist for efficiently sharing and combining adjusted LD for genome-wide meta-analysis of common variants in GWAS or xQTL studies. APEX summary data can be combined across multiple studies for meta-analysis in APEX mode *meta* for joint and conditional regression analysis, or accessed and combined through an R interface for use with other packages. Further details are given in Supplementary Materials.

Secondary xQTL signal discovery

We implemented stepwise regression algorithms to detect multiple conditionally independent genetic association signals (Supplementary Figure 9) using either individual-level data or sumstat and vcov files. At each iteration, we evaluate signal-level significance using an omnibus p-value to test the null hypothesis that no remaining variants are associated with the trait, calculated as $p^O = F\left\{\sum_{j\in U} w_j F^{-1}(p_{j|S})\right\}$, where S and U are the current sets of selected and unselected variants, $p_{j|S}$ is the conditional p-value for variant j given selected variants S , w_j is the weight for variant j (normalized so that $\sum_{j\in U}\ w_j = 1$ at iteration), and F is the CDF of the standard Cauchy distribution. If $p^O < \alpha$, where α is a specified threshold, we select the most significant variant in U (adding it to S and removing it from U) and continue; otherwise, we retain the current set S and exit. Further details and extensions are given in Supplementary Materials.

Data sources

LCL eQTL genotype data

Genotype data from the 1000 Genomes Project Phase 3 in NCBI build 38 were obtained from the International Genome Sample Resource (IGSR) webpage[255]. WGS-based genotype data for the GTEx project v8 were obtained from dbGaP under accession number (phg 001219.v1); variants and samples with >15% missingness were excluded. Remaining missing genotype calls were imputed as best-guess hard call genotypes using the phasing software Eagle[256]. Genotype data from the HapMap project in NCBI build 36 from the Broad Institute webpage. This data set included 1,379,607 autosomal variants; to increase the number of variants overlapping the other studies, HapMap genotypes were imputed with the 1000 Genomes Project Phase 3 reference panel using Minimac3[69]; imputed variants were filtered with Mach-Rsq > 0.3. A final list of 10,930,386 variants, the intersection of variants across the three studies, was used for meta-analysis. Kinship matrices and genetic principal component covariates were calculated using PLINK 2[252].

	Sample size	Genotype data	Total no. variants	Expression data	Total no. transcripts
GTEx v8	147	WGS	12,232,655	RNA-seq	22,759
Geuvadis	454	WGS	31,331,216	RNA-seq	17,815
HapMap	518	Genotyped and imputed	29,539,804	Expression microarray	16,329

Table 5.0.1. Descriptive statistics for LCL eQTL data sets. Summary of LCL data sets analyzed. For HapMap, we report the number of imputed variants. For all studies, we report the number of variants before filtering. Processing and filtering procedures for each study are described in Methods.

LCL gene expression data

RNA-seq expression data from the Geuvadis consortium, which performed RNA-seq on LCLs for a subset of samples in the 1000 Genomes Project, were obtained from the IGSR webpage[211]. RNA-seq expression data from LCLs for GTEx v8 participants were obtained from dbGaP under accession number (phe000037.v1). LCL expression microarray data for 618 individuals in the HapMap 3 study[221] were obtained from ArrayExpress[257]; Illumina probe identifiers were mapped to Ensembl gene identifiers using the illuminaHumanv2 Bioconductor R package. Genes that were lowly expressed (count ≤ 5) in ≥ 25% of individuals were excluded. Expression microarray measurements and RNA-seq TPMs were rank-normal transformed within each study[211].

Prior to association analysis of gene expression traits, we applied two-stage rank normalization by (a) applying a rank-normal transformation to each trait, (b) calculating trait residuals by regressing out technical covariates, and (c) applying a second rank-normal transformation to these trait residuals. This procedure is performed internally in APEX for *cis* and *trans* association analysis; we also performed two-stage rank-normalization in R for analysis using external software packages (e.g., finemapping analysis).

We identified 76 individuals overlapping between the HapMap and Geuvadis studies among those with LCL expression and genotype data in each study. We removed these 76 individuals from the HapMap data sets prior to analysis (and retained them in Geuvadis) to ensure that no participants were duplicated between studies. The reported sample sizes (Table 1) reflect these exclusions. We identified no other duplicates.

LCL eQTL study protocols and informed consent

Geuvadis protocols were approved by the 1000 Genomes Project Steering Committee and institutional review boards (IRBs) or ethics committees, and written informed consent forms were signed by all Geuvadis participants (`https://www.internationalgenome.org/about/`). HapMap protocols were approved by IRBs or ethics committees by all involved institutions, and written informed consent forms were signed by all participants[258]. GTEx protocols were approval by local IRBs or ORSP by all involved institutions[259]; informed consent was provided by next of kin for all participants (all human donors in the GTEx project were deceased)[31].

UK Biobank genotype data

Genotype data from the UK Biobank study were obtained under Application Number 52008. UK Biobank protocols were approved by the National Research Ethics Service Committee and written informed consent were signed by the participants. Marker variants were filtered by including only autosomal SNPs with genotype missingness < 1% that passed all batch-wise genotype quality control steps[260] (590,606 variants after filtering). We randomly selected a multi-ethnic subset of 10,000 UK Biobank participants for analysis, among which 4,000 were Irish, 3,000 were South Asian (Indian, Pakistani, and Bangladeshi), and 3,000 were African and Caribbean (all self-reported). We generated an ancestry-adjusted sparse genetic relatedness matrix (GRM) using LD-pruned MAF > 0.01 variants in R by projecting out genotype PCs from genotypes and setting GRM elements to 0 for > 4th degree estimated relatives (genetic correlation < 0.044). LD pruning used pairwise r^2 < 0.1 in sliding windows of 50 SNPs moving 5 SNPs at a time.

Data analysis and simulation procedures

Molecular trait simulation procedures

To evaluate Type I error rates of association test statistics, we simulated expression data under the null hypothesis of no single-variant genetic associations in the Geuvadis study. We used the empirical covariance between expression and technical covariates and simulate covariance of expression residuals to simulate expression with a realistic correlation

structure (Supplementary Figures 1-2). Specifically, in each replicate, we simulated the row vector of expression across genes for participant i as a multivariate normal distribution with mean $(\hat{\alpha}_{\mathbf{1}}, \ldots, \hat{\alpha}_{\mathbf{p}})^{\top} \mathbf{C}_i^{\top}$ and variance $\tilde{\Sigma}$, where $\mathbf{C}_i$ is the i^{th} row vector of from technical covariates **C** (genotype PCs, gender, batch, ethnicity indicator), $\hat{\alpha}_{\mathbf{j}} = \left(\mathbf{C}^{\top}\mathbf{C}\right)^{-1} \mathbf{C}^{\top}\mathbf{Y}_j$ is the estimated effects of technical covariates on gene j expression $\mathbf{Y}_j$ (column vector), and $\tilde{\Sigma} = \frac{1}{n-1}\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^{\top}$ is the sample covariance matrix of expression residuals across genes where $\tilde{\mathbf{Y}} = (\mathbf{I} - \mathbf{C}\left(\mathbf{C}^{\top}\mathbf{C}\right)^{-1}\mathbf{C}^{\top})\mathbf{Y}$. In each simulation replicate, we re-calculated the inferred covariates (ePC, eFA, or PEER) from the simulated expression matrix.

We simulated expression data in the UK Biobank study to assess the computational performance of linear mixed models (LMMs) for xQTL analysis in large cohorts, which will be critical to identify rare and small-effect xQTL variants and molecular traits that contribute to heritable diseases. In these experiments, we simulated each trait independently from a multivariate normal distribution with mean $\mathbf{C}\alpha$, where **C** is the matrix of genotype PCs, and variance $h^2\mathbf{K} + (1 - h^2)\mathbf{I}$ where **K** is the sparse genetic relatedness matrix. We simulated the covariate effects α from an independent normal distribution, and pseudo-heritability parameter h^2 from a uniform distribution.

LCL eQTL fine-mapping analysis

We performed Bayesian finemapping of gene expression traits using the susieR package with both individual-level and summary-level data[228]. For each gene, we analyzed all variants within a 1 Mbp window of the transcription start site. We used the *susie::susie* and *susie::susie_suff_stat* functions in the susieR package to finemap from individual-level data and summary statistics respectively using the default L = 10 maximum number of causal variants. To correct for technical covariates in individual-level analysis, we used residualized genotype and expression matrices calculated as $\tilde{\mathbf{G}} = \mathbf{PG}$ and $\tilde{\mathbf{Y}} = \mathbf{PY}$ respectively, where **G** and **Y** are the genotype and expression matrices, $\mathbf{P} = \mathbf{I} - \mathbf{C}\left(\mathbf{C}^{\top}\mathbf{C}\right)^{-1}\mathbf{C}^{\top}$, and **C** is the matrix of technical covariates. To correct for technical covariates in summary-based analysis, we calculated the vector of score statistics for each trait t as $\hat{\sigma}_{\mathbf{t}}^{-\mathbf{2}}\mathbf{G}^{\top}\mathbf{P}\mathbf{Y}_t$ where $\hat{\sigma}_{\mathbf{t}}^{\mathbf{2}} = \frac{1}{n-m}\mathbf{Y}_t^{\top}\mathbf{P}\mathbf{Y}_t$, and the variance-covariance matrix either as $\mathbf{V}_t = \hat{\sigma}_{\mathbf{t}}^{-\mathbf{2}}\mathbf{G}^{\top}\mathbf{P}\mathbf{G}$ (for adjusted LD) or $\mathbf{V}_t^U = \hat{\sigma}_{\mathbf{t}}^{-\mathbf{2}}\mathbf{G}^{\top}(\mathbf{I} - \mathbf{1}\left(\mathbf{1}^{\top}\mathbf{1}\right)^{-1}\mathbf{1}^{\top})\mathbf{G}$ (for unadjusted LD) as described in Re-

sults. To use external LD as a proxy for the variance-covariance matrix, we calculated the proxy-LD variance-covariance matrix as $\mathbf{V}_t^P = \mathbf{A}_t^{1/2}\mathbf{R}^P\mathbf{A}_t^{1/2}$, where $\mathbf{A}_t$ is a diagonal matrix with the diagonal entries of $\mathbf{V}_t$, and $\mathbf{R}^P$ is the proxy LD matrix, calculated as the sample correlation matrix of adjusted genotypes from the reference panel.

LCL eQTL enrichment analysis

We used binomial logistic regression models to assess functional enrichment of LCL eQTLs. The mean model was specified $logit[P(t_j = 1)] = \mathbf{c}_j^\top\alpha + x_j\gamma$, where the outcome was defined as $t_j = 1$ if variant j is in high LD (r^2 > 0.8) with a lead LCL eQTL variant for any gene and $t_j = 0$ otherwise, where lead eQTL variants were identified using stepwise regression (described above). The scalar x_j denotes the feature of interest (e.g., $x_j = 1$ if variant j overlaps a lymphoid-specific DHS and $x_j = 0$ otherwise), and the covariate vector $\mathbf{c}_j$ included an intercept and cubic b-spline terms for log-transformed minor allele frequency (MAF) and distance to nearest transcription start site (TSS). We included all variants that were tested for *cis* association (within 1 Mbp of TSS for any tested gene).

Figures

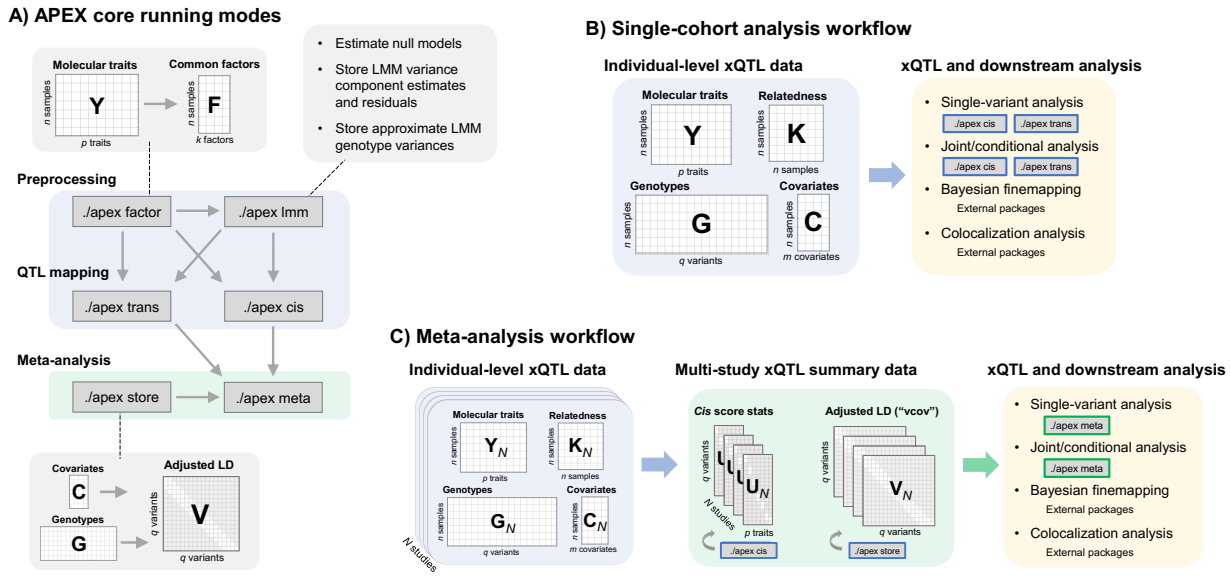


Figure 5.0.1. APEX toolkit for molecular QTL mapping and meta-analysis. A: Mode *factor* provides factor analysis to infer shared technical and biological factors across traits. In QTL mapping (modes *cis* and *trans*), inferred factor covariates can be modeled as fixed effects (by appending matrix **F** to covariate matrix **C**) or random effects (using mode *lmm*). Mode *lmm* enables rapid linear mixed model (LMM) association analysis (in modes *cis* and *trans*) by precomputing and storing variance component estimates, LMM trait residuals, and approximate LMM genotypic variances. Mode *store* generates compact adjusted LD files for accurate multiple-variant analysis from summary statistics (using mode *meta* for meta-analysis). **B:** Individual-level molecular trait, genotype, and covariate data (and optional genetic relatedness matrix) are used as input for single-variant and joint/conditional association analysis across traits (APEX modes *cis* and *trans*). These data can also be used for Bayesian finemapping and colocalization analysis using external software packages. **C:** Each study generates summary data files (single-variant score statistics using mode *cis* and adjusted LD matrices using mode *store*) from individual-level data. These summary files can be used for single-variant and joint/conditional association meta-analysis in mode *meta*, or combined using the *Apex2R* interface to create input data for Bayesian finemapping and colocalization analysis using external packages.

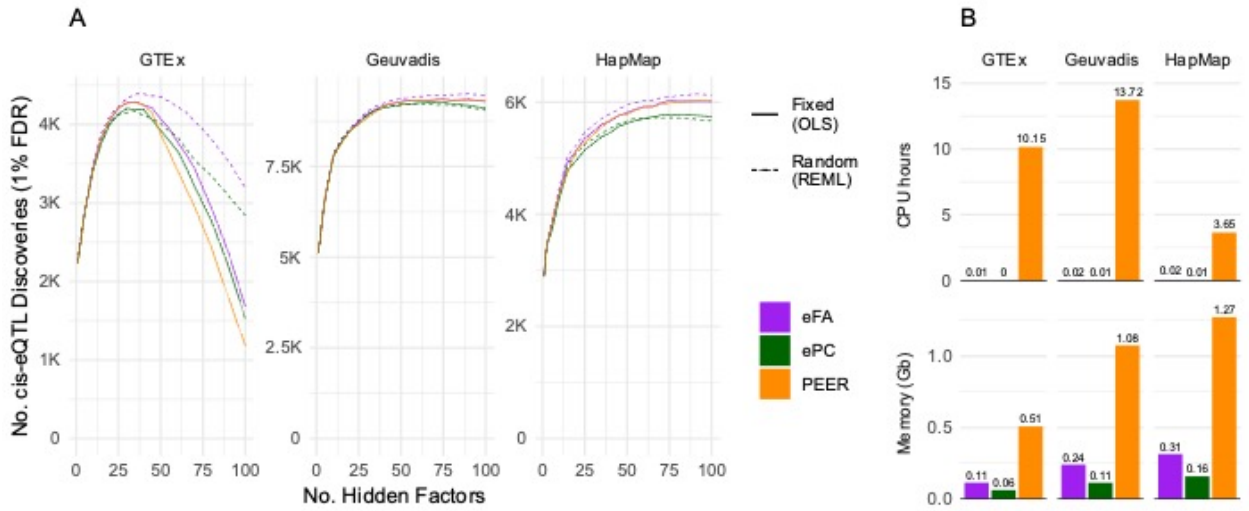


Figure 5.0.2. Rapid factor analysis and linear mixed models for *cis*-eQTL analysis. A: Number of LCL cis-eQTL discoveries at 1% FDR as a function of the number of hidden factors (x axis) inferred using PEER, factor analysis (eFA), or principal components analysis (ePC) across 3 studies. ePC and eFA covariate effects were estimated either as fixed effects (using OLS) or random effects (using REML) in association analysis using APEX. PEER covariates effects were estimated as fixed effects. **B:** Total running time (CPU hours) and maximum memory usage to generate ePC, eFA, and PEER covariates across models with 5, 10, 20, 40, 60, 80, and 100 latent factors. All jobs used a single CPU core. ePC and eFA covariates were calculated using APEX; PEER covariates were calculated using the PEER R package version 1.3 with a maximum of 1000 iterations.

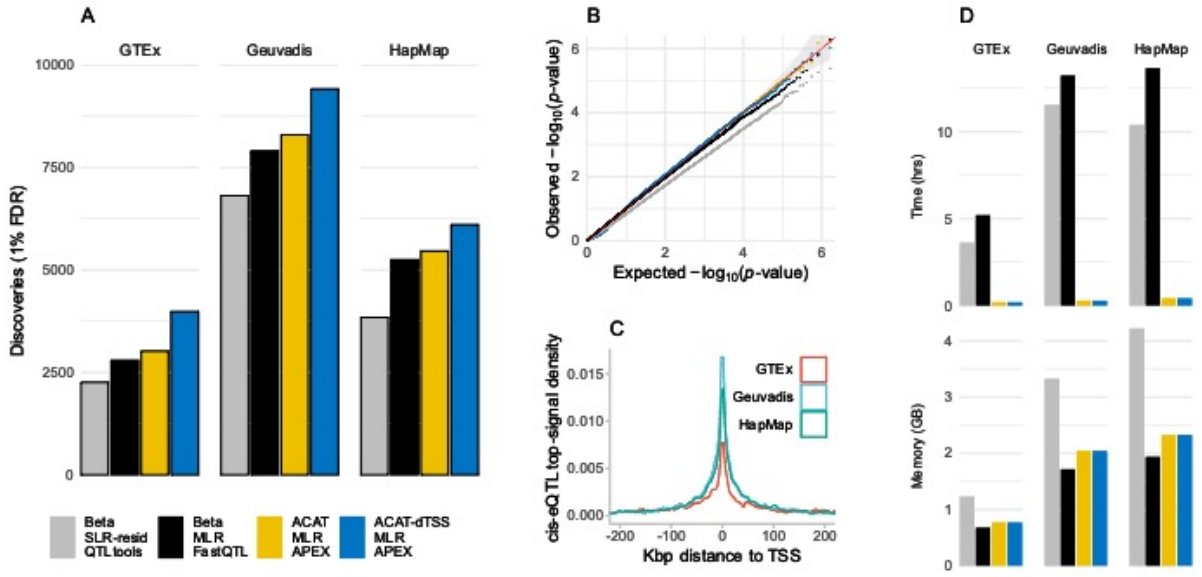


Figure 5.0.3. Fast and powerful *cis*-eQTL omnibus tests. A: ACAT and dTSS weights increase eGene discoveries. Gene-level *cis*-eQTL discoveries for each LCL data set at 1% FDR. Because all methods maintain calibrated Type I error rates in simulations (panel B), a larger number of discoveries suggests greater statistical power. Note that the number of tested genes varies across the three studies (Figure 4). **B**: Calibration of permutation-based and ACAT p-values. Q-Q plots for each method in simulations under the null hypothesis using genotype and expression data from Geuvadis. Traits were simulated using the observed correlation structure of gene expression, and expression PC covariates were re-calculated from simulated expression values in each replicate (Methods). P-values for all methods maintain calibrated or conservative Type I error rates, and SLR-resid permutation-based p-values are notably conservative. **C:** eQTL enrichment by dTSS. Density of chromosomal distance between top cis-eVariant and TSS across genes for each study. *Cis*-eVariants are strongly enriched nearer the TSS. **D:** CPU time and memory for eGene discovery. Analyses were run sequentially across chromosomes with 1 CPU; we report maximum memory usage and total running time across all 22 autosomes for each of the 4 methods.

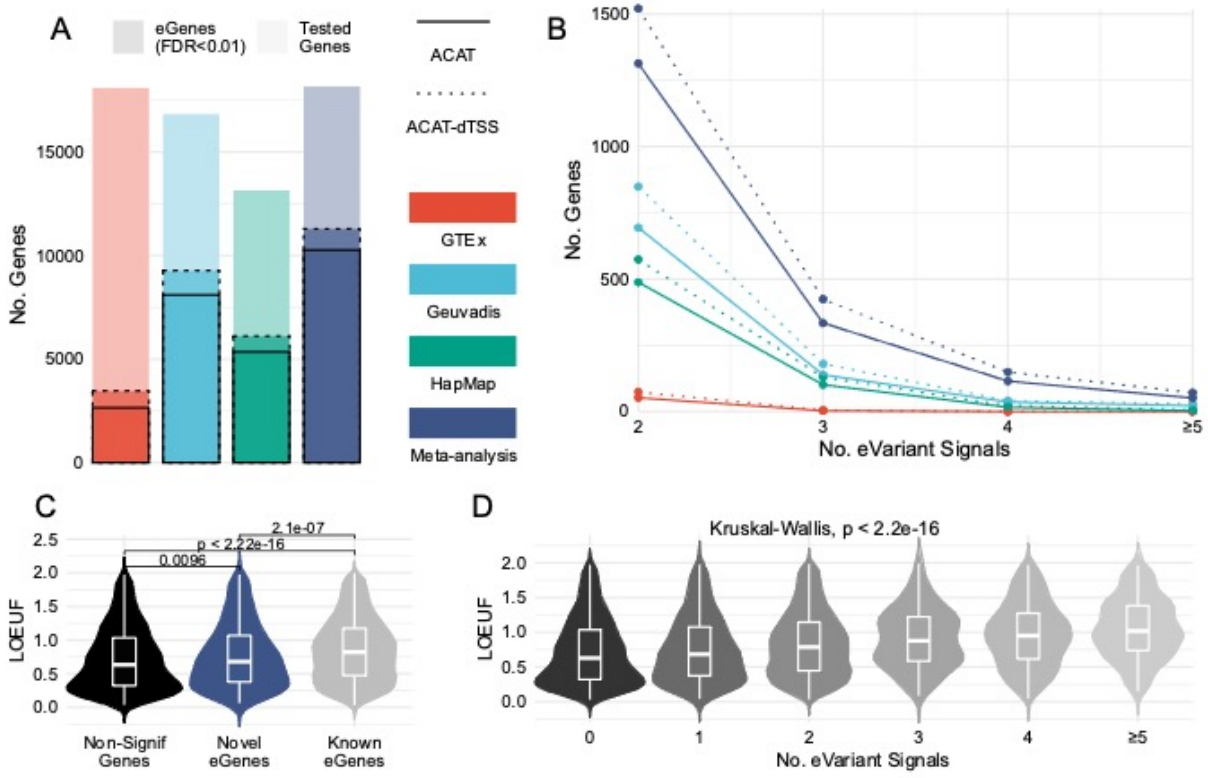


Figure 5.0.4. -analysis identifies novel primary and secondary *cis*-eQTLs. **A:** Meta-analysis and dTSS weights increase eGene discoveries. eGenes detected in LCL cis-eQTL analysis across studies and meta-analysis. Colored bars show total numbers of tested genes, and outlined bars show numbers of eGenes (cis-eQTL genes) detected at 1% FDR using unweighted ACAT (solid line) and or distance to transcription start site (dTSS) weighted ACAT (dashed line). dTSS weights increased eGene discoveries by 30.6% for GTEx, 14.4% for Geuvadis, 14.1% for HapMap, and 10.0% for meta-analysis. **B:** Meta-analysis and dTSS weights increase secondary eQTL discoveries. Secondary cis-eQTL variant discoveries across studies and meta-analysis. Shown are numbers of genes with 2, 3, 4, or $\geq$ 5 LCL eQTL eVariant signals detected at 1% FDR using unweighted (solid line) and dTSS-weighted ACAT. dTSS weights increased secondary signal discoveries by 43.6% for GTEx, 23.3% for Geuvadis, 20.4% for HapMap, and 19.3% for meta-analysis. **C:** Meta-analysis detects *cis*-eQTLs for constrained genes. Loss of function (LoF) observed/expected upper bound fraction (LOEUF) is a metric of genetic constraint; constrained genes have smaller LOEUF. LOEUF densities are shown for the 11,750 genes present in all (3 out of 3) studies, divided into 3 categories: (a) no cis-eQTLs detected at 1% FDR (2,659 "non-signif" genes), (b) $\geq$ 1 eQTL detected in meta-analysis but not individual studies (693 "novel eGenes"), and (c) $\geq$ 1 eQTL detected by $\geq$ 1 individual study (8,398 "known eGenes"). Both novel and non-significant genes have significantly lower LOEUF than known eGenes, suggesting greater constraint. **D: Fewer secondary *cis*-eQTLs are detected for constrained genes.** LOEUF densities for genes with 0, 1, ... $\geq$ 5 significant eVariants detected by stepwise regression in meta-analysis (1% FDR), shown for genes present in 3 out of 3 studies. Genes with more eVariants tend to have higher LOEUF (less constraint), as expected.

In panels C and D, box hinges show inter-quartile ranges (IQR) with median lines; whiskers show the highest and lowest values <1.5 * IQR from the hinge. Density plots were calculated with default settings using *geom_violin* from the ggplot2 R package (version 3.3.2)

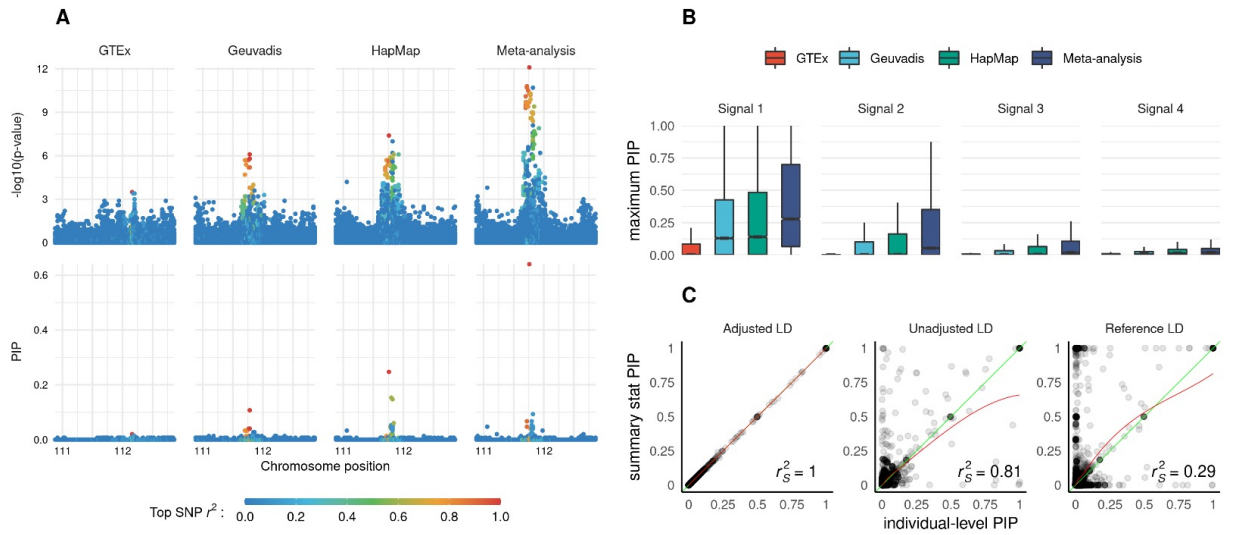


Figure 5.0.5. Accurate QTL fine-mapping from summary statistics. APEX xQTL sumstat and vcov files enable accurate multiple-variant analyses without individual-level data. Here, we illustrate Bayesian finemapping from APEX summary statistics data using the *susieR* package and *Apex2R* interface to access sumstat and vcov files. **A:** Finemapping *cis*-eQTLs from summary statistics. *cis*-eQTL p-values (upper panel) and posterior inclusion probabilities (PIPs) for *cis* variants at the *FYN* locus (6.p22) are shown across the three studies and meta-analysis. Meta-analysis increases signal strength (upper panels) and precision identifying putative causal variants (lower panels). **B:** Meta-analysis increases finemapping precision. We finemapped 9,787 genes present each of the 3 studies from APEX sumstat and vcov summary data files using the *susieR* package. For each gene, we assigned each variant to its most likely signal cluster (highest posterior probability), and calculated the maximum PIP across variants within each signal cluster. Boxplots show the distribution of the maximum PIP within the 1st, 2nd, 3rd and 4th signal cluster across genes for each study. Box hinges show inter-quartile ranges (IQR) with median lines; whiskers show the highest and lowest values <1.5 * IQR from the hinge. Maximum PIPs tend to increase with sample size, as expected. **C:** APEX sumstat and vcov files enable accurate finemapping from summary statistics. Concordance of PIPs across 71 genes using individual-level data (x axis) vs summary statistics (y axis) from HapMap with covariate-adjusted HapMap LD (left), HapMap LD not adjusted for covariates (middle), or proxy LD from Geuvadis (right) adjusted for similar covariates. PIPs from summary statistics using APEX vcov files (adjusted LD) are nearly numerically equivalent with individual-level analysis. PIPs using unadjusted or proxy LD are less concordant with individual-level analysis (Spearman r^2 0.81 or 0.29 respectively).

Supplementary Figures

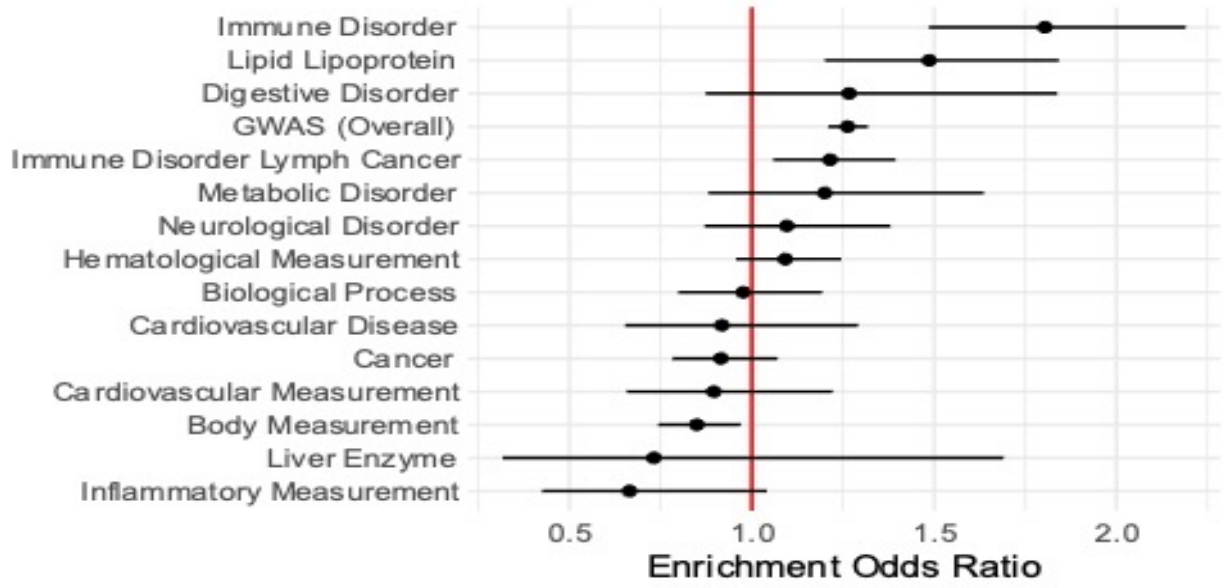


Figure 5.0.6. LCL eQTL enrichment for categories of traits in the NHGRI-EBI GWAS Catalog, adjusted for minor allele frequency (MAF) (log-transformed cubic b-spline) and distance to nearest transcription start site (TSS) (log-transformed cubic b-spline) as described in Methods. LCL eQTLs show strongest enrichment with immune disorders. Shown are enrichment odds ratios (exponentiated logistic regression coefficients, ± 2 standard errors) estimated separately by including a single GWAS trait category per model.

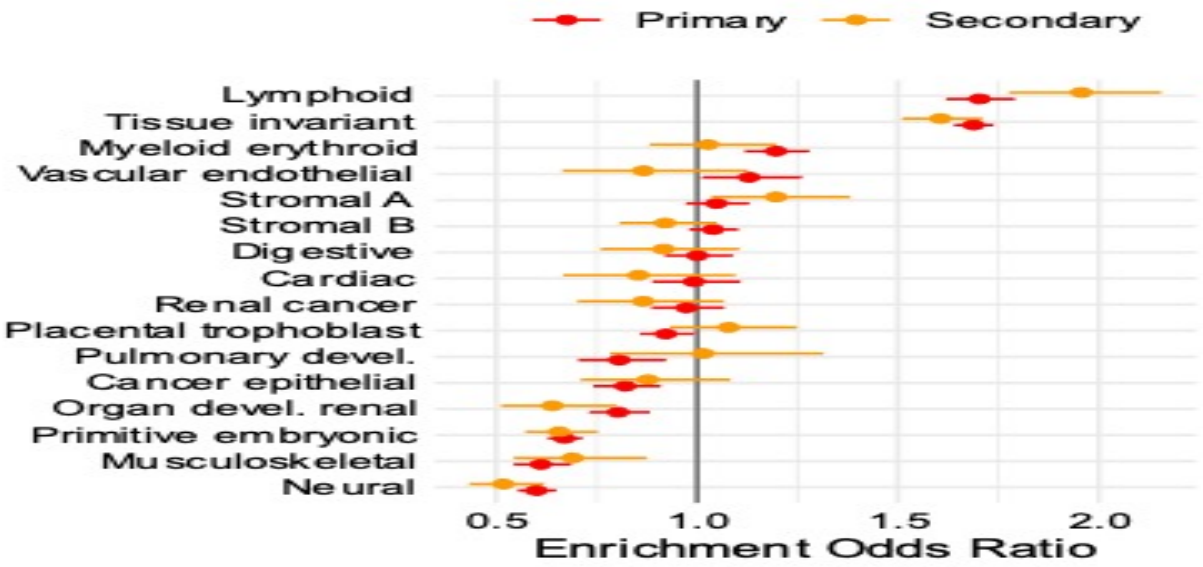


Figure 5.0.7. Primary and secondary LCL eQTL enrichment in tissue-specific DNase I hypersensitive sites (DHSs)[248], adjusted for minor allele frequency (MAF) (log-transformed cubic b-spline) and distance to nearest transcription start site (TSS) (log-transformed cubic b-spline) as described in Methods. Shown are enrichment odds ratios (exponentiated logistic regression coefficients, ± 2 standard errors) estimated separately by including a single DHS tissue category per model.

Bibliography

[1] Hon-Cheong So, Allen H. S. Gui, Stacey S. Cherny, et al. “Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases”. *Genetic Epidemiology* 2011.

[2] David J. Hunter. “Gene–environment interactions in human diseases”. *Nature Reviews Genetics* 2005.

[3] Wenfei Jin, Pengfei Qin, Haiyi Lou, et al. “A systematic characterization of genes underlying both complex and Mendelian diseases”. *Human Molecular Genetics* 2012.

[4] Doug Speed, Gibran Hemani, Michael R. Johnson, et al. “Improved heritability estimation from genome-wide SNPs”. *American Journal of Human Genetics* 2012.

[5] Eppie M. Yiu and Andrew J. Kornberg. “Duchenne muscular dystrophy”. *Journal of Paediatrics and Child Health* 2015.

[6] J. M. Rommens, M. C. Iannuzzi, B. Kerem, et al. “Identification of the cystic fibrosis gene: chromosome walking and jumping”. *Science* 1989.

[7] J. R. Riordan, J. M. Rommens, B. Kerem, et al. “Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA”. *Science* 1989.

[8] B. Kerem, J. M. Rommens, J. A. Buchanan, et al. “Identification of the cystic fibrosis gene: genetic analysis”. *Science* 1989.

[9] “A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington’s disease chromosomes. The Huntington’s Disease Collaborative Research Group”. *Cell* 1993.

[10] J. R. Gulcher, A. Kong, and K. Stefansson. “The role of linkage studies for common diseases”. *Current Opinion in Genetics & Development* 2001.

[11] Eric Londin, Priyanka Yadav, Saul Surrey, et al. “Use of Linkage Analysis, Genome-Wide Association Studies, and Next-Generation Sequencing in the Identification of Disease-Causing Mutations”. *Pharmacogenomics: Methods and Protocols*. Ed. by Federico Innocenti and Ron H.N. van Schaik. Methods in Molecular Biology. 2013.

[12] B. K. Sinha, D. P. Monga, and S. Prasad. “Studies on the role of macrophages in experimental candidosis in mice”. *Mykosen* 1987.

[13] N. J. Risch. “Searching for genetic determinants in the new millennium”. *Nature* 2000.

[14] Shuquan Rao, Yao Yao, and Daniel E. Bauer. “Editing GWAS: experimental approaches to dissect and exploit disease-associated genetic variation”. *Genome Medicine* 2021.

[15] N. Risch and K. Merikangas. “The future of genetic studies of complex human diseases”. *Science* 1996.

[16] Leonid Kruglyak. “The road to genome-wide association studies”. *Nature Reviews. Genetics* 2008.

[17] Psychiatric GWAS Consortium Coordinating Committee, Sven Cichon, Nick Craddock, et al. “Genomewide association studies: history, rationale, and prospects for psychiatric disorders”. *The American Journal of Psychiatry* 2009.

[18] Annalisa Buniello, Jacqueline A. L. MacArthur, Maria Cerezo, et al. “The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019”. *Nucleic Acids Research* D1 2019.

[19] Melinda C. Mills and Charles Rahal. “A scientometric review of genome-wide association studies”. *Communications Biology* 2019.

[20] Peter M. Visscher, Matthew A. Brown, Mark I. McCarthy, et al. “Five Years of GWAS Discovery”. *The American Journal of Human Genetics* 2012.

[21] Lu Chen, Bing Ge, Francesco Paolo Casale, et al. “Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells”. *Cell* 2016.

[22] Karsten Suhre, Mark I. McCarthy, and Jochen M. Schwenk. “Genetics meets proteomics: perspectives for large population-based studies”. *Nature Reviews. Genetics* 2021.

[23] *IDF Diabetes Atlas 9th edition 2019*. URL: `https://diabetesatlas.org/en/` (visited on 05/26/2021).

[24] Yan Zheng, Sylvia H. Ley, and Frank B. Hu. “Global aetiology and epidemiology of type 2 diabetes mellitus and its complications”. *Nature Reviews Endocrinology* 2018.

[25] Mark P. Keller, YounJeong Choi, Ping Wang, et al. “A gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility”. *Genome Research* 2008.

[26] Laura J. Scott, Michael R. Erdos, Jeroen R. Huyghe, et al. “The genetic regulatory signature of type 2 diabetes in human skeletal muscle”. *Nature Communications* 2016.

[27] D. Leland Taylor, Anne U. Jackson, Narisu Narisu, et al. “Integrative analysis of gene expression, DNA methylation, physiological traits, and genetic variation in human skeletal muscle”. *Proceedings of the National Academy of Sciences of the United States of America* 2019.

[28] Mete Civelek, Ying Wu, Calvin Pan, et al. “Genetic Regulation of Adipose Gene Expression and Cardio-Metabolic Traits”. *American Journal of Human Genetics* 2017.

[29] Chelsea K. Raulerson, Arthur Ko, John C. Kidd, et al. “Adipose Tissue Gene Expression Associations Reveal Hundreds of Candidate Genes for Cardiometabolic Traits”. *The American Journal of Human Genetics* 2019.

[30] Emma Nilsson, Per Anders Jansson, Alexander Perfilyev, et al. “Altered DNA methylation and differential expression of genes influencing metabolism and inflammation in adipose tissue from subjects with type 2 diabetes”. *Diabetes* 2014.

[31] GTEx Consortium. “The GTEx Consortium atlas of genetic regulatory effects across human tissues”. *Science* 2020.

[32] Paul W. Franks. *The Complex Interplay of Genetic and Lifestyle Risk Factors in Type 2 Diabetes: An Overview*. Scientifica. 2012. URL: `https://www.hindawi.com/journals/scientifica/2012/482186/` (visited on 01/19/2021).

[33] Anubha Mahajan, Daniel Taliun, Matthias Thurner, et al. “Fine-mapping of an expanded set of type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps”. *Nature genetics* 2018.

[34] Peter M. Visscher, Naomi R. Wray, Qian Zhang, et al. “10 Years of GWAS Discovery: Biology, Function, and Translation”. *American Journal of Human Genetics* 2017.

[35] Jason M. Torres, Moustafa Abdalla, Anthony Payne, et al. “A Multi-omic Integrative Scheme Characterizes Tissues of Action at Loci Associated with Type 2 Diabetes”. *American Journal of Human Genetics* 2020.

[36] Ambrish K. Srivastava. “Challenges in the treatment of cardiometabolic syndrome”. *Indian Journal of Pharmacology* 2012.

[37] Anne-Karien M. de Waard, Monika Hollander, Joke C. Korevaar, et al. “Selective prevention of cardiometabolic diseases: activities and attitudes of general practitioners across Europe”. *European Journal of Public Health* 2019.

[38] Birgit Gustafson, Ann Hammarstedt, Christian X. Andersson, et al. “Inflamed adipose tissue: a culprit underlying the metabolic syndrome and atherosclerosis”. *Arteriosclerosis, Thrombosis, and Vascular Biology* 2007.

[39] B. Chowdhury, L. Sjöström, M. Alpsten, et al. “A multicompartment body composition technique based on computerized tomography”. *International Journal of Obesity and Related Metabolic Disorders: Journal of the International Association for the Study of Obesity* 1994.

[40] Allison J. Richard, Ursula White, Carrie M. Elks, et al. “Adipose Tissue: Physiology to Metabolic Dysfunction”. *Endotext*. Ed. by Kenneth R. Feingold, Bradley Anawalt, Alison Boyce, et al. South Dartmouth (MA): MDText.com, Inc., 2000.

[41] X. Yang and U. Smith. “Adipose tissue distribution and risk of metabolic disease: does thiazolidinedione-induced adipose tissue redistribution provide a clue to the answer?” *Diabetologia* 2007.

[42] Fredrik Karpe and Katherine E. Pinnick. “Biology of upper-body and lower-body adipose tissue–link to whole-body phenotypes”. *Nature Reviews. Endocrinology* 2015.

[43] Biljana Atanasovska, Vinod Kumar, Jingyuan Fu, et al. “GWAS as a Driver of Gene Discovery in Cardiometabolic Diseases”. *Trends in endocrinology and metabolism: TEM* 2015.

[44] Valur Emilsson, Gudmar Thorleifsson, Bin Zhang, et al. “Genetics of gene expression and its effect on disease”. *Nature* 2008.

[45] Elin Grundberg, Kerrin S. Small, Åsa K. Hedman, et al. “Mapping cis- and trans-regulatory effects across multiple tissues in twins”. *Nature genetics* 2012.

[46] Hassan Foroughi Asl, Husain A. Talukdar, Alida S. D. Kindt, et al. “Expression quantitative trait Loci acting across multiple tissues are enriched in inherited risk for coronary artery disease”. *Circulation. Cardiovascular Genetics* 2015.

[47] Oscar Franzén, Raili Ermel, Ariella Cohain, et al. “Cardiometabolic Risk Loci Share Downstream Cis- and Trans-Gene Regulation Across Tissues and Diseases”. *Science* 2016.

[48] Adeline R. Whitney, Maximilian Diehn, Stephen J. Popper, et al. “Individuality and variation in gene expression patterns in human blood”. *Proceedings of the National Academy of Sciences of the United States of America* 2003.

[49] Roby Joehanes, Xiaoling Zhang, Tianxiao Huan, et al. “Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies”. *Genome Biology* 2017.

[50] Holger Kirsten, Hoor Al-Hasani, Lesca Holdt, et al. “Dissecting the genetics of the human transcriptome identifies novel trait-related trans-eQTLs and corroborates the regulatory relevance of non-protein coding loci”. *Human Molecular Genetics* 2015.

[51] Rick Jansen, Jouke-Jan Hottenga, Michel G. Nivard, et al. “Conditional eQTL analysis reveals allelic heterogeneity of gene expression”. *Human Molecular Genetics* 2017.

[52] Rudolf S. N. Fehrmann, Ritsert C. Jansen, Jan H. Veldink, et al. “Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA”. *PLoS genetics* 2011.

[53] Ettje F. Tigchelaar, Alexandra Zhernakova, Jackie A. M. Dekens, et al. “Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics”. *BMJ open* 2015.

[54] Solveig K. Sieberts, Thanneer M. Perumal, Minerva M. Carrasquillo, et al. “Large eQTL meta-analysis reveals differing patterns between cerebral cortical and cerebellar brain regions”. *Scientific Data* 2020.

[55] Urmo Võsa, Annique Claringbould, Harm-Jan Westra, et al. “Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis”. *bioRxiv* 2018. Publisher: Cold Spring Harbor Laboratory Section: New Results.

[56] Inez Trouwborst, Suzanne M. Bowser, Gijs H. Goossens, et al. “Ectopic Fat Accumulation in Distinct Insulin Resistant Phenotypes; Targets for Personalized Nutritional Interventions”. *Frontiers in Nutrition* 2018.

[57] Yurong Xin, Jinrang Kim, Haruka Okamoto, et al. “RNA Sequencing of Single Human Islet Cells Reveals Type 2 Diabetes Genes”. *Cell Metabolism* 2016.

[58] Klev Diamanti, Marco Cavalli, Gang Pan, et al. “Intra- and inter-individual metabolic profiling highlights carnitine and lysophosphatidylcholine pathways as key molecular defects in type 2 diabetes”. *Scientific Reports* 2019.

[59] Esther Phielix and Marco Mensink. “Type 2 diabetes mellitus and skeletal muscle metabolic function”. *Physiology & Behavior* 2008.

[60] M. Snel, J. T. Jonker, J. Schoones, et al. “Ectopic fat and insulin resistance: pathophysiology and effect of diet and lifestyle interventions”. *International Journal of Endocrinology* 2012.

[61] Unai Galicia-Garcia, Asier Benito-Vicente, Shifa Jebari, et al. “Pathophysiology of Type 2 Diabetes Mellitus”. *International Journal of Molecular Sciences* 2020.

[62] Clinton C. Mason, Robert L. Hanson, Vicky Ossowski, et al. “Bimodal distribution of RNA expression levels in human skeletal muscle tissue”. *BMC Genomics* 2011.

[63] Mete Civelek, Raffi Hagopian, Calvin Pan, et al. “Genetic regulation of human adipose microRNA expression and its consequences for metabolic traits”. *Human Molecular Genetics* 2013.

[64] Alexander W. Drong, George Nicholson, Åsa K. Hedman, et al. “The Presence of Methylation Quantitative Trait Loci Indicates a Direct Genetic Influence on the Level of DNA Methylation in Adipose Tissue”. *PLOS ONE* 2013.

[65] Petr Volkov, Anders H. Olsson, Linn Gillberg, et al. “A Genome-Wide mQTL Analysis in Human Adipose Tissue Identifies Genetic Variants Associated with DNA Methylation, Gene Expression and Metabolic Traits”. *PloS One* 2016.

[66] Ani Manichaikul, Josyf C. Mychaleckyj, Stephen S. Rich, et al. “Robust relationship inference in genome-wide association studies”. *Bioinformatics (Oxford, England)* 2010.

[67] John Novembre, Toby Johnson, Katarzyna Bryc, et al. “Genes mirror geography within Europe”. *Nature* 2008.

[68] Po-Ru Loh, Petr Danecek, Pier Francesco Palamara, et al. “Reference-based phasing using the Haplotype Reference Consortium panel”. *Nature Genetics* 2016.

[69] Sayantan Das, Lukas Forer, Sebastian Schönherr, et al. “Next-generation genotype imputation service and methods”. *Nature Genetics* 2016.

[70] Stephen W. Hartley and James C. Mullikin. “QoRTs: a comprehensive toolset for quality control and data processing of RNA-Seq experiments”. *BMC Bioinformatics* 2015.

[71] Goo Jun, Matthew Flickinger, Kurt N. Hetrick, et al. “Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data”. *American Journal of Human Genetics* 2012.

[72] Martin J. Aryee, Andrew E. Jaffe, Hector Corrada-Bravo, et al. “Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays”. *Bioinformatics (Oxford, England)* 2014.

[73] Sarah Dedeurwaerder, Matthieu Defrance, Martin Bizet, et al. “A comprehensive overview of Infinium HumanMethylation450 data processing”. *Briefings in Bioinformatics* 2014.

[74] Jean-Philippe Fortin, Elana Fertig, and Kasper Hansen. “shinyMethyl: interactive quality control of Illumina 450k DNA methylation arrays in R”. *F1000Research* 2014.

[75] Joel Rozowsky, Robert R. Kitchen, Jonathan J. Park, et al. “exceRpt: A Comprehensive Analytic Platform for Extracellular RNA Profiling”. *Cell Systems* 2019.

[76] Sam Griffiths-Jones. “The microRNA Registry”. *Nucleic Acids Research* Database issue 2004.

[77] Oliver Stegle, Leopold Parts, Richard Durbin, et al. “A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies”. *PLoS computational biology* 2010.

[78] Nick Patterson, Alkes L. Price, and David Reich. “Population Structure and Eigenanalysis”. *PLOS Genetics* 2006.

[79] Olivier Delaneau, Halit Ongen, Andrew A. Brown, et al. “A complete tool set for molecular QTL discovery and analysis”. *Nature Communications* 2017.

[80] John D. Storey and Robert Tibshirani. “Statistical significance for genomewide studies”. *Proceedings of the National Academy of Sciences* 2003.

[81] Yoav Benjamini and Yosef Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing”. *Journal of the Royal Statistical Society. Series B (Methodological)* 1995.

[82] Xiaoquan Wen, Yeji Lee, Francesca Luca, et al. “Efficient Integrative Multi-SNP Association Analysis via Deterministic Approximation of Posteriors”. *American Journal of Human Genetics* 2016.

[83] Xiaoquan Wen, Roger Pique-Regi, and Francesca Luca. “Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization”. *PLOS Genetics* 2017.

[84] Milton Pividori, Padma S. Rajagopal, Alvaro Barbeira, et al. “PhenomeXcan: Mapping the genome to the phenome through the transcriptome”. *bioRxiv* 2019.

[85] Arushi Varshney, Laura J. Scott, Ryan P. Welch, et al. “Genetic regulatory signatures underlying islet gene expression and type 2 diabetes”. *Proceedings of the National Academy of Sciences* 2017.

[86] Maren E. Cannon, Kevin W. Currin, Kristin L. Young, et al. “Open Chromatin Profiling in Adipose Tissue Marks Genomic Regions with Functional Roles in Cardiometabolic Traits”. *G3 (Bethesda, Md.)* 2019.

[87] M. Wabitsch, R. E. Brenner, I. Melzner, et al. “Characterization of a human preadipocyte cell strain with high capacity for adipose differentiation”. *International Journal of Obesity and Related Metabolic Disorders: Journal of the International Association for the Study of Obesity* 2001.

[88] Marie P. Fogarty, Maren E. Cannon, Swarooparani Vadlamudi, et al. “Identification of a Regulatory Variant That Binds FOXA1 and FOXA2 at the CDC123/CAMK1D Type 2 Diabetes GWAS Locus”. *PLOS Genetics* 2014.

[89] Seunggeun Lee, Wei Sun, Fred A. Wright, et al. “An improved and explicit surrogate variable analysis procedure by coefficient adjustment”. *Biometrika* 2017.

[90] Craig A. Glastonbury, Alexessander Couto Alves, Julia S. El-Sayed Moustafa, et al. “Cell-Type Heterogeneity in Adipose Tissue Is Associated with Complex Traits and Reveals Disease-Relevant Cell-Specific eQTLs”. *American Journal of Human Genetics* 2019.

[91] Michael I. Love, Wolfgang Huber, and Simon Anders. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. *Genome Biology* 2014.

[92] Vikram Agarwal, George W. Bell, Jin-Wu Nam, et al. “Predicting effective microRNA target sites in mammalian mRNAs”. *eLife* 2015.

[93] Ioannis S. Vlachos, Maria D. Paraskevopoulou, Dimitra Karagkouni, et al. “DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions”. *Nucleic Acids Research* Database issue 2015.

[94] Arbel Harpak, Anand Bhaskar, and Jonathan K. Pritchard. “Mutation Rate Variation is a Primary Determinant of the Distribution of Allele Frequencies in Humans”. *PLoS Genetics* 2016.

[95] Jedidiah Carlson, Adam E. Locke, Matthew Flickinger, et al. “Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans”. *Nature Communications* 2018.

[96] François Aguet, Alvaro N. Barbeira, Rodrigo Bonazzola, et al. “The GTEx Consortium atlas of genetic regulatory effects across human tissues”. *bioRxiv* 2019.

[97] François Aguet, Andrew A. Brown, Stephane E. Castel, et al. “Genetic effects on gene expression across human tissues”. *Nature* 2017.

[98] Jeffrey A. Simon and Robert E. Kingston. “Occupying chromatin: Polycomb mechanisms for getting to genomic targets, stopping transcriptional traffic, and staying put”. *Molecular cell* 2013.

[99] *RFT1 RFT1 homolog [Homo sapiens (human)] - Gene - NCBI*. URL: `https://www.ncbi.nlm.nih.gov/gene/91869` (visited on 01/17/2021).

[100] Cassandra N. Spracklen, Momoko Horikoshi, Young Jin Kim, et al. “Identification of type 2 diabetes loci in 433,540 East Asian individuals”. *bioRxiv* 2019.

[101] Rengna Yan, Shanshan Lai, Yang Yang, et al. “A novel type 2 diabetes risk allele increases the promoter activity of the muscle-specific small ankyrin 1 gene”. *Scientific Reports* 2016.

[102] Tom G. Richardson, Eleanor Sanderson, Tom M. Palmer, et al. “Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: A multivariable Mendelian randomisation analysis”. *PLOS Medicine* 2020.

[103] Ayush Giri, Jacklyn N. Hellwege, Jacob M. Keaton, et al. “Trans-ethnic association study of blood pressure determinants in over 750,000 individuals”. *Nature Genetics* 2019.

[104] Josée Dupuis, Claudia Langenberg, Inga Prokopenko, et al. “New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk”. *Nature Genetics* 2010.

[105] Kajsa Sjöholm, Jenny Palming, Theodore C. Lystig, et al. “The expression of inhibin beta B is high in human adipocytes, reduced by weight loss, and correlates to factors implicated in metabolic disease”. *Biochemical and Biophysical Research Communications* 2006.

[106] R. Wijayarathna and D.M. de Kretser. “Activins in reproductive biology and beyond”. *Human Reproduction Update* 2016.

[107] Dylan D Thomas, Barbara E Corkey, Nawfal W Istfan, et al. “Hyperinsulinemia: An Early Indicator of Metabolic Dysfunction”. *Journal of the Endocrine Society* 2019.

[108] Yun-Hung Chen, Yu-Chien Lee, Yu-Chung Tsao, et al. “Association between high-fasting insulin levels and metabolic syndrome in non-diabetic middle-aged and elderly populations: a community-based study in Taiwan”. *BMJ open* 2018.

[109] C. Haas, E. Hanson, A. Kratzer, et al. “Selection of highly specific and sensitive mRNA biomarkers for the identification of blood”. *Forensic Science International: Genetics* 2011.

[110] Suaad Alshehhi and Penelope R. Haddrill. “Evaluating the effect of body fluid mixture on the relative expression ratio of blood-specific RNA markers”. *Forensic Science International* 2020.

[111] Carlos J. Pirola, Tomas Fernández Gianotti, Gustavo O. Castaño, et al. “Circulating microRNA signature in non-alcoholic fatty liver disease: from serum non-coding RNAs to liver histology and disease pathogenesis”. *Gut* 2015.

[112] Zhihong Zhu, Futao Zhang, Han Hu, et al. “Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets”. *Nature Genetics* 2016.

[113] Farhad Hormozdiari, Martijn van de Bunt, Ayellet V. Segrè, et al. “Colocalization of GWAS and eQTL Signals Detects Target Genes”. *American Journal of Human Genetics* 2016.

[114] Farhad Hormozdiari, Emrah Kostem, Eun Yong Kang, et al. “Identifying causal variants at loci with multiple signals of association”. *Genetics* 2014.

[115] Amanda Dobbyn, Laura M. Huckins, James Boocock, et al. “Landscape of Conditional eQTL in Dorsolateral Prefrontal Cortex and Co-localization with Schizophrenia GWAS”. *American Journal of Human Genetics* 2018.

[116] Gonzalo S. Nido, Fiona Dick, Lilah Toker, et al. “Common gene expression signatures in Parkinson’s disease are driven by changes in cell composition”. *Acta Neuropathologica Communications* 2020.

[117] Brandon Jew, Marcus Alvarez, Elior Rahmani, et al. “Accurate estimation of cell composition in bulk expression through robust integration of single-cell information”. *Nature Communications* 2020.

[118] Mee-Sup Yoon. “The Role of Mammalian Target of Rapamycin (mTOR) in Insulin Signaling”. *Nutrients* 2017.

[119] Xingjun Huang, Guihua Liu, Jiao Guo, et al. “The PI3K/AKT pathway in obesity and type 2 diabetes”. *International Journal of Biological Sciences* 2018.

[120] Shih-Yin Tsai, Ariana A. Rodriguez, Somasish G. Dastidar, et al. “Increased 4E-BP1 Expression Protects against Diet-Induced Obesity and Insulin Resistance in Male Mice”. *Cell Reports* 2016.

[121] N. Hoggard, M. Cruickshank, K. M. Moar, et al. “Inhibin betaB expression in murine adipose tissue and its regulation by leptin, insulin and dexamethasone”. *Journal of Molecular Endocrinology* 2009.

[122] Monkol Lek, Konrad J. Karczewski, Eric V. Minikel, et al. “Analysis of protein-coding genetic variation in 60,706 humans”. *Nature* 2016.

[123] Adam E. Locke, Karyn Meltz Steinberg, Charleston W. K. Chiang, et al. “Exome sequencing of Finnish isolates enhances rare-variant association power”. *Nature* 2019.

[124] Elina Salmela, Tuuli Lappalainen, Ingegerd Fransson, et al. “Genome-wide analysis of single nucleotide polymorphisms uncovers population structure in Northern Europe”. *PloS One* 2008.

[125] Jian Yang, Teresa Ferreira, Andrew P Morris, et al. “Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits”. *Nature genetics* 2012.

[126] Luís Henrique Corrêa, Gabriella Simões Heyn, and Kelly Grace Magalhaes. “The Impact of the Adipose Organ Plasticity on Inflammation and Cancer Progression”. *Cells* 2019.

[127] Alan Chait and Laura J. den Hartigh. “Adipose Tissue Distribution, Inflammation and Its Metabolic Consequences, Including Diabetes and Cardiovascular Disease”. *Frontiers in Cardiovascular Medicine* 2020. Publisher: Frontiers.

[128] Daniel C. Berry, Drew Stenesen, Daniel Zeve, et al. “The developmental origins of adipose tissue”. *Development (Cambridge, England)* 2013.

[129] Gideon R. Hajer, Timon W. van Haeften, and Frank L. J. Visseren. “Adipose tissue dysfunction in obesity, diabetes, and vascular diseases”. *European Heart Journal* 2008.

[130] Jan-Bernd Funcke, Julia von Schnurbein, Belinda Lennerz, et al. “Monogenic forms of childhood obesity due to mutations in the leptin gene”. *Molecular and Cellular Pediatrics* 2014.

[131] Gilberto Paz-Filho, Claudio Mastronardi, Ma-Li Wong, et al. “Leptin therapy, insulin sensitivity, and glucose homeostasis”. *Indian Journal of Endocrinology and Metabolism* Suppl 3 2012.

[132] Lavinia Woodward, Ioannis Akoumianakis, and Charalambos Antoniades. “Unravelling the adiponectin paradox: novel roles of adiponectin in the regulation of cardiovascular disease”. *British Journal of Pharmacology* 2017.

[133] Takashi Kadowaki, Toshimasa Yamauchi, Naoto Kubota, et al. “Adiponectin and adiponectin receptors in insulin resistance, diabetes, and the metabolic syndrome”. *The Journal of Clinical Investigation* 2006.

[134] Junghyo Jo, Oksana Gavrilova, Stephanie Pack, et al. “Hypertrophy and/or Hyperplasia: Dynamics of Adipose Tissue Growth”. *PLoS computational biology* 2009.

[135] Andrew D. Hildreth, Feiyang Ma, Yung Yu Wong, et al. “Single-cell sequencing of human white adipose tissue identifies new cell states in health and obesity”. *Nature Immunology* 2021.

[136] Michele Longo, Federica Zatterale, Jamal Naderi, et al. “Adipose Tissue Dysfunction as Determinant of Obesity-Associated Metabolic Complications”. *International Journal of Molecular Sciences* 2019.

[137] José J. Fuster, Noriyuki Ouchi, Noyan Gokce, et al. “Obesity-Induced Changes in Adipose Tissue Microenvironment and Their Impact on Cardiovascular Disease”. *Circulation Research* 2016.

[138] Pavankumar Patel and Nicola Abate. “Role of subcutaneous adipose tissue in the pathogenesis of insulin resistance”. *Journal of Obesity* 2013.

[139] Lindsay Fernández-Rhodes, Kristin L. Young, Adam G. Lilly, et al. “Importance of Genetic Studies of Cardiometabolic Disease in Diverse Populations”. *Circulation Research* 2020.

[140] Tim D. Spector and Frances M. K. Williams. “The UK Adult Twin Registry (TwinsUK)”. *Twin Research and Human Genetics: The Official Journal of the International Society for Twin Studies* 2006.

[141] Julia S. El-Sayed Moustafa, Anne U. Jackson, Sarah M. Brotman, et al. “ACE2 expression in adipose tissue is associated with COVID-19 cardio-metabolic risk factors and cell type composition”. *medRxiv* 2020.

[142] Raffaella Cancello, Corneliu Henegar, Nathalie Viguerie, et al. “Reduction of macrophage infiltration and chemoattractant gene expression changes in white adipose tissue of morbidly obese subjects after surgery-induced weight loss”. *Diabetes* 2005.

[143] Jian-Wen Han, Hou-Feng Zheng, Yong Cui, et al. “Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus”. *Nature Genetics* 2009.

[144] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, et al. “STAR: ultrafast universal RNA-seq aligner”. *Bioinformatics* 2013.

[145] Jennifer Harrow, Adam Frankish, Jose M. Gonzalez, et al. “GENCODE: the reference human genome annotation for The ENCODE Project”. *Genome Research* 2012.

[146] Alfonso Buil, Andrew Anand Brown, Tuuli Lappalainen, et al. “Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins”. *Nature Genetics* 2015.

[147] Mark D. Robinson and Alicia Oshlack. “A scaling normalization method for differential expression analysis of RNA-seq data”. *Genome Biology* 2010.

[148] Montserrat Esteve Ràfols. “Adipose tissue: cell heterogeneity and functional diversity”. *Endocrinologia Y Nutricion: Organo De La Sociedad Espanola De Endocrinologia Y Nutricion* 2014.

[149] Koichiro Taguchi, Kazuo Kajita, Yoshihiko Kitada, et al. “Role of small proliferative adipocytes: possible beige cell progenitors”. *The Journal of Endocrinology* 2020.

[150] Corbin Quick, Li Guan, Zilin Li, et al. “A versatile toolkit for molecular QTL mapping and meta-analysis at scale”. *bioRxiv* 2020.

[151] Halit Ongen, Alfonso Buil, Andrew Anand Brown, et al. “Fast and efficient QTL mapper for thousands of molecular phenotypes”. *Bioinformatics* 2016.

[152] Cristen J. Willer, Yun Li, and Gonçalo R. Abecasis. “METAL: fast and efficient meta-analysis of genomewide association scans”. *Bioinformatics* 2010.

[153] Yaowu Liu, Sixing Chen, Zilin Li, et al. “ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies”. *The American Journal of Human Genetics* 2019.

[154] Yaowu Liu and Jun Xie. “Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures”. *Journal of the American Statistical Association* 2020.

[155] Sara L Pulit, Charli Stoneman, Andrew P Morris, et al. “Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry”. *Human Molecular Genetics* 2019.

[156] Majid Nikpay, Anuj Goel, Hong-Hee Won, et al. “A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease”. *Nature Genetics* 2015.

[157] Ying Wu, K. Alaine Broadaway, Chelsea K. Raulerson, et al. “Colocalization of GWAS and eQTL signals at loci with multiple signals identifies additional candidate genes for body fat distribution”. *Human Molecular Genetics* 2019.

[158] Claudia Giambartolomei, Damjan Vukcevic, Eric E. Schadt, et al. “Bayesian test for colocalisation between pairs of genetic association studies using summary statistics”. *PLoS genetics* 2014.

[159] Yajuan Qi, Zihui Xu, Qinglei Zhu, et al. “Myocardial loss of IRS1 and IRS2 causes heart failure and is controlled by p38α MAPK during insulin resistance”. *Diabetes* 2013.

[160] Yun Wang, Patsy M. Nishina, and Jürgen K. Naggert. “Degradation of IRS1 leads to impaired glucose uptake in adipose tissue of the type 2 diabetes mouse model TALLYHO/Jng”. *The Journal of Endocrinology* 2009.

[161] Thomas J. Hoffmann, Elizabeth Theusch, Tanushree Haldar, et al. “A large electronic-health-record-based genome-wide study of serum lipids”. *Nature Genetics* 2018.

[162] Genevieve L. Wojcik, Mariaelisa Graff, Katherine K. Nishimura, et al. “Genetic analyses of diverse populations improves discovery for complex traits”. *Nature* 2019.

[163] Tom G. Richardson, Eleanor Sanderson, Tom M. Palmer, et al. “Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: A multivariable Mendelian randomisation analysis”. *PLoS medicine* 2020.

[164] Pietari Ripatti, Joel T. Rämö, Nina J. Mars, et al. “Polygenic Hyperlipidemias and Coronary Artery Disease Risk”. *Circulation. Genomic and Precision Medicine* 2020.

[165] Dragana Vuckovic, Erik L. Bao, Parsa Akbari, et al. “The Polygenic and Monogenic Basis of Blood Traits and Diseases”. *Cell* 2020.

[166] Lingyan Xiao, Dongyang Shi, Hui Zhang, et al. “Association between single nucleotide polymorphism rs11057401 of CCDC92 gene and the risk of coronary heart disease (CHD)”. *Lipids in Health and Disease* 2018.

[167] Tetsuya Saito, Tadao Shibasaki, and Susumu Seino. “Involvement of Exoc3l, a protein structurally related to the exocyst subunit Sec6, in insulin secretion”. *Biomedical Research (Tokyo, Japan)* 2008.

[168] Margaret R. Heider and Mary Munson. “Exorcising the Exocyst Complex”. *Traffic* 2012.

[169] Sonja I. Berndt, Stefan Gustafsson, Reedik Mägi, et al. “Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture”. *Nature Genetics* 2013.

[170] Dmitry Shungin, Thomas W. Winkler, Damien C. Croteau-Chonka, et al. “New genetic loci link adipose and insulin biology to body fat distribution”. *Nature* 2015.

[171] Tom G Richardson, Eleanor Sanderson, Benjamin Elsworth, et al. “Use of genetic variation to separate the effects of early and later life adiposity on disease risk: mendelian randomisation study”. *The BMJ* 2020.

[172] Gudmar Thorleifsson, G. Bragi Walters, Daniel F. Gudbjartsson, et al. “Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity”. *Nature Genetics* 2009.

[173] Torgny Karlsson, Mathias Rask-Andersen, Gang Pan, et al. “Contribution of genetics to visceral adiposity and its relation to cardiovascular and metabolic disease”. *Nature Medicine* 2019.

[174] Yalda Jamshidi, Harold Snieder, Dongliang Ge, et al. “The SH2B gene is associated with serum leptin and body fat in normal female twins”. *Obesity (Silver Spring, Md.)* 2007.

[175] S. Robiou-du-Pont, A. Bonnefond, L. Yengo, et al. “Contribution of 24 obesity-associated genetic variants to insulin resistance, pancreatic beta-cell function and type 2 diabetes risk in the French population”. *International Journal of Obesity (2005)* 2013.

[176] Anna-Lena Volckmar, Florian Bolze, Ivonne Jarick, et al. “Mutation screen in the GWAS derived obesity gene SH2B1 including functional analyses of detected variants”. *BMC medical genomics* 2012.

[177] Funda E. Orkunoglu-Suer, Brennan T. Harmon, Heather Gordish-Dressman, et al. “MC4R variant is associated with BMI but not response to resistance training in young females”. *Obesity (Silver Spring, Md.)* 2011.

[178] Liangyou Rui. “SH2B1 regulation of energy balance, body weight, and glucose metabolism”. *World Journal of Diabetes* 2014.

[179] Loic Yengo, Julia Sidorenko, Kathryn E. Kemper, et al. “Meta-analysis of genome-wide association studies for height and body mass index in 700000 individuals of European ancestry”. *Human Molecular Genetics* 2018.

[180] Tarunveer S. Ahluwalia, Bram P. Prins, Mohammadreza Abdollahi, et al. “Genome-wide association study of circulating interleukin 6 levels identifies novel loci”. *Human Molecular Genetics* 2021.

[181] *ATP2A1 ATPase sarcoplasmic/endoplasmic reticulum Ca2+ transporting 1 [Homo sapiens (human)] - Gene - NCBI*. URL: https://www.ncbi.nlm.nih.gov/gene/487 (visited on 04/13/2021).

[182] Sara Carmo-Silva, Clevio Nobrega, Luís Pereira de Almeida, et al. “Unraveling the Role of Ataxin-2 in Metabolism”. *Trends in endocrinology and metabolism: TEM* 2017.

[183] Karla P. Figueroa, Sadaf Farooqi, Kristopher Harrup, et al. “Genetic variance in the spinocerebellar ataxia type 2 (ATXN2) gene in children with severe early onset obesity”. *PloS One* 2009.

[184] Heng Jiao, Lingxiao Zeng, Shengsheng Yang, et al. “Knockdown EIF3C Suppresses Cell Proliferation and Increases Apoptosis in Pancreatic Cancer Cell”. *Dose-Response* 2020. Publisher: SAGE Publications Inc.

[185] Hsin-Yi Lee, Chi-Kuan Chen, Chun-Ming Ho, et al. “EIF3C-enhanced exosome secretion promotes angiogenesis and tumorigenesis of human hepatocellular carcinoma”. *Oncotarget* 2018.

[186] Samaneh Farashi, Thomas Kryza, Judith Clements, et al. “Post-GWAS in prostate cancer: from genetic association to biological contribution”. *Nature Reviews. Cancer* 2019.

[187] Daria V. Zhernakova, Patrick Deelen, Martijn Vermaat, et al. “Identification of context-dependent expression quantitative trait loci in whole blood”. *Nature Genetics* 2017.

[188] Yukie Kashima, Yoshitaka Sakamoto, Keiya Kaneko, et al. “Single-cell sequencing techniques from individual to multiomics analyses”. *Experimental & Molecular Medicine* 2020.

[189] Mgp van der Wijst, D. H. de Vries, H. E. Groot, et al. “The single-cell eQTLGen consortium”. *eLife* 2020.

[190] Rupali P. Patwardhan, Choli Lee, Oren Litvin, et al. “High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis”. *Nature Biotechnology* 2009.

[191] Alexandre Melnikov, Anand Murugan, Xiaolan Zhang, et al. “Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay”. *Nature Biotechnology* 2012.

[192] Rupali P. Patwardhan, Joseph B. Hiatt, Daniela M. Witten, et al. “Massively parallel functional dissection of mammalian enhancers in vivo”. *Nature Biotechnology* 2012.

[193] Yuchen Pan, Todd A. Duncombe, Colleen A. Kellenberger, et al. “High-Throughput Electrophoretic Mobility Shift Assays for Quantitative Analysis of Molecular Binding Reactions”. *Analytical Chemistry* 2014.

[194] Ronen Ben Jehuda, Yuval Shemer, and Ofer Binah. “Genome Editing in Induced Pluripotent Stem Cells using CRISPR/Cas9”. *Stem Cell Reviews and Reports* 2018.

[195] Dirk Hockemeyer and Rudolf Jaenisch. “Induced Pluripotent Stem Cells Meet Genome Editing”. *Cell Stem Cell* 2016.

[196] Alice B. Popejoy and Stephanie M. Fullerton. “Genomics is failing on diversity”. *Nature News* 2016. Section: Comment.

[197] Giorgio Sirugo, Scott M. Williams, and Sarah A. Tishkoff. “The Missing Diversity in Human Genetic Studies”. *Cell* 2019.

[198] Alicia R. Martin, Masahiro Kanai, Yoichiro Kamatani, et al. “Clinical use of current polygenic risk scores may exacerbate health disparities”. *Nature Genetics* 2019.

[199] Susanne B. Haga. “Impact of limited population diversity of genome-wide association studies”. *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 2010.

[200] Yun J. Sung, Thomas W. Winkler, Lisa de Las Fuentes, et al. “A Large-Scale Multi-ancestry Genome-wide Study Accounting for Smoking Behavior Identifies Multiple Significant Loci for Blood Pressure”. *American Journal of Human Genetics* 2018.

[201] Candelaria Vergara, Chloe L. Thio, Eric Johnson, et al. “Multi-Ancestry Genome-Wide Association Study of Spontaneous Clearance of Hepatitis C Virus”. *Gastroenterology* 2019.

[202] Ioanna Ntalla, Lu-Chen Weng, James H. Cartwright, et al. “Multi-ancestry GWAS of the electrocardiographic PR interval identifies 202 loci underlying cardiac conduction”. *Nature Communications* 2020.

[203] Zhaohui Du, Niels Weinhold, Gregory Chi Song, et al. “A meta-analysis of genome-wide association studies of multiple myeloma among men and women of African ancestry”. *Blood Advances* 2020.

[204] Lauren S. Mogil, Angela Andaleon, Alexa Badalamenti, et al. “Genetic architecture of gene expression traits across diverse populations”. *PLoS genetics* 2018.

[205] Lulu Shang, Jennifer A. Smith, Wei Zhao, et al. “Genetic Architecture of Gene Expression in European and African Americans: An eQTL Mapping Study in GENOA”. *American Journal of Human Genetics* 2020.

[206] Abhay Hukku, Milton Pividori, Francesca Luca, et al. “Probabilistic colocalization of genetic variants from complex and molecular traits: promise and limitations”. *American Journal of Human Genetics* 2021.

[207] Bernard Ng, Charles C White, Hans-Ulrich Klein, et al. “An xQTL map integrates the genetic architecture of the human brain’s transcriptome and epigenome”. *Nature neuroscience* 2017.

[208] Zhanye Zheng, Dandan Huang, Jianhua Wang, et al. “QTLbase: an integrative resource for quantitative trait loci across multiple human molecular phenotypes”. *Nucleic acids research* D1 2020.

[209] Jordana T Bell, Athma A Pai, Joseph K Pickrell, et al. “DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines”. *Genome biology* 2011. Type: Journal Article.

[210] Leopold Parts, Oliver Stegle, John Winn, et al. “Joint Genetic Analysis of Gene Expression Data with Inferred Cellular Phenotypes”. *PLOS Genetics* 2011. Publisher: Public Library of Science.

[211] Tuuli Lappalainen, Michael Sammeth, Marc R. Friedländer, et al. “Transcriptome and genome sequencing uncovers functional variation in humans”. *Nature* 2013.

[212] Eric R. Gamazon, Heather E. Wheeler, Kaanan P. Shah, et al. “A gene-based association method for mapping traits using reference transcriptome data”. *Nature Genetics* 2015.

[213] Ping Zeng, Ting Wang, and Shuiping Huang. “Cis-SNPs Set Testing and PrediXcan Analysis for Gene Expression Data using Linear Mixed Models”. *Scientific Reports* 2017.

[214] Alexander Gusev, Nicholas Mancuso, Hyejung Won, et al. “Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights”. *Nature Genetics* 2018.

[215] Nicholas Mancuso, Simon Gayther, Alexander Gusev, et al. “Large-scale transcriptome wide association study identifies new prostate cancer risk regions”. *Nature Communications* 2018.

[216] Hai Fang, ULTRA-DD Consortium, Hans De Wolf, et al. “A genetics-led approach defines the drug target landscape of 30 immune-related traits”. *Nature Genetics* 2019.

[217] Ioanna Tachmazidou, Konstantinos Hatzikotoulas, Lorraine Southam, et al. “Identification of new therapeutic targets for osteoarthritis through genome-wide analyses of UK Biobank data”. *Nature Genetics* 2019.

[218] Áine Duffy, Marie Verbanck, Amanda Dobbyn, et al. “Tissue-specific genetic features inform prediction of drug side effects in clinical trials”. *Science Advances* 2020.

[219] Hyun Min Kang, Jae Hoon Sul, Susan K Service, et al. “Variance component model to account for sample structure in genome-wide association studies”. *Nature genetics* 2010.

[220] Oliver Stegle, Leopold Parts, Matias Piipari, et al. “Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses”. *Nature protocols* 2012.

[221] Barbara E. Stranger, Stephen B. Montgomery, Antigone S. Dimas, et al. “Patterns of Cis Regulatory Variation in Diverse Human Populations”. *PLOS Genetics* 2012.

[222] Hyun Min Kang, Chun Ye, and Eleazar Eskin. “Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots”. *Genetics* 2008.

[223] Art B Owen and Jingshu Wang. “Bi-cross-validation for factor analysis”. *Statistical Science* 2016.

[224] Andrey Ziyatdinov, Miquel Vázquez-Santiago, Helena Brunel, et al. “lme4qtl: linear mixed models with flexible covariance structure for genetic studies of related individuals”. *BMC bioinformatics* 2018.

[225] Jakris Eu-Ahsunthornwattana, E. Nancy Miller, Michaela Fakiola, et al. “Comparison of methods to account for relatedness in genome-wide association studies with family-based data”. *PLoS genetics* 2014.

[226] Corbin Quick, Xiaoquan Wen, Gonçalo Abecasis, et al. “Integrating comprehensive functional annotations to boost power and accuracy in gene-based association analysis”. *PLOS Genetics* 2020.

[227] Shuang Feng, Dajiang Liu, Xiaowei Zhan, et al. “RAREMETAL: fast and powerful meta-analysis for rare variants”. *Bioinformatics* 2014.

[228] Gao Wang, Abhishek Sarkar, Peter Carbonetto, et al. “A simple new approach to variable selection in regression, with application to genetic fine mapping”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2020.

[229] Christian Benner, Aki S Havulinna, Marjo-Riitta Järvelin, et al. “Prospects of fine-mapping trait-associated genomic regions by using summary statistics from genome-wide association studies”. *The American Journal of Human Genetics* 2017.

[230] Christian Benner, Chris CA Spencer, Aki S Havulinna, et al. “FINEMAP: efficient variable selection using summary data from genome-wide association studies”. *Bioinformatics* 2016.

[231] Xiaowei Zhan, Youna Hu, Bingshan Li, et al. “RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data”. *Bioinformatics* 2016. Type: Journal Article.

[232] Xihao Li, Zilin Li, Hufeng Zhou, et al. “Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale”. *Nature Genetics* 2020.

[233] Y. Kim, K. Xia, R. Tao, et al. “A meta-analysis of gene expression quantitative trait loci in brain”. *Translational Psychiatry* 2014.

[234] G Guennebaud and B Jacob. “Eigen: a c++ linear algebra library. URL http://eigen. tuxfamily. org” 2014.

[235] Y Qiu. “Spectra C++ Library For Large Scale Eigenvalue Problems. URL https: //spectralib.org/” 2020.

[236] Petr Danecek, Adam Auton, Goncalo Abecasis, et al. “The variant call format and VCFtools”. *Bioinformatics* 2011.

[237] Richard A Gibbs, John W Belmont, Paul Hardenbol, et al. “The international HapMap project” 2003.

[238] Wei Zhang, Mark J Ratain, and M Eileen Dolan. “The HapMap resource is providing new insights into ourselves and its application to pharmacogenomics”. *Bioinformatics and biology insights* 2008.

[239] 1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, et al. “A global reference for human genetic variation”. *Nature* 2015.

[240] Michael E. Tipping and Christopher M. Bishop. “Probabilistic Principal Component Analysis”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 1999.

[241] Longda Jiang, Zhili Zheng, Ting Qi, et al. *A resource-efficient tool for mixed model association analysis of large-scale data*. Nature Publishing Group, 2019.

[242] Po-Ru Loh, George Tucker, Brendan K Bulik-Sullivan, et al. “Efficient Bayesian mixed-model analysis increases association power in large cohorts”. *Nature genetics* 2015.

[243] Han Chen and Matthew P Conomos. “GMMAT-package: Generalized Linear Mixed Model Association Tests” 2020.

[244] Stephanie M Gogarten, Tamar Sofer, Han Chen, et al. “Genetic association testing using the GENESIS R/Bioconductor package”. *Bioinformatics* 2019.

[245] Po-Ru Loh. “BOLT-LMM v2.3.4 User Manual.URL https://alkesgroup.broadinstitute.org/BOLT-LMM/downloads/BOLT-LMM_v2.3.4_manual.pdf” 2019.

[246] Pranav Yajnik and Michael Boehnke. “Power loss due to testing association between covariate-adjusted traits and genetic variants”. *Genetic Epidemiology* 2020.

[247] Konrad J. Karczewski, Laurent C. Francioli, Grace Tiao, et al. “The mutational constraint spectrum quantified from variation in 141,456 humans”. *Nature* 2020.

[248] Wouter Meuleman, Alexander Muratov, Eric Rynes, et al. “Index and biological spectrum of human DNase I hypersensitive sites”. *Nature* 2020.

[249] Antonio Fabio Di Narzo, Haoxiang Cheng, Jianwei Lu, et al. “Meta-eQTL: a tool set for flexible eQTL meta-analysis”. *BMC Bioinformatics* 2014.

[250] Yeji Lee, Francesca Luca, Roger Pique-Regi, et al. “Bayesian Multi-SNP Genetic Association Analysis: Control of FDR and Use of Summary Statistics”. *bioRxiv* 2018.

[251] Alkes L Price, Nick J Patterson, Robert M Plenge, et al. “Principal components analysis corrects for stratification in genome-wide association studies”. *Nature genetics* 2006.

[252] Christopher C. Chang, Carson C. Chow, Laurent Cam Tellier, et al. “Second-generation PLINK: rising to the challenge of larger and richer datasets”. *GigaScience* 2015.

[253] Christoph Lippert, Jennifer Listgarten, Ying Liu, et al. “FaST linear mixed models for genome-wide association studies”. *Nature methods* 2011.

[254] Gulnara R Svishcheva, Tatiana I Axenovich, Nadezhda M Belonogova, et al. “Rapid variance components–based method for whole-genome association analysis”. *Nature genetics* 2012.

[255] Susan Fairley, Ernesto Lowy-Gallego, Emily Perry, et al. “The International Genome Sample Resource (IGSR) collection of open human genomic variation resources”. *Nucleic Acids Research* D1 2020.

[256] *Fast and accurate long-range phasing in a UK Biobank cohort | Nature Genetics*. URL: `https://www.nature.com/articles/ng.3571` (visited on 02/05/2021).

[257] *ArrayExpress update – from bulk to single-cell expression data | Nucleic Acids Research | Oxford Academic*. URL: `https://academic.oup.com/nar/article/47/D1/D711/5144130` (visited on 01/04/2021).

[258] Charles Rotimi, Mark Leppert, Ichiro Matsuda, et al. “Community engagement and informed consent in the International HapMap project”. *Community Genetics* 2007.

[259] Latarsha J. Carithers, Kristin Ardlie, Mary Barcus, et al. “A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project”. *Biopreservation and Biobanking* 2015.

[260] Clare Bycroft, Colin Freeman, Desislava Petkova, et al. “The UK Biobank resource with deep phenotyping and genomic data”. *Nature* 2018.