EVALUATION OF PREDIXCAN FOR ASSOCIATING LIPIDS WITH GENES

AN EVALUATION OF THE PREDIXCAN METHOD FOR THE IDENTIFICATION OF LIPID ASSOCIATED GENES

By JOANNE E. I. GITTENS, B.SC., PH.D., M.B.A.

A thesis submitted in partial fulfilment of the requirements for
the degree of Master of Science

School of Graduate Studies
McMaster University
Hamilton, Ontario, Canada

Master of Science (2017) – Mathematics & Statistics

McMaster University, Hamilton, Ontario, Canada

TITLE:

An Evaluation of the PrediXcan Method for
the Identification of Lipid Associated Genes

AUTHOR:

Joanne E. I. Gittens
B.Sc. (University of Guelph),
Ph.D. (University of Western Ontario),
M.B.A. (Rotman School of Management, University of Toronto)

SUPERVISOR:

Professor Angelo J. Canty

NUMBER OF PAGES: 103

Abstract

PrediXcan, an imputed gene expression-trait association method, was compared to multiple linear regressions (MLR) of single nucleotide polymorphisms (SNPs) using the quantitative phenotypes serum total cholesterol (TC), low-density lipoprotein cholesterol (LDL), high-density lipoprotein cholesterol (HDL) and triglycerides (TG). The gene expression prediction models were trained using transcriptome- and genome-wide data from Depression Genes and Networks (DGN whole blood) and Genotype-Tissue Expression (GTEx) Project (GTEx whole blood, GTEx pancreas and GTEx liver). Linear combinations of the effect sizes derived using elastic net or least absolute shrinkage and selection operator (LASSO) with genotypes from 1304 European patients from the Diabetes Control and Complications Trial (DCCT) were used to estimate the genetically regulated expression (GReX) for genes. Different gene expression predictors were present in each training set. The 10-fold cross-validated predictive performance, estimated GReX, and p values from associations for matched genes were weakly correlated across training sets and strongly correlated for models derived using elastic net and LASSO. MLR models had more significant associations than PrediXcan models and larger inflation factors for p values. A comparison of p values for matched genes between PrediXcan and MLR models showed weak correlations but strong evidence for LDL and HDL associations with genes at locus 1p13.3 and 16q13, respectively.

Keywords

PrediXcan, gene-based association, single nucleotide polymorphism, lipids, expression quantitative trait loci, imputed gene expression, cholesterol, lipoproteins, triglycerides, type 1 diabetes, insulin-dependent diabetes mellitus, Diabetes Control and Complications Trial, regression analysis, linear models

To Maria and Nate

May you grow to love learning as much as I do.

Acknowledgements

I met with Dr. Narayanaswamy Balakrishnan early in 2015 to discuss the possibility of pursuing a graduate program in statistics and I thank him for not only accepting me into the program but also suggesting courses to take in the preceding months. Dr. Roman Viveros-Aguilera and Dr. Paul McNicholas furthered my curiosity of statistical theory and methods and I am grateful for their kind provision of time to discuss course material in greater depth.

With profound gratitude, I thank Dr. Angelo Canty for the opportunity to be his student and have an interesting applied statistics thesis. I cannot thank him enough for the thought-provoking questions, helpful suggestions, computational assistance, and careful review of my thesis. A special thank you to the past and present members of the SickKids Research Institute: Dr. Andrew Paterson, Dr. Delnaz Roshandel, Dr. Sareh Keshavarzi, and Ms. Chen Di Liao and Lunenfeld-Tanenbaum Research Institute at Mount Sinai Hospital: Dr. Shelley Bull for their constructive comments and assistance with data access. I also wish to thank Dr. Roman Viveros-Aguilera and Dr. Gregory Pond for serving on my thesis examination committee and my family for their love and support.

Contents

List of Tables

List of Figures

List of Abbreviations

IDDM	Insulin-Dependent Diabetes Mellitus
DCCT	Diabetes Control and Complications Trial
TG	Triglycerides
TC	Total Cholesterol
QTL	Quantitative Trait Locus
HDL	High-Density Lipoprotein Cholesterol
LDL	Low-Density Lipoprotein Cholesterol
GWAS	Genome-wide Association Study
LD	Linkage Disequilibrium
HWE	Hardy-Weinberg Equilibrium
SNP	Single Nucleotide Polymorphism
MAF	Minor Allele Frequency
LASSO	Least Absolute Shrinkage and Selection Operator
EN	Elastic Net
GTEx	Genotype-Tissue Expression
DGN	Depression Genes and Networks
GReX	Genetically Regulated Expression
CEPH	Centre d'Etude du Polymorphisme Humain

Chapter 1

Introduction

1.1 Genome-wide Association Studies

The genome is the genetic material of an individual and in humans it is comprised of 22 homologous autosomal chromosomes and a pair of sex chromosomes. The genetic instructions of a person are encoded in double stranded deoxyribonucleic acid (DNA) via sequences of nucleotides containing one of four nitrogenous bases: cytosine (C), guanine (G), adenine (A) and thymine (T). The human genome spans approximately 3.2 billion base pairs of protein coding and noncoding regions over approximately 20,000 genes (Ezkurdia et al., 2014). Genome-wide association studies (GWAS) explore the whole genome for DNA sequence variations termed single nucleotide polymorphisms (SNP) that are associated with a disease or trait (Edwards et al., 2005; Klein et al., 2005). Regions on a chromosome marked by a SNP (or gene) that correlate with a quantitative phenotype are called quantitative trait loci (QTL).

Homologous alleles may be homozygous (AA or aa) for the dominant (common) or recessive (rare) allele, respectively, or heterozygous (Aa) for both forms of the allele. The prevalence of certain variants segregates with populations, phenotypes and disease and can be quantified through calculations of the frequency of the rare (minor) allele (International HapMap 3 Consortium et al., 2010; 1000 Genomes Project Consortium

et al., 2010). Most common variants with minor allele frequency (MAF) of $\geqslant 5\%$ have small effects and thus marginally affect transcription, translation and subsequent downstream mechanisms. Hence, the common disorders that arise from these common SNPs are usually the result of many low penetrant variants (Manolio et al., 2009).

GWAS begins with the acquisition of high quality genome-wide data through the use of genotyping arrays and stringent quality control measures that help to minimize the GWAS false positive rate (Anderson et al., 2010). Linkage disequilibrium (LD), or the non-random inheritance of sets of SNPs on a chromosome in haplotype blocks, is exploited by GWAS (McCarthy et al., 2008). SNPs at the same or different loci are said to be in LD if their joint genotype distribution differs from the product of their marginal genotype distributions. In an effort to query the entire human genome without representing every SNP on an array, genotyping platforms include markers directly and indirectly using SNPs in LD. The level of correlation required to be tagged on an array is generally greater than 0.8 and the fraction of common SNPs that are captured determines the global coverage (Li et al., 2008). Rare variants with MAF $\leqslant 0.05$ (or 0.01) are normally excluded from analyses because statistical tests do not preform well for such low frequency SNPs (Carlson et al., 2003; Spencer et al., 2009).

Genotype calling algorithms are used to determine the three possible genotypes given the probe intensity for the SNPs in the sample. Genotype calls that do not meet the threshold—set to keep the call rate and number of errors within the range of tolerance—become missing calls because they cannot be accurately assigned the genotypes AA, Aa or aa. The total number of called genotypes and the quality of each call can vary by marker and sample (Anderson et al., 2010). Missing calls may suggest poor DNA quality or problems with the genotyping process. These samples or SNPs may be eliminated from the data set or retained by estimating

the posterior expected value for the missing markers using haplotypes from HapMap or 1000 genomes reference panels in a process called imputation (Li et al., 2009). Dosages for the SNPs are calculated from the called and imputed genotypes and they take on continuous values from 0 to 2. Dosages of 0 and 2 denote genotypes AA and aa, respectively. Sample contamination may be assessed through measures of the heterozygosity rate of the called genotypes (Anderson et al., 2010). Comparisons between the reproducibility of calls and the recorded and genotyped gender can further aid assessments of the genotype call integrity (Zeng et al., 2015).

GWAS results may be confounded by subgroup related factors. Tests for Hardy-Weinberg equilibrium (HWE) can assess population stratification (differences in the frequency of the minor allele among population members) due to genetically distinct subgroups in the sample (Anderson et al., 2010). The HWE model follows a binomial(2,MAF) distribution. Deviations from HWE can be assessed using Pearson's χ^2 test of the observed allele frequencies from the called genotypes and the expected HWE allele frequencies for the population (Balding, 2006). It is customary to remove SNPs showing strong evidence against HWE.

Principal component (PC) analysis is also widely used to identify individuals with different genetic backgrounds (Price et al., 2006; Anderson et al., 2010). Linearly uncorrelated PCs are calculated from the genotype matrix using singular value decomposition and reduced to the PCs of the population structure, in decreasing order of importance. The top two PCs, which describe ancestry effects, may be used to exclude or control for individuals with distinct PCs.

Linear regression models are used for quantitative (continuous) traits and they follow either a dominant (at least 1 minor allele is needed for disease risk; Aa or aa), recessive (both minor alleles are needed for disease risk; aa), multiplicative (Aa and aa confer a and a^2 disease risk, respectively) or additive (disease risk is proportional to

the number of copies of the minor allele) genetic model (Lettre et al., 2007). In single SNP GWAS, phenotypes are regressed onto each SNP individually and the estimate of the SNP-effect is tested for statistical significance and ranked according to the p value. The confidence of these estimates coincides with the sample size and thus the coefficients of the SNPs have smaller errors with larger sample sizes. Variation in complex traits is a result of genetic and environmental factors and the portion due to all of the causal genetic variants is termed heritability ($\hat{h}^2$) (Wray et al., 2013). The coefficient of determination (R^2) measures the amount of variation in the trait that is explained by the predictors in the model and in models with only genetic variants an upper bound is $\hat{h}^2$.

Multiple hypotheses are tested in GWAS and in single variant associations they may number $\geqslant 1$ million comparisons. Family-wise error rate (FWER) is the probability of making at least one type 1 error in a group of comparisons. The type 1 error becomes $1-(1-\alpha)^N$ for N approximately independent statistical tests unless the family of comparisons is corrected (Shaffer, 1995). Bonferroni correction (α/N) is commonly used in GWAS to control for family-wise error (Johnson et al., 2010) and hypotheses are generally tested at a genome wide significance level of $\sim 5 \times 10^{-8}$ (Risch et al., 1996). SNPs with p values less than this threshold are those selected for validation.

Quantile-quantile (Q-Q) plots facilitate the visualization of the distribution of p values from a GWAS study relative to the expected (uniformly distributed) p values under the hypothesis of no association, in rank order (Turner, 2014). The majority of variants follow the line $y = x$ since it is unlikely for many SNPs to be associated with the trait. The $-\log_{10}$ p value is commonly used to emphasize the large-effect loci that deviate from the diagonal at the far upper-right. Some variants will have small p values due to chance and thus it is only those p values that deviate sub-

stantially from the null distribution that are of particular interest; however, spurious associations may also have very small p values and thus claims of significance should follow from study replication and literature validation. Deviations from $y = x$ suggest quality control, cryptic relatedness (the unknown relationship between two or more persons) or population stratification problems, which can be evaluated by estimating the genomic control inflation factor for the distribution of p values (λ_{gc}) according to

$$\lambda_{gc} = \frac{\text{median}(w_1, w_2, ..., w_N)}{0.455}, \tag{1.1}$$

where $w_1, w_2, \cdots, w_N$ are the observed N (asymptotic) one-degree of freedom chi-squared test statistics and 0.455 is the expected median of the chi-squared distribution with one degree of freedom (Devlin and Roeder, 1999; Zheng et al., 2006). The λ_{gc} can also be calculated from p values converted to one degree of freedom chi-squared test statistics. The λ_{gc} should be close to one and a $\lambda_{gc} > 1$ suggests an inflation of p values and an increase in the type 1 error rate. Inflated test statistics can be corrected by dividing the test statistics by the inflation factor λ_{gc}. Histograms can also be used to assess the uniformity of p values. The results of GWAS are visualized using Manhattan plots (Turner, 2014) where SNPs are plotted by their genomic position and $-\log_{10}$ p value using a gradient of colours by chromosome number. The majority of insignificant variants cluster en masse at the base of the plot and the few significantly associated SNPs present at or above the Bonferroni corrected α level.

1.2 Gene-based Association Studies

While single-SNP analyses consider the marginal effect of a SNP on a phenotype, multi-marker associations capture the cumulative effect of SNPs that alone are either

not, weakly or moderately associated with the phenotype (Monir and Zhu, 2017). Consequently, multi-marker associations may explain more of the variation in the trait under the null hypothesis of no association because they contain $p > 1$ predictors in the model (Wray et al., 2013). They also result in fewer multiple comparisons and facilitate exploratory and secondary analyses of genome-wide data. In multi-marker associations, SNPs are aggregated into relevant sets that, in turn, are associated with a phenotype (Tregout et al., 2009). Relevance for inclusion in a set may be as simple as the strength of the p value from previous associations or as complex as the putative relationship of the SNPs to a biological pathway. Q-Q and Manhattan plots are used to visualize the distribution of p values in analogous fashion to those in GWAS and given that a set of SNPs is more likely to be causal than a single SNP, the customary Bonferroni corrections of $0.05/N$ (where N is now the number of SNP sets) more than adequately correct for multiple comparisons in multi-marker association studies. Similarly, calculations of the inflation factor for p values follow using the number of SNP sets for N.

In gene-based approaches, the relevant SNPs are those within and around a gene and in some cases these windows may overlap other gene windows. Thus, a SNP may appear in more than one statistical test for N genes, calling into question not only the assumption of independent hypotheses but also the suitability of the stringent Bonferroni multi-testing burden. In addition, the number of variants in a SNP set (gene size) can vary by gene and the SNPs within an aggregate can be correlated with one another (Mooney and Wilmot, 2015).

All of the variants of a gene need not be considered for association because the LD structure of genes can enable complete information to be obtained from only a subset of haplotype-tagged variants (Browning and Browning, 2007). The variable selection methods ridge, LASSO (least absolute shrinkage and selection operator) and

elastic net are commonly used in genetic association studies. They are based on the properties of ordinary least squares (OLS), where minimizing the residual squared error for a $(\mathbf{X}, \mathbf{y})$ data set with $n \times p$ matrix $\mathbf{X}$ results in the best linear unbiased estimator

$$\hat{\boldsymbol{\beta}} = \text{argmin}\Big\{ \sum_{i=1}^{n} \Big(y_i - \sum_j \beta_j x_{ij} \Big)^2 \Big\} = \text{argmin}||\boldsymbol{y} - \boldsymbol{X}\beta||^2 \tag{1.2}$$

however, the variance of coefficients increases with the dimensionality and collinearity of the data set and in certain circumstances may be too large. Ridge regression is a modelling method that reduces the variance of OLS estimates through a constraint on the coefficients

$$\sum_{j=1}^{p} \beta_j^2 = ||\boldsymbol{\beta}||^2 \leq t. \tag{1.3}$$

that shrinks them to continuous non-zero values. LASSO is a technique that combines the principles of ridge regression (shrinkage of coefficients) and variable selection (dropping regressors by setting coefficients to 0) to achieve a parsimonious model (Tibshirani, 1996) using the constraint on the coefficients

$$\sum_{j=1}^{p} |\beta_j| = ||\boldsymbol{\beta}||_1 \leq t. \tag{1.4}$$

Another continuos shrinkage and variable selection method termed elastic net (EN) combines the L^1-norm penalty of LASSO and L^2-norm penalty of ridge regression for $0 < \alpha < 1$ and places the following constraint on coefficients

$$\alpha||\boldsymbol{\beta}||_1 + (1-\alpha)||\boldsymbol{\beta}||^2 \leq t. \tag{1.5}$$

In contrast to LASSO, which indiscriminately selects a sparse set of SNPs from a group

of highly, pairwise correlated predictors and at most n variables when the number of explanatory variables p exceeds the number of observations n, EN selects groups of correlated variables (Zou and Hastie, 2005).

The accuracy of a predictive model can be assessed using cross-validation when an independent, external validation sample is not available. In cross-validation, a data set is equally partitioned into k random folds with one fold being the testing (validation) set and the other $k-1$ folds being the training (discovery) set. The average of k rounds of cross-validation, using different testing-training set combinations and every testing set exactly once, provides a good estimate of the predictive performance of the model. The estimated effect sizes $\hat{\beta}_i$ from the p variants selected from the discovery sample and one of $x_i = 0, 1, 2$ from the validation sample (Wray et al., 2013) can be used according to

$$\hat{y} = \sum_{i=1}^{p} \hat{\beta}_i x_i.$$

1.3 Transcriptome-wide Association Studies

The transcriptome contains all of the ribonucleic acids (RNA) from a cell type at a particular time, and the messenger RNAs (mRNAs) of the transcriptome signify the actively expressing genes in the cell type. Expression quantitative trait loci (eQTLs) are genomic regions that influence mRNA levels via the cis-regulation of their gene neighbours or the trans-regulation of distally located genes (Wittkopp and Kalay, 2012; Gilad et al., 2008), and polymorphisms in these regulatory regions affect transcript and protein abundance. Analyses of the relationships between SNPs and gene expression were conducted to understand the functional consequences of genetic variants (Lappalainen et al., 2013) and they are key intermediary steps to linking the genetics of gene expression to phenotypes (Albert and Kruglyak, 2015). Many

complex disorders associate with eQTL (Nicolae et al., 2010), and transcriptome-wide association studies (TWAS) seek to identify the relationships by associating the expression levels of tens of thousands of genes with traits. TWAS depends on the availability of gene expression data from the tissues of interest and due to the paucity of certain tissues some studies are not feasible. To circumvent this problem, groups developed methods to identify eQTL without directly measuring gene expression in a process called gene expression imputation (Gusev et al., 2016; Gamazon et al., 2015). Reference transcriptome and genome data sets of measured gene expression and genetic variation, respectively, were used to impute the cis-genetic component of expression in independent GWAS data sets. Gusev et al. (2016) used a Bayesian linear mixed model that performed shrinkage of the SNP effects but not variable selection. Gamazon et al. (2015) used regression models that shrunk coefficients and selected variables. Similar to gene-based association studies, regression methods, Bonferroni corrections, Q-Q and Manhattan plots, and genomic control for the inflation factor for p values are used to detect significant transcriptome-wide associations.

1.4 PrediXcan

PrediXcan is an imputed gene expression-trait association method that uses genome-wide data from individuals (Gamazon et al., 2015) rather than GWAS summary statistics (Gusev et al., 2016). All cis-acting common SNPs (MAF > 0.05) in HWE (p value > 0.05) from HapMap reference genomic data were aggregated into sets for autosomal genes if the SNPs were within 1 Mbp of the start and stop transcription sites. SNP sets were then regressed against the expression for the gene using reference transcriptome data sets to identify eQTL. Parsimonious additive genetic models were achieved

using elastic net and LASSO of the form

$$Y_g = \sum_k w_{k,g} X_k + \epsilon \tag{1.6}$$

where Y_g was the expression trait for gene g, $w_{k,g}$ was the effect size of variant k for gene g, X_k was the dosage of variant k, and ϵ represented other effectors of gene expression. SNPs associated with the expression of the gene were stored in the PredictDB data repository along with their effect sizes (weights). The PredictDB data repository contained training sets derived using elastic net and LASSO and reference transcriptome and genome data sets from 40 human tissue samples (and two transformed cells) across 24 organs. The reference data sets included the deceased donors of the Genotype-Tissue Expression (GTEx) Project (Lonsdale et al., 2013; GTEx Consortium et al., 2015) and living donors of the Depression Genes and Networks (DGN) (Battle et al., 2014). Approximately 66% of the GTEx donors were male: $\sim$ 84% were white and $\sim$ 14% were African American. The majority of GTEx donors were between 50-70 years of age and the cause of death for those between 60-71 years was heart (37.6%) and cerebrovascular (24.7%) disease (online resources: Gamazon et al. (2015)). DGN whole blood was from donors of European ancestry (Battle et al., 2014).

Gamazon and colleagues suggested that a linear combination of the dosages for a set of SNPs (X_k) with the weights of each variant ($\hat{w}_{k,g}$) as constants according to

$$\widehat{GReX}_g = \sum_k \hat{w}_{k,g} X_k \tag{1.7}$$

where $\widehat{GReX}_g$ was the estimated genetically regulated expression for a gene, could approximate the transcriptome of an individual from an independent genomic data set.

Furthermore, they proposed that $\widehat{GReX}_g$ could be associated with any phenotype of interest using regression and they illustrated their methods using expression predictors with a 10-fold cross-validated $R^2_{prediction} > 0.01$ from DGN whole blood derived using elastic net.

1.5 Lipids

Blood lipids include cholesterol (TC), low-density lipoprotein cholesterol (LDL), high-density lipoprotein cholesterol (HDL) and triglycerides (TG). Since high concentrations of TC and LDL and low levels of HDL are risk factors for cardiovascular disease many genetic association studies have searched for and discovered significantly associated QTLs (Despres et al., 2000; Prospective Studies Collaboration et al., 2007; Surakka et al., 2015; Ma et al., 2010; Zhang et al., 2015; Teslovich et al., 2010; Willer et al., 2013). Poorly controlled insulin dependent diabetes mellitus (IDDM) can present with high concentrations of LDL and TG, low concentrations of HDL, and cardiomyopathy induced heart failure (Vergs, 2009; Ritchie et al., 2017), suggesting underlying genetic determinants. Previous studies demonstrated a positive relationship with hyperglycaemia, dyslipidemia (Guy et al., 2009), and risk of heart failure in diabetic individuals (Iribarren et al., 2001; Boudina and Abel, 2007). Given the role of insulin in the regulation of lipid metabolism, it is not surprising that individuals with diabetes due to the autoimmune (or idiopathic) loss of the insulin producing islet of Langerhans β-cells in the pancreas have dyslipidemia without insulin therapy (Vergs, 2009).

1.6 Diabetes Control and Complications Trial (DCCT)

The Diabetes Control and Complications Trial (DCCT) was a randomized clinical trial involving 29 medical centres from the United States and Canada (DCCT Research Group et al., 1993). It was designed to test the hypothesis that an intensive insulin therapy regime would lead to significantly different glucose control outcomes and rates of appearance (or progression) of retinopathy and other IDDM-related diseases (DCCT Research Group et al., 1986). Two regimes (intensive and conventional) and two patient cohorts (primary prevention and secondary intervention; Table 1.1) were studied in 1441 IDDM (type 1 diabetic) patients recruited from 1983 to 1989. The primary prevention cohort included 726 patients without retinopathy at baseline and the secondary intervention cohort had 730 patients with mild-to-moderate non-proliferative retinopathy at baseline. Randomization was stratified by cohort, medical centre and age (13-17 years and 18-39 years) at entry and patients were part of the study for 3-9 years or 6.5 years on average. The characteristics of the patients are presented in Table 1.1. The intensive treatment group, with the glucose control outcome of maintaining close to normal blood glucose levels, received $\geqslant 3$ daily insulin injections (or continuous subcutaneous insulin infusion) to keep blood glucose at target glycemic values. The conventional treatment group, with the glucose control goal being sustained clinical well-being and no adverse diabetic events, received 1-2 daily insulin injections; glycemic targets were not set for the conventional treatment group but glycated hemoglobin HbA_{1c} (an assay of blood glucose concentration) was not permitted to exceed 13.11% (DCCT Research Group et al., 1993).

Table 1.1: Eligibility criteria for DCCT patients.

Age		13-39 years
HbA_{1c}		$> 6.6\%$
Serum creatinine		$\leqslant 1.2$ mg/dl
Basal C-peptide		< 0.2 nmol/l
Hypertension		No
Hypercholesterolemia		No
Severe Medical Conditions		No
	Primary Prevention	Secondary Intervention
Insulin therapy	1-5 years	1-15 years
Retinopathy (fundus)	No	$\geqslant 1$ microaneurysm (< level P2)
Albuminuria	< 40 mg/24h	< 200mg/24h

Blood glucose, HbA_{1c}, and blood pressure were measured quarterly for the conventional group and monthly for the intensive group. Stereo fundus photographs were taken biannually, and urinary albumin along with serum creatine, total cholesterol (TC), high-density lipoprotein cholesterol (HDL), triglycerides (TG) were collected annually following a $\geqslant 8$ h overnight fast (DCCT Research Group et al., 1986). Annual measures for LDL were calculated using Friedewald's formula (Friedewald et al., 1972). The visit completion rate exceeded 95%. The median for HbA_{1c} for the intensive treatment group fell to approximately 7% from 8.7% following 6 months of treatment and remained in this range for the remainder of the trial. The HbA_{1c} for the conventional treatment group rose from 8.5% to a sustained median value of approximately 9% after one year of treatment as shown in Figure 1.1. The results of the study showed that intensive treatment delayed the development and progression of retinopathy, nephropathy and neuropathy.

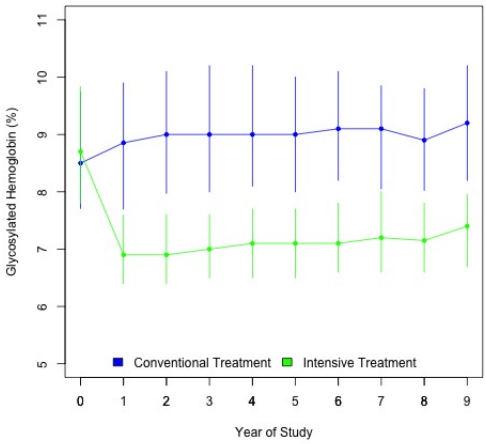

Figure 1.1: Medians of all annual glycosylated hemoglobin measurements in DCCT patients with IDDM receiving intensive or conventional therapy. Vertical lines mark the 25th and 75th percentiles of the yearly values. The conventional treatment measurements are shown in blue and the intensive treatment measurements are shown in green.

1.7 DCCT Data Set

DNA was collected from DCCT patients with their written informed consent and used for genome-wide genotyping with the Human1M beadchip (Illumina® Inc., San Diego, CA, USA). The Illumina Human1M beadchip was based on the HapMap II reference data set and 93% of the common SNPs from CEU (Utah residents with Northern and Western European ancestry from the CEPH collection) were tagged at $R^2 \geqslant 0.8$. Genotypes were called using BeadStudio/GenomeStudio software (Illumina® Inc., San Diego, CA, USA) and thereafter analyzed using PLINK v1.07 (http://pngu.mgh.harvard.edu/purcell/plink/). No data were removed because of a low genotype call rate; however, individuals with discrepancies between the reported and genotyped sex or previously reported genotypes were removed. The genotype concordance of 24 duplicate samples was 99.9995% (at a call rate threshold of

Table 1.2: 1304 Caucasian DCCT patients by treatment.

		Conventional	Intensive
Gender	Male	363	332
	Female	304	305
Cohort	Primary Prevention	344	307
	Secondary Intervention	323	330
Age	DCCT baseline	26.5 ± 7.1	27.2 ± 7.1

The sample size for each treatment group is presented across gender, cohort and age in years (Paterson et al., 2010).

0.988) and the mean heterozygosity across the genome for each individual was between 0.25-0.32. Two probands were removed following further tests for quality and cryptic relatedness. Autosomal SNPs were excluded from analyses if they deviated from HWE ($p < 10^{-8}$) or showed significant association with sex (Paterson et al., 2010). Untyped SNPs were imputed using methods and software described in (Howie et al., 2009) and the 1000 Genomes (phase 1 version 3) integrated variant release (March 2012). Only individuals who clustered with CEU and TSI (Toscani in Italy) from phase III of the International HapMap Project in PC analysis and thus were European Caucasians were used for GWAS, and only SNPs that were imputed with high certainty (INFO $\geqslant$0.8; IMPUTE version 2) were included in the study.

1.8 Rationale for the Thesis

The relationship between blood lipids, glucose homeostasis, IDDM and cardiovascular disease is well reported (Vergs, 2009). Furthermore, the mechanisms by which dyslipidemia and heart failure arise in IDDM are beginning to unfold along with the development of target-specific therapeutic agents (Siebel et al., 2015). GWAS have identified many QTLs (Teslovich et al., 2010; Kurano et al., 2016) but the identifi-

cation of eQTLs in TWAS has yet to be performed in the setting of IDDM and the models proposed for imputed gene expression-trait association require further validation.

The PredictDB data repository, DCCT data set, and PrediXcan model provide an excellent opportunity to examine the relationship between lipid traits and expression relevant SNPs at the individual level and on the genetic backdrop of IDDM. Exploring how different models impact the outcomes of TWAS can facilitate the development of efficient methods for this nascent field and aid the subsequent interpretation of statistical findings. Such investigation could help tease apart the genetic determinants of dyslipidemia in type 1 diabetes and enable progress in personalized medicine.

1.9 Objectives of the Thesis

1.9.1 Objective 1

The first objective, addressed in Chapter Two, was to describe and compare the characteristics of four publicly available PredictDB training sets (GTEx liver, GTEx pancreas, GTEx whole blood and DGN whole blood) derived using elastic net and LASSO. These training sets were selected because they relate to the etiology of IDDM.

1.9.2 Objective 2

The second objective, addressed in Chapter Three, was to describe the four lipids measured in DCCT and estimate the genetically regulated expression (GReX) of genes for European Caucasian DCCT patients. The estimated GReX values were compared across models and training sets to ascertain the impact of different variable selection methods and reference data sets on the estimated GReX. The estimated GReX were

associated with one of TC, LDL, $\sqrt{\text{HDL}}$, $\log_{10}$ TG, with and without covariates (age, gender, duration of IDDM, cohort, treatment, and the interaction between cohort and treatment). The p values from the Student's t for lipid associations with the estimated GReX were compared across models and training sets.

1.9.3 Objective 3

The final objective, addressed in Chapter Four, was to test for significant eQTL using multiple linear regression (MLR). Models with and without covariates were examined and the F-test was used to see if the lipid trait could be explained by one or more of the SNPs in the gene expression predictor. The MLR results were compared to the PrediXcan results of Chapter 3.

Chapter 2

PredictDB Training Sets

2.1 Methods

PredictDB training sets. PredictDB training sets were retrieved from the PredictDB data repository (http://hakyimlab.org/predictdb/) on March 1, 2016. The training sets: GTEx liver (n=97), GTEx pancreas (n=149), GTEx whole blood (n=338) and DGN whole blood (n=922) from elastic net and LASSO models (Appendices: Table A.1) were extracted using the DB Browser for SQLite Version 3.9.2 and read into R version 3.3.1 (2016-06-21) for descriptive analyses. Venn diagrams were constructed using package 'VennDiagram' from the CRAN repository (Chen and Boutros, 2011).

2.2 Results

DGN whole blood covered the most genes followed by GTEx whole blood, GTEx pancreas and GTEx liver in decreasing order for both elastic net and LASSO models (Table 2.1) and relative to the sample size used by Gamazon and colleagues to develop

Table 2.1: Summary statistics for the predictive performance and number of SNPs in the expression predictors across training sets.

Training Set	Model	Variable	N	Mean	St. Dev.	Min	Max
DGN whole blood	Elastic net	$R^2_{prediction}$	11,538	0.124	0.166	0.000	0.909
		SNPs	11,538	28.725	23.984	1	222
	LASSO	$R^2_{prediction}$	11,520	0.125	0.166	0.000	0.911
		SNPs	11,520	17.055	18.626	1	192
GTEx whole blood	Elastic net	R^2	10,215	0.046	0.088	0.000	0.769
		SNPs	10,215	16.805	16.094	1	128
	LASSO	$R^2_{prediction}$	10,067	0.047	0.089	0.000	0.767
		SNPs	10,067	11.729	12.632	1	362
GTEx pancreas	Elastic net	R^2	9,793	0.058	0.098	0.000	0.774
		SNPs	9,793	17.957	18.472	1	211
	LASSO	$R^2_{prediction}$	9,613	0.059	0.099	0.000	0.779
		SNPs	9,613	11.407	12.983	1	164
GTEx liver	Elastic net	R^2	8,561	0.049	0.079	0.000	0.669
		SNPs	8,561	16.928	19.062	1	167
	LASSO	$R^2_{prediction}$	8,402	0.051	0.081	0.000	0.679
		SNPs	8,402	10.379	12.664	1	136

$R^2_{prediction}$, 10-fold cross-validated predictive performance; SNPs, the number of SNPs in the expression predictor; N, the number of expression predictors in the training set; LASSO, least absolute shrinkage and selection operator.

the predictor models. In addition, DGN whole blood had the highest mean and max 10-fold cross-validated R^2 for predictive performance ($R^2_{prediction}$). GTEx whole blood had the lowest mean $R^2_{prediction}$ despite the fact that it was trained using the second largest sample size. The $R^2_{prediction}$ for matched expression predictors were similar between elastic net and LASSO models for all DGN whole blood and GTEx training sets (Figure 2.1) and thus the elastic net and LASSO models for gene expression predicted equally well. A comparison of the $R^2_{prediction}$ between pairs of training sets demonstrated weak correlations for matched expression predictors from all training set pairs except DGN whole blood and GTEx whole blood, which had moderate correlations (Figure 2.2).

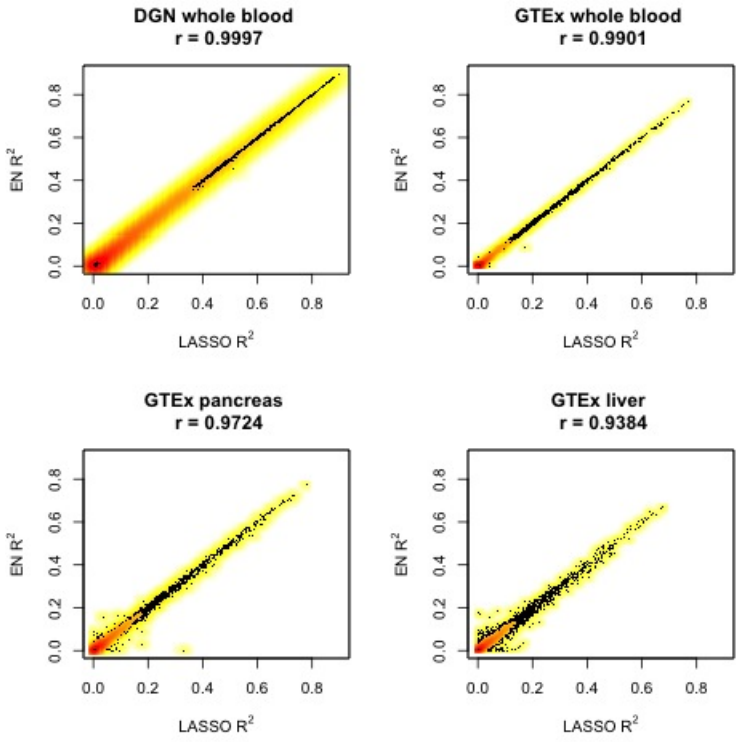

Figure 2.1: Scatter plots of the 10-fold cross-validated R^2 for predictive performance for matched expression predictors derived using EN (elastic net; y-axis) and LASSO (least absolute shrinkage and selection operator; x-axis). Spearman's rank correlation coefficient (r) is shown. The red line represents $y = x$.

Thus, the performance of an expression predictor varied with the training set selected and DGN whole blood did not always have the strongest predictive performance for gene expression.

In accordance with the selection of groups of correlated variables, elastic net models contained on average more SNPs in the expression predictor than LASSO models for all training sets (Figure 2.3), demonstrated by the higher point density in the upper left half of the elastic net versus LASSO scatter plots.

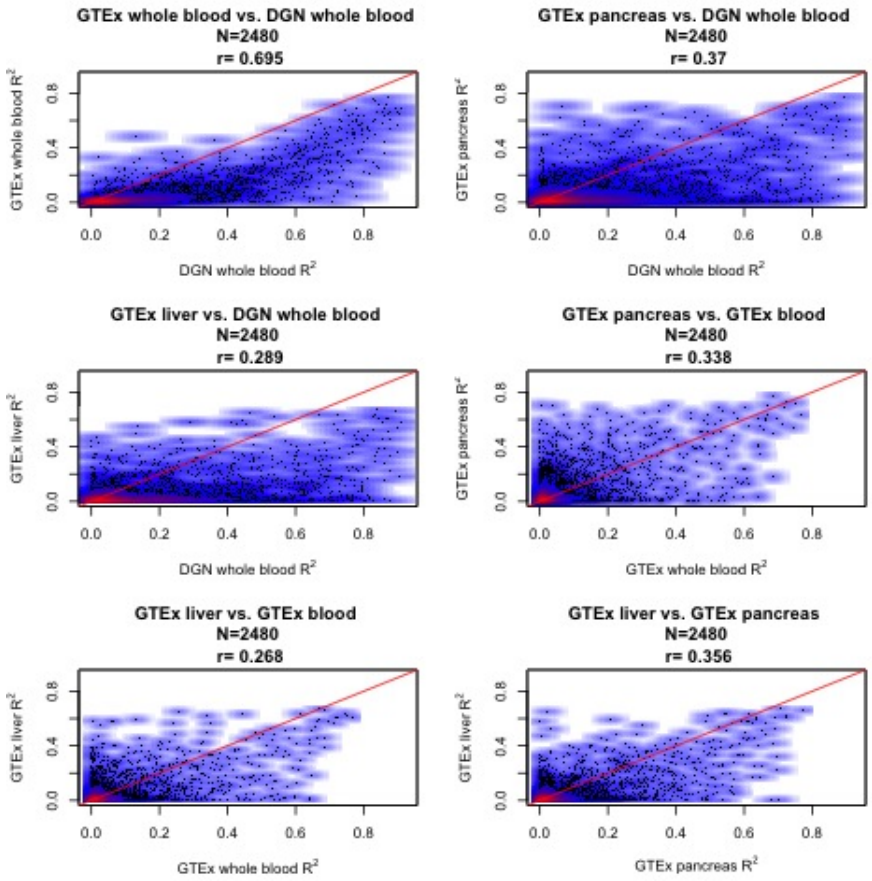

Figure 2.2: Scatter plots of the 10-fold cross-validated R^2 for predictive performance for matched expression predictors from training sets. The same expression predictors N were examined in each plot. Spearman's rank correlation coefficient (r) is shown. The red line represents $y = x$.

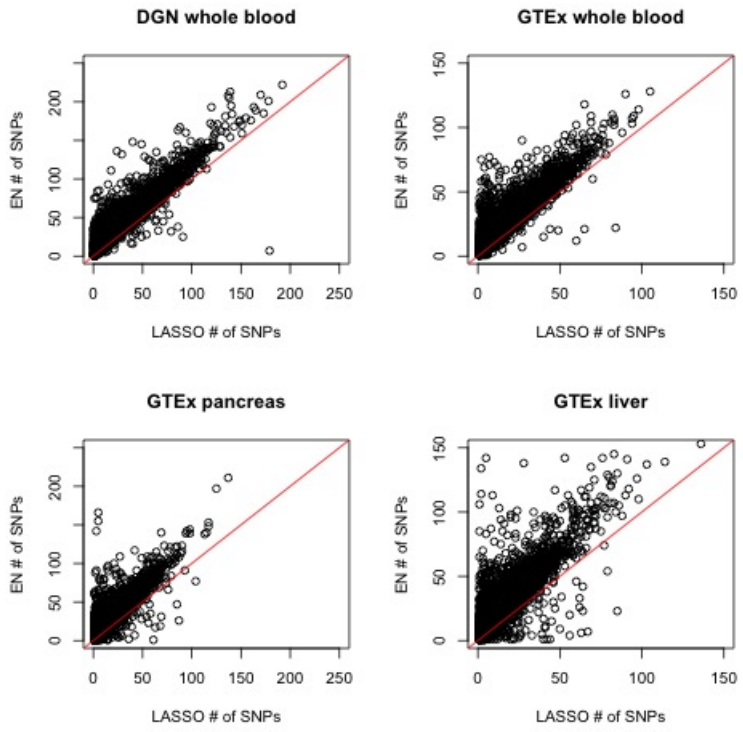

Figure 2.3: Scatter plots of the number of SNPs in each expression predictor derived using EN (elastic net; y-axis) and LASSO (least absolute shrinkage and selection operator; x-axis). The red line represents $y = x$.

The fact that elastic net models had more SNPs was further demonstrated in Table 2.2 and Figure 2.4, with DGN whole blood expression predictors containing more SNPs than the other GTEx training set expression predictors. The disparity between the number of SNPs in Tables 2.2 between total and unique SNPs was due to the multiple copies of the same SNP that existed in every PredictDB training set. A close inspection of the prevalence of multiple SNP copies is shown in Table 2.3, and it is a consequence of the assigned overlap of SNPs between neighbouring genes.

Table 2.2: Number of SNPs, unique SNPs, and unique expression predictors in the PredictDB training sets.

Training Set	Model	Total SNPs	Unique SNPs	Unique Genes
DGN whole blood	Elastic net	331,425	249696	11538
	LASSO	224,632	173624	13171
GTEx whole blood	Elastic net	171,659	152042	10212
	LASSO	118,078	106155	10064
GTEx pancreas	Elastic net	175,852	156979	9788
	LASSO	109,658	99960	9609
GTEx liver	Elastic net	144,920	131300	8558
	LASSO	87,206	80303	8400

LASSO, least absolute shrinkage and selection operator.

At most 9 gene expression predictors contained the same SNP in GTEx training sets and as many as 26 gene expression predictors contained the same SNP in DGN whole blood. The range of SNP weights varied by training set, being narrower for DGN whole blood and GTEx whole blood. GTEx pancreas and GTEx liver had extreme SNP weights for less than 0.2% of the variants (Table 2.4).

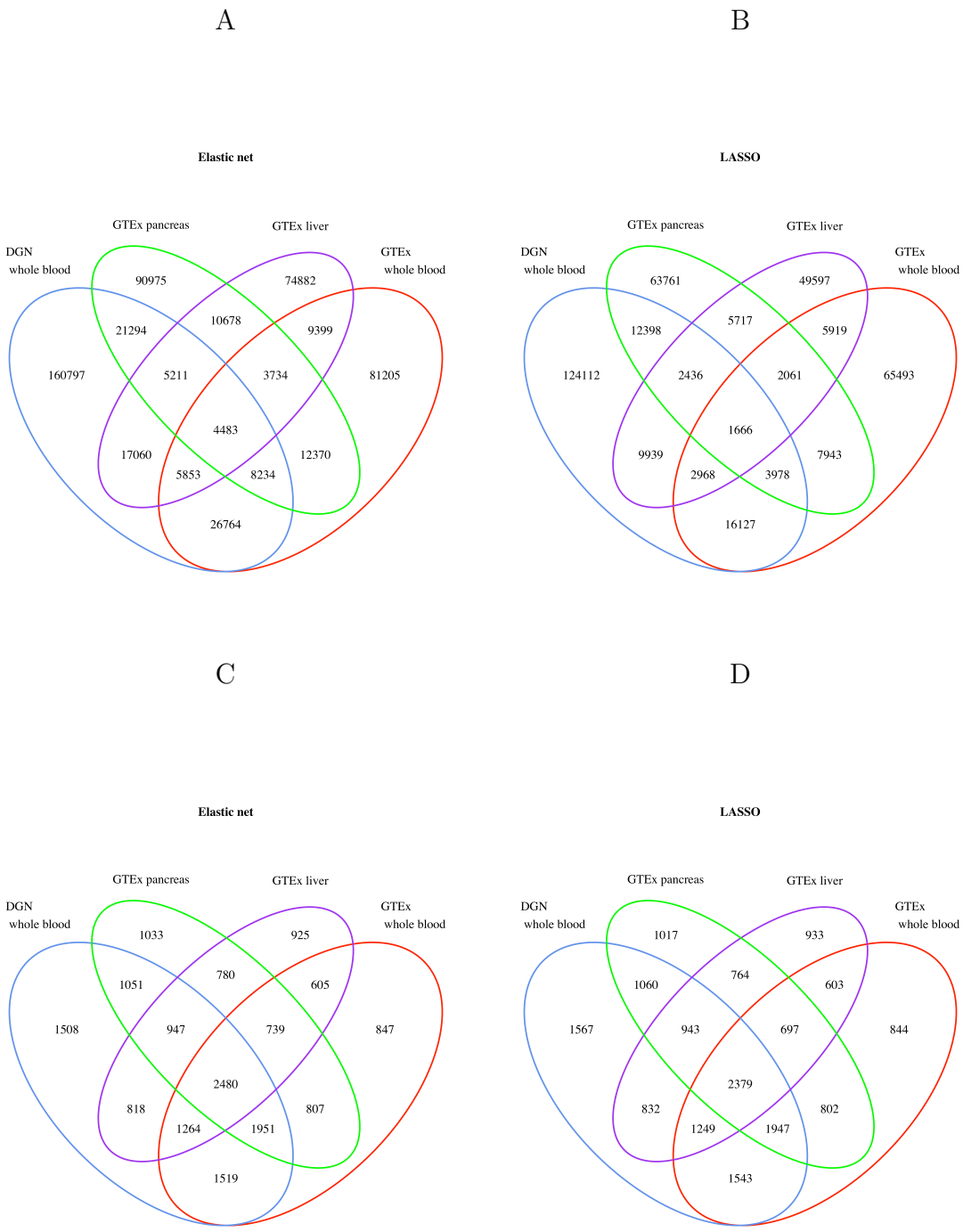

Figure 2.4: Venn diagrams for the number of unique SNPs (A and B) and unique gene expression predictors (C and D) across training sets. LASSO, least absolute shrinkage and selection operator

Table 2.3: Count of SNPs in PredictDB training sets that were present in one or more expression predictors.

SNP Copy	Training Set							
	GTEx liver		GTEx pancreas		GTEx whole blood		DGN whole blood	
	EN	LASSO	EN	LASSO	EN	LASSO	EN	LASSO
1	119712	74321	141073	91675	134925	96253	195925	139824
2	9977	5218	13624	7191	13491	8378	37574	23901
3	1299	638	1798	866	2039	1179	10225	6225
4	229	99	363	169	398	247	3380	2070
5	61	24	74	40	118	64	1350	834
6	18	2	25	12	49	17	566	345
7	4	1	14	2	18	15	268	175
8			5	4	3	1	152	87
9			3	1	1	1	92	53
10							46	23
11							35	27
12							16	14
13							16	13
14							12	6
15							9	7
16							9	5
17							8	6
18							2	3
19							5	2
20							2	2
21							1	1
22							1	
23							1	
24								
25								1
26							1	

The majority of SNPs in PredictDB training sets were present in only one expression predictor. One SNP appeared in 26 expression predictors from DGN whole blood derived using elastic net. EN, elastic net; LASSO, least absolute shrinkage and selection operator.

A few expression predictors had two IDs for a single Human Genome Organization (HUGO) symbol and thus two values for $R^2_{prediction}$ and the number of SNPs. Two expression predictors in GTEx liver ($CCDC177$, $GOLGA6L9$), four expression

Table 2.4: Summary statistics for SNP weights in the training sets.

Training Set	Model	N	Percentile		St. Dev	Min	Max
			0.1th	99.9th			
DGN whole blood	Elastic net	331425	-0.54	0.57	0.07	-2.64	2.23
	LASSO	224632	-0.89	0.96	0.11	-2.96	3.66
GTEx whole blood	Elastic net	171659	-0.57	0.55	0.09	-4.43	4.98
	LASSO	118078	-0.83	0.81	0.19	-37.83	26.69
GTEx pancreas	Elastic net	175852	-0.90	0.91	3.06	-975.98	350.19
	LASSO	109658	-1.26	1.23	4.39	-1163.52	393.79
GTEx liver	Elastic net	144920	-1.87	1.74	4.90	-817.06	628.24
	LASSO	87206	-3.79	3.06	6.58	-887.94	982.94

N, the number of SNPs in the training set; LASSO, least absolute shrinkage and selection operator.

predictors in GTEx pancreas ($F11R$, $GOLGA6L9$, $MR0H7$, $TMEM236$), and three expression predictors in GTEx whole blood ($DCAF8$, $F11R$, $ZNF763$) were duplicated in elastic net and LASSO models and one expression predictor, $FAM47E$, was duplicated in only GTEx liver and GTEx pancreas of elastic net. DGN whole blood training sets did not contain any duplications. Counts of the unique expression predictors (having only one ID and HUGO symbol) are presented in Table 2.2.

It may be tempting to select only DGN whole blood derived using elastic net for subsequent calculations of the score for gene expression, $\widehat{GReX}_g$, and association with a trait since it covered the most genes, had the largest number of SNPs, and the highest $R^2_{prediction}$ of the training sets. Such an approach is reasonable since there was expression predictor and SNP overlap among the training sets (Figure 2.4), and the majority of expression predictors and SNPs in LASSO models were also present in elastic net models. Moreover, proportionally more expression predictors intersected elastic net and LASSO for DGN whole blood relative to the other training sets (Table 2.5). Nonetheless, the extent of overlap paled in comparison to the collective number

Table 2.5: The number of gene expression predictors and SNPs within elastic net but not LASSO (EN), elastic net and LASSO (EN and LASSO), and LASSO but not elastic net (LASSO) across training sets.

Training Set	Variable	EN	EN and LASSO	LASSO
DGN whole blood	Genes	27	11511	9
	SNPs	78229	171467	2157
GTEx whole blood	Genes	200	10012	52
	SNPs	47647	104395	1760
GTEx pancreas	Genes	316	9472	137
	SNPs	61336	95643	4317
GTEx liver	Genes	350	8208	192
	SNPs	56553	74747	5556

EN, elastic net; LASSO, least absolute shrinkage and selection operator.

of expression predictors (or SNPs) present in training sets other than DGN whole blood. Thus, associations studies using more than one Predict DB training set in the calculation of the estimated genetically regulated expression for a gene may prove beneficial to the discovery of eQTL following the appropriate Bonferroni correction for the total number of tests.

Chapter 3

Lipid associations using PrediXcan

3.1 Methods

Lipids. The quantitative serum traits total cholesterol (TC), high-density lipoprotein cholesterol (HDL) and triglycerides (TG) were collected following an overnight, 8 hour fast at baseline and annually for the following 9 years (DCCT Research Group et al., 1986). LDL was estimated using Friedewald's formula

$$LDL = TC - HDL - \frac{TG}{5}$$

as described in Friedewald et al. (1972). The calculated LDL of fifteen measurements from 13 patients with TG concentrations in excess of 400 mg/dl were set to missing because Friedewald's formula does not reliably estimate LDL when TG concentrations are high. Lipid concentrations larger than the (one-tailed) 99.5 percentile—TC (293 mg/dl), LDL (207 mg/dl), HDL (94 mg/dl), TG (344.58 mg/dl) were set to the lipid concentrations previously listed. Thirty-eight out of 8172 (TC), 39 out of 8157 (LDL), 38 out of 8172 (HDL), and 41 out of 8172 (TG) extreme values were set to

the 99.5 percentile. In addition, HDL was transformed via the square root and TG was transformed via the common (base 10) logarithm. A total of at most 9 annual lipid measurements were averaged for each individual to determine the mean lipid measurement per patient. Baseline lipid measurements were not used in the averaged values.

Estimating GReX for genes for DCCT patients. DCCT patient dosages were extracted using SNPs from DGN whole blood, GTEx whole blood, GTEx pancreas, and GTEx liver derived from elastic net (EN) or least absolute shrinkage and selection operator (LASSO) according to scripts presented in the appendicies. The estimated genetically regulated expression (GReX) was determined according to

$$\widehat{GReX}^{M}_{g,T} = \sum_{k \in S^{M}_{g,T}} \hat{w}^{M}_{k,g,T} X_k \tag{3.1}$$

with model M: EN or LASSO, training set T: DGN whole blood, GTEx whole blood, GTEx pancreas, or GTEx liver, $S^{M}_{g,T}$: set of SNPs in the expression predictor for gene g from T and M, $\hat{w}^{M}_{k,g,T}$: the estimated weight for variant k from g, T and M, and X_k: the dosage of SNP k. Genotypes were put in terms of the PredictDB effect alleles.

Lipid Associations. For each of the four lipid traits: TC, LDL, $\sqrt{\text{HDL}}$ or $\log_{10}$ TG the mean across time was calculated for each individual and used as the phenotype. Each phenotype was then associated with $\widehat{GReX}^{M}_{g,T}$ according to the simple linear regression model

$$Y_{lipid} = \beta^{M}_{0,g,T} + \beta^{M}_{1,g,T} \widehat{GReX}^{M}_{g,T} + \epsilon^{M}_{g,T} \tag{3.2}$$

to see if the estimated GReX for a gene accounts for the variability in the lipid trait.

The Student's t with 1302 degrees of freedom was used to test

$$H_0: \quad \beta^M_{1,g,T} = 0 \quad \text{versus} \quad H_1: \quad \beta^M_{1,g,T} \neq 0.$$

The contribution of the $\widehat{GReX}_g$ to the lipid trait after adjustment for the covariates age (C_1), duration of IDDM (C_2), gender (C_3), cohort (C_4), treatment (C_5) and the interaction between cohort and treatment (C_6) was tested using the Student's t with 1296 degrees of freedom and the linear regression model

$$Y_{lipid} = \beta^M_{0,g,T} + \beta^M_{1,g,T}\widehat{GReX}^M_{g,T} + \gamma_1 C_1 + \cdots + \gamma_6 C_6 + \epsilon^M_{g,T}. \tag{3.3}$$

Lipids were regressed against the covariates in the models

$$Y_{lipid} = \gamma_0 + \gamma_1 C_1 + \cdots + \gamma_6 C_6 + \epsilon$$

to determine the proportion of the variance in the lipid trait that can be explained by the covariates alone. Each model assumed that the ϵ's were independent and normally distributed with $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2$. The binary variables were gender—1 for male and 2 for female, cohort—0 for primary prevention and 1 for secondary intervention, and treatment—0 for conventional and 1 for intensive. The lm function in R version 3.3.0 (2016-05-03) was used to fit the linear models.

A Bonferroni significance threshold of 1.53×10^{-7} was used to account for the collective number of tests across models and training sets (Figure 3.1) and those tests meeting or exceeding the threshold were considered significant. A less stringent (within training set) Bonferroni correction of 4.33×10^{-6} highlighted suggestive associations. Manhattan plots were used to visualize the significant (red line) and suggestive (blue line) p-values from the linear regressions and Q-Q plots and histograms were

used to observe the distribution of p-values. The genomic control inflation factor for p values, λ_{gc}, was calculated for the number of tests N in each training set. Scatter plots were used to compare the $-\log p$ values from $\widehat{GReX}_g^{EN}$ and $\widehat{GReX}_g^{LASSO}$ lipid associations. Scatter plots were also used to compare models 3.2 and 3.3 (without and with covariates). Internally studentized residuals were used to check the linearity and normality assumptions of tests that were significant. Points with leverage values greater than $2(p+1)/n$ were considered to have high leverage. Cook's distance was used to measure the influence of the ith observation. Values greater than the 50% point of the F distribution with degrees of freedom $p+1$ and $n-p-1$ were regarded as influential (Chatterjee and Hadi, 2006). In all cases the sample $n = 1304$ and the number of explanatory variables $p = 1$ or 7. Manhattan and Q-Q plots were constructed using qqman (Turner, 2014) and λ_{gc} was determined using GenABEL (Aulchenko et al., 2007).

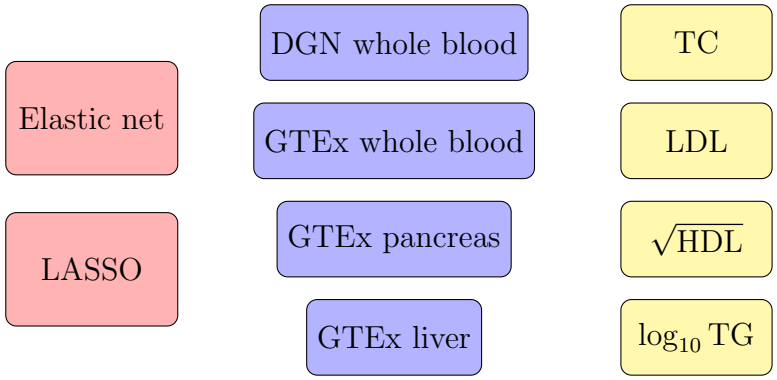

Figure 3.1: Overview of the models (red), training sets (blue) and phenotypes (yellow) used for associations. All possible combinations of one from each colour ($2 \times 4 \times 4 = 32$ models). A significance threshold of 1.53×10^{-7} represented a Bonferroni correction of $0.05/32N$ where $N = 10,215$ expression predictors. TC, total cholesterol; LDL, low-density lipoprotein cholesterol; HDL, high-density lipoprotein cholesterol; TG, triglycerides

3.2 Results

3.2.1 Phenotypes

To model the dependence of lipid traits on the scores for the estimated genetically regulated expression, the mean across time for each of the four lipid traits was calculated for the 1304 DCCT patients of European ancestry. The distribution of patients across the measurements (as shown in the table of Figure 3.2) reflected the traits TC, $\sqrt{\text{HDL}}$, and log_{10}TG. Eight of the patients had fewer than four annual measurements for all phenotypes. The distribution of the mean by annual measure is shown at the bottom of Figure 3.2 and summary statistics for the phenotypes are presented in Table 3.1. Each of the mean lipids followed a normal distribution after winsorization to the 99.5 percentile and transformation on the square root and $\log_{10}$ scale for HDL and TG, respectively (Figure 3.3). Mean lipids by cohort, treatment and gender were similar across categories except for the $\sqrt{\text{HDL}}$ trait which showed higher median levels for females in agreement with other studies (Davis et al., 1996). The covariates accounted for more of the variation in $\sqrt{\text{HDL}}$ relative to the other lipid traits (Table 3.2) and the p value for the gender coefficient was 3.07×10^{-47}. The coefficients of age and cohort were also significant at $p = 4.48 \times 10^{-8}$ and $p = 0.00061$, respectively. Mean serum lipids for DCCT patients in the conventional and intensive treatment regimes were similar (Figure 3.4) despite differences in the extent of glycemic control (Figure 1.1).

Measurements	1	2	3	4	5	6	7	8	9
Patients	3	3	2	154	348	281	212	78	223

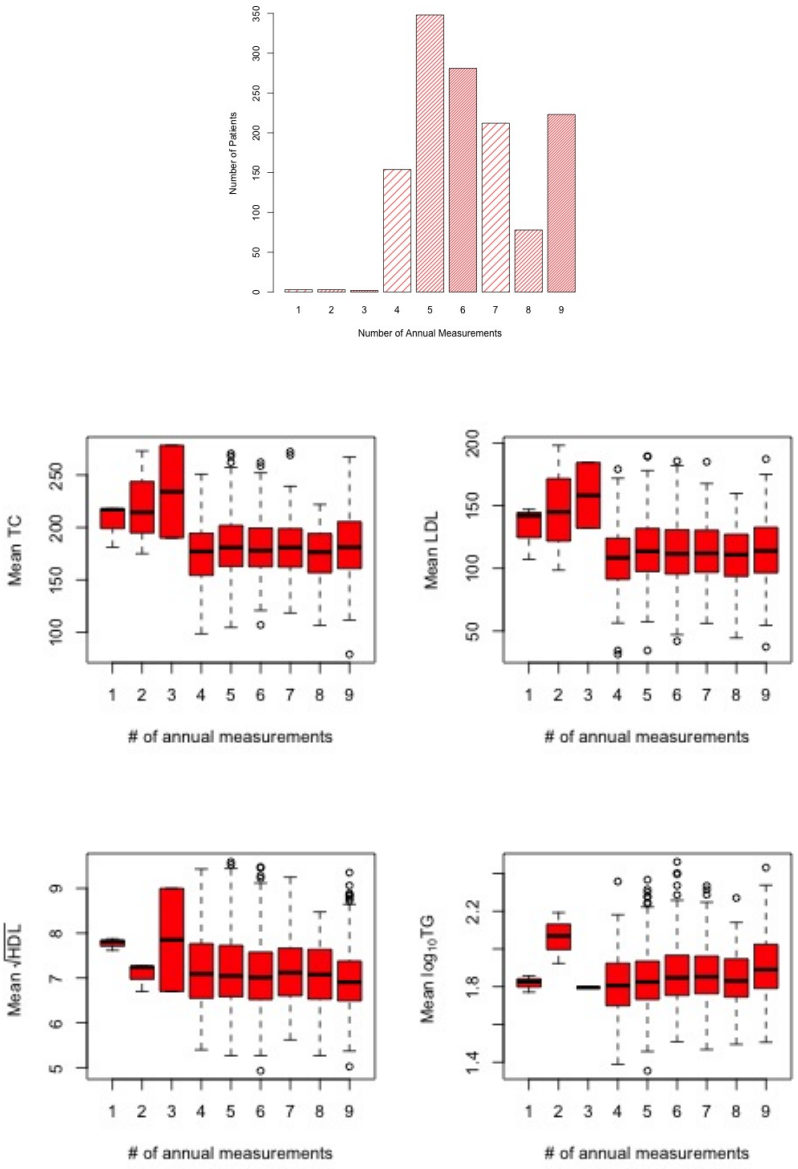

Figure 3.2: Distribution of patient mean lipids. Mean represents the average of all measurements for a patient (excluding baseline measurements) for each lipid trait.

Table 3.1: Summary statistics for the lipid traits.

Lipid	Mean	St. Dev.	Min	Max
TC	181.158	29.001	78.778	278.333
LDL	113.463	25.954	31.500	198.000
$\sqrt{\text{HDL}}$	7.114	0.787	4.932	9.589
$\log_{10}$ TG	1.859	0.164	1.354	2.461

TC, total cholesterol; LDL, low-density lipoprotein cholesterol; HDL, high-density lipoprotein cholesterol; TG, triglycerides.

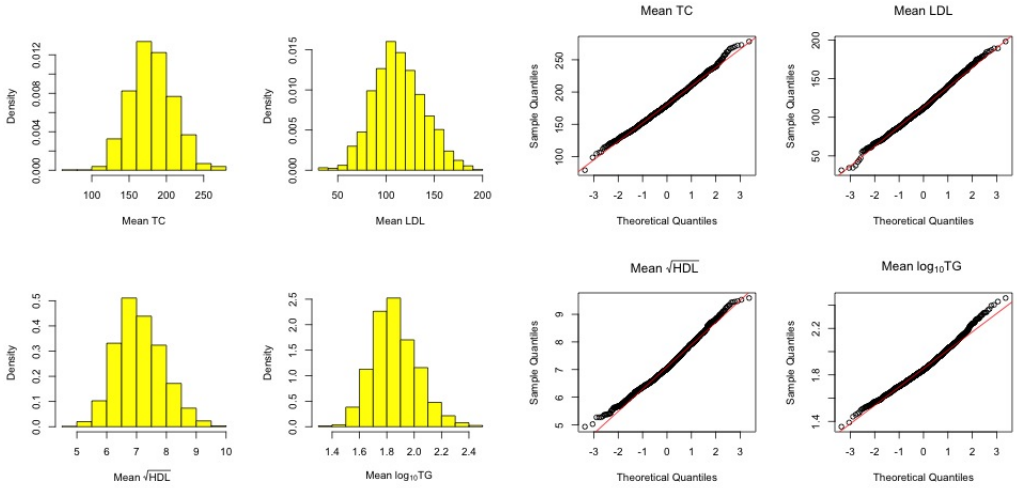

Figure 3.3: Distribution of the patient mean lipid traits used for associations. Theoretical quantiles are from a normal distribution. TC, total cholesterol; LDL, low-density lipoprotein cholesterol; HDL, high-density lipoprotein cholesterol; TG, triglycerides.

Table 3.2: The R^2, F statistic and p value for models of mean lipid traits regressed against covariates.

Lipid	R^2	F	P
TC	0.054	12.295	1.70E-13
LDL	0.044	9.907	1.04E-10
$\sqrt{\text{HDL}}$	0.168	43.657	9.40E-49
$\log_{10}$ TG	0.042	9.513	2.99E-10

TC, total cholesterol; LDL, low-density lipoprotein cholesterol; HDL, high-density lipoprotein cholesterol; TG, triglycerides; R^2, coefficient of determination of the model.

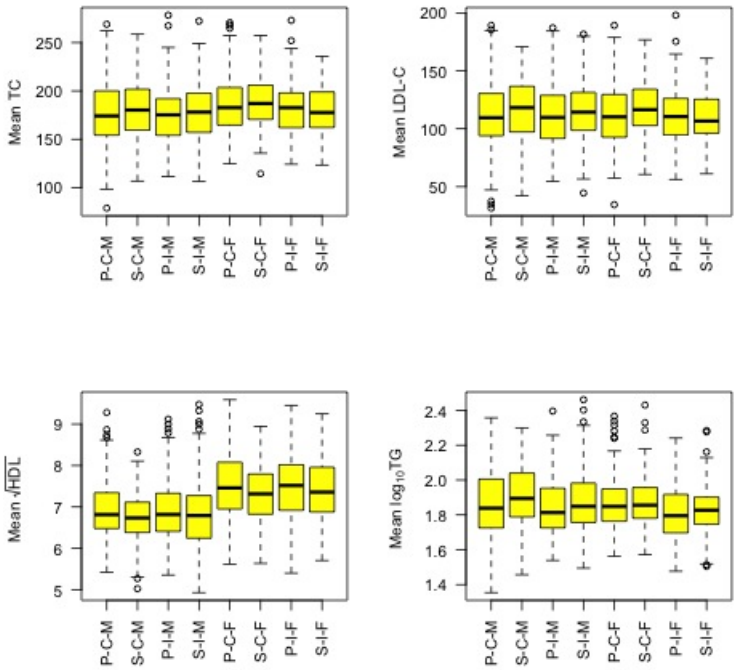

Figure 3.4: Mean lipid traits by cohort, treatment and gender. Mean represents the average of all possible measurements for a patient (excluding baseline measurements) for each lipid trait. Categories are cohort-treatment-gender: cohort, P is primary prevention and S is secondary intervention; treatment, C is conventional and I is intensive; gender, M is male and F is female.

Table 3.3: Count of SNPs by minor allele frequency (MAF) category, training set and model.

Training Set	Model	MAF (%)						Total
		< 1	$1-5$	$5-25$	$25-50$	$50-75$	$75-100$	
DGN whole blood	EN	2	2235	103473	69834	46206	27918	249668
	LASSO	1	1635	73768	47893	31038	19267	173602
GTEx whole blood	EN	13602	17878	49461	33049	21595	16449	152034
	LASSO	10029	13678	34734	22041	14238	11430	106150
GTEx pancreas	EN	15104	18628	50667	33738	21833	17000	156970
	LASSO	9763	13216	32685	20355	13072	10865	99956
GTEx liver	EN	9771	15936	43744	29101	18085	14653	131290
	LASSO	5656	11151	27360	16866	10507	8756	80296

EN, elastic net; LASSO, least absolute shrinkage and selection operator; MAF, minor allele frequency.

3.2.2 Imputed gene expression

The SNPs extracted from the DCCT data set using DGN whole blood contained fewer SNPs with a minor allele frequency (MAF) $\leq 5\%$. Therefore, fewer rare variants were involved in the lipid associations that used this training set (Table 3.3). The results of Chapter 2 showed that most of the variants in LASSO models were also present in elastic net models. To see if differences in the set of SNPs for an expression predictor resulted in different $\widehat{GReX}_g$, the $\widehat{GReX}_g$ derived using elastic net and LASSO for five randomly selected DCCT patients were compared. The correlation between elastic net and LASSO for all of the training sets was very strong suggesting that a similar $\widehat{GReX}_{g,T}$ was determined by the models. Similarly, the p values from associations across the genome were highly correlated (Figure 3.5). In contrast, scatter plots for $\widehat{GReX}_{g,T}^{EN}$ showed no correlation for all training set pairs except DGN whole blood and GTEx whole blood, which had very weak correlations (Figure 3.6). Weak correlations between p values from associations of training set pairs were also observed (Figure 3.7). Thus, the results of an association change with the selection of training set but not the selection of prediction model.

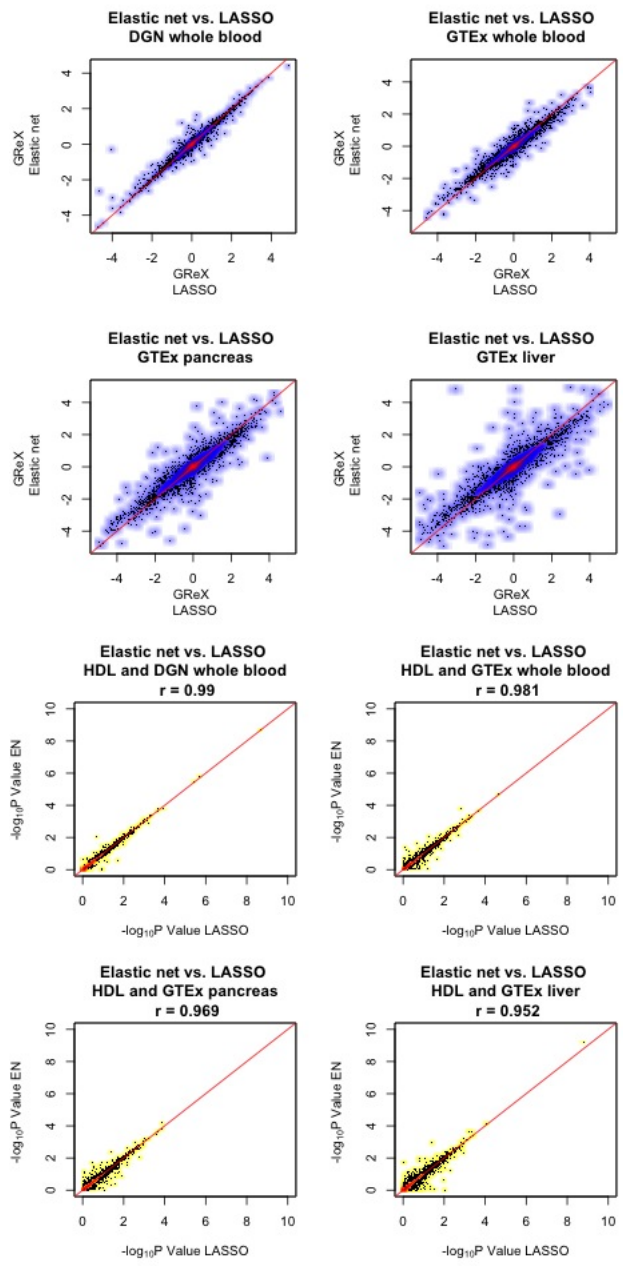

Figure 3.5: Comparison of the estimated genetically regulated expression (GReX) (Top) and $-\log p$ values for associations with $\sqrt{\text{HDL}}$ (Bottom) using expression predictors derived from elastic net and LASSO across the genome for matched genes for one randomly selected patient. EN, elastic net; LASSO, least absolute shrinkage and selection operator. Red line is $y = x$.

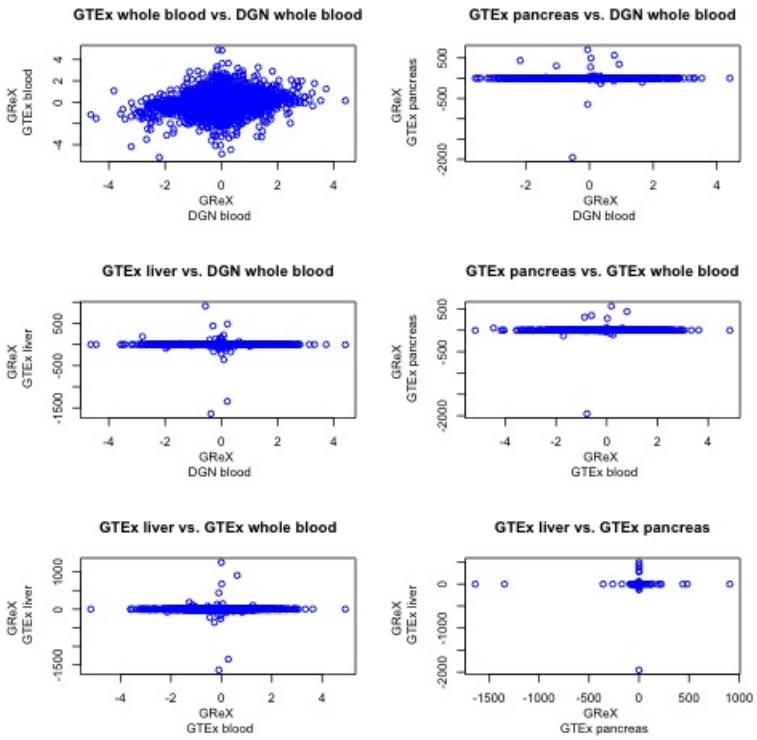

Figure 3.6: Comparison of the estimated genetically regulated expression (GReX) for matched genes for one randomly selected patient. Note the different scale for the plot of GTEx whole blood vs. DGN whole blood.

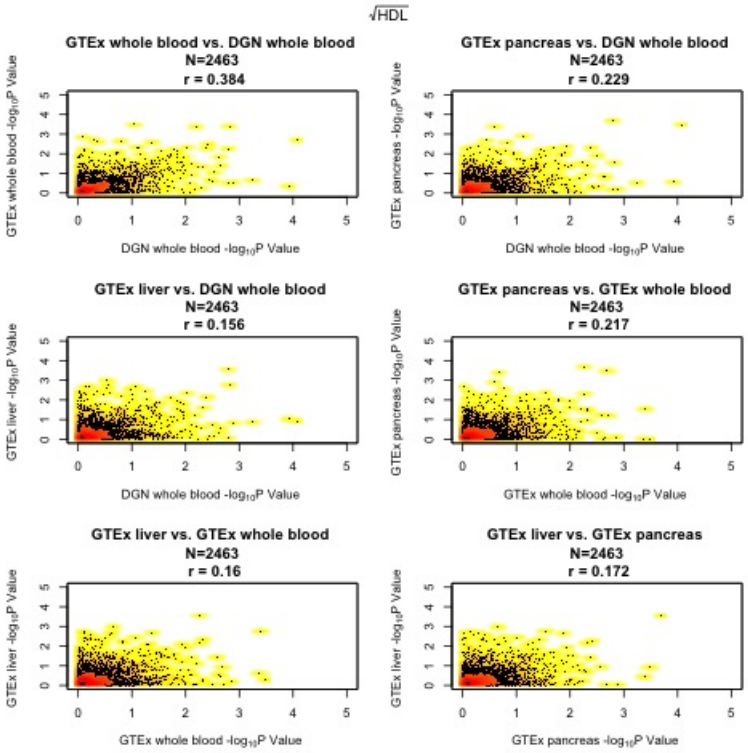

Figure 3.7: Comparison of the matched $-\log p$ values from 2463 associations with $\sqrt{\text{HDL}}$.

The number of t-tests varied by model and training set. With respect to elastic net, there were 11538, 10201, 9775 and 8557 for DGN whole blood, GTEx whole blood, GTEx pancreas and GTEx liver, respectively. The number of t-tests for LASSO for DGN whole blood, GTEx whole blood, GTEx pancreas and GTEx liver were 11520, 10053, 9591 and 8386, respectively. Some $\widehat{GReX}_g$ from GTEx whole blood, GTEx pancreas and GTEx liver had the same value for every patient because the genotype for each SNP in the set did not vary in the sample of patients (Table 3.4) and thus regressions were not possible.

Table 3.4: Summary statistics for the number of SNPs in the gene expression predictors with zero variance.

Training Set	Model	Genes	Number of SNPs			
			Mean	St. Dev.	Min	Max
GTEx whole blood	EN	14	1.571	0.852	1	3
	LASSO	14	1.357	0.633	1	3
GTEx pancreas	EN	18	1.722	1.708	1	8
	LASSO	22	1.273	0.456	1	2
GTEx liver	EN	4	1.500	1.000	1	3
	LASSO	16	1.438	0.814	1	3

EN, elastic net; LASSO, least absolute shrinkage and selection operator; Genes, gene expression predictor; MAF, minor allele frequency.

3.2.3 Associations

The null hypothesis was rejected more using model 3.3 with covariates, which marginally decreased the p values of suggestive or significant associations from model 3.2. The Spearman's rank correlation coefficient (r) between the $-\log p$ values of models 3.2 and 3.3 were strong for all associations except those involving $\sqrt{\text{HDL}}$, which showed moderate correlations (Figure 3.8) owing to the fact that covariates explain variation $\sqrt{\text{HDL}}$ to a larger extent than the other lipid traits, as previously shown. Model 3.3 associations with $\sqrt{\text{HDL}}$ also had a few outlying p values.

An expression predictor for a gene on Chromosome 4, $TADA2B$, from GTEx whole blood was suggestive with TC after adjusting for covariates. Similarly, associations with TC for expression predictors for genes on Chromosome 1 ($SORT1$; GTEx liver derived using LASSO and $CELSR2$; GTEx pancreas) were suggestive after adjusting for covariates. With respect to associations with LDL, expression predictors for genes on Chromosome 1 were either significant ($CELSR2$; GTEx pancreas) or suggestive ($PSRC1$; DGN whole blood and $SORT1$, $PSRC1$, $CELSR2$; GTEx liver) with LDL after adjusting for covariates (Figure 3.9).

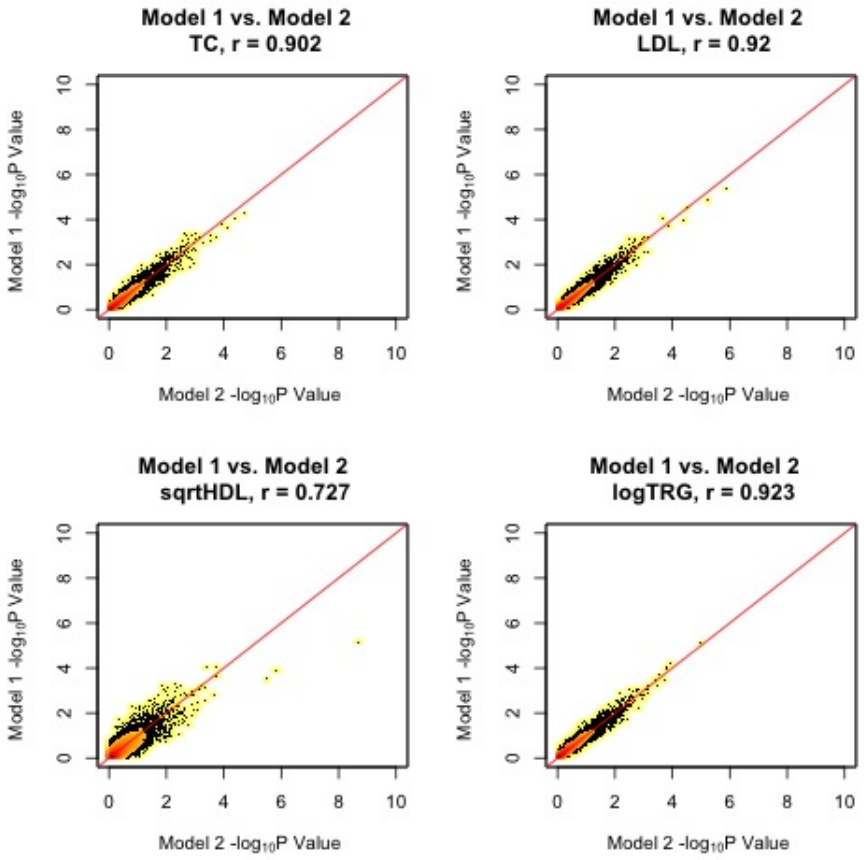

Figure 3.8: Comparison of $-\log p$ values across the genome for matched expression predictors between models without (Model 1) and with covariates (Model 2). Associations for all lipid traits with the expression predictors from DGN whole blood derived using elastic net are shown. Red line is $y = x$ and Spearman's rank correlation coefficient (r) is presented above each plot.

Association with $\sqrt{\text{HDL}}$ and an expression predictor for a gene, $NLRC5$, on Chromosome 16 from GTEx liver exceeded the significance threshold using model 3.2 (elastic net) and model 3.3 (elastic net and LASSO), however, expression predictors for this gene from the other training sets were not associated with $\sqrt{\text{HDL}}$, possibly owing to training set-specific differences in the gene expression predictor. An expression predictor for another gene on Chromosome 16, $SLC12A3$ was also significantly associated with $\sqrt{\text{HDL}}$ after adjustment for covariates, and two other expression predictors for genes on Chromosome 16, $CETP$ and $HERPUD1$, were suggestively associated with $\sqrt{\text{HDL}}$ using model 3.3 only (Figure 3.10). Adjusting for covariates did not enable suggestive or significant associations with $\log_{10}\text{TG}$, but all training sets showed a possible relationship with one or more expression predictors on Chromosome 11 with $\log_{10}\text{TG}$.

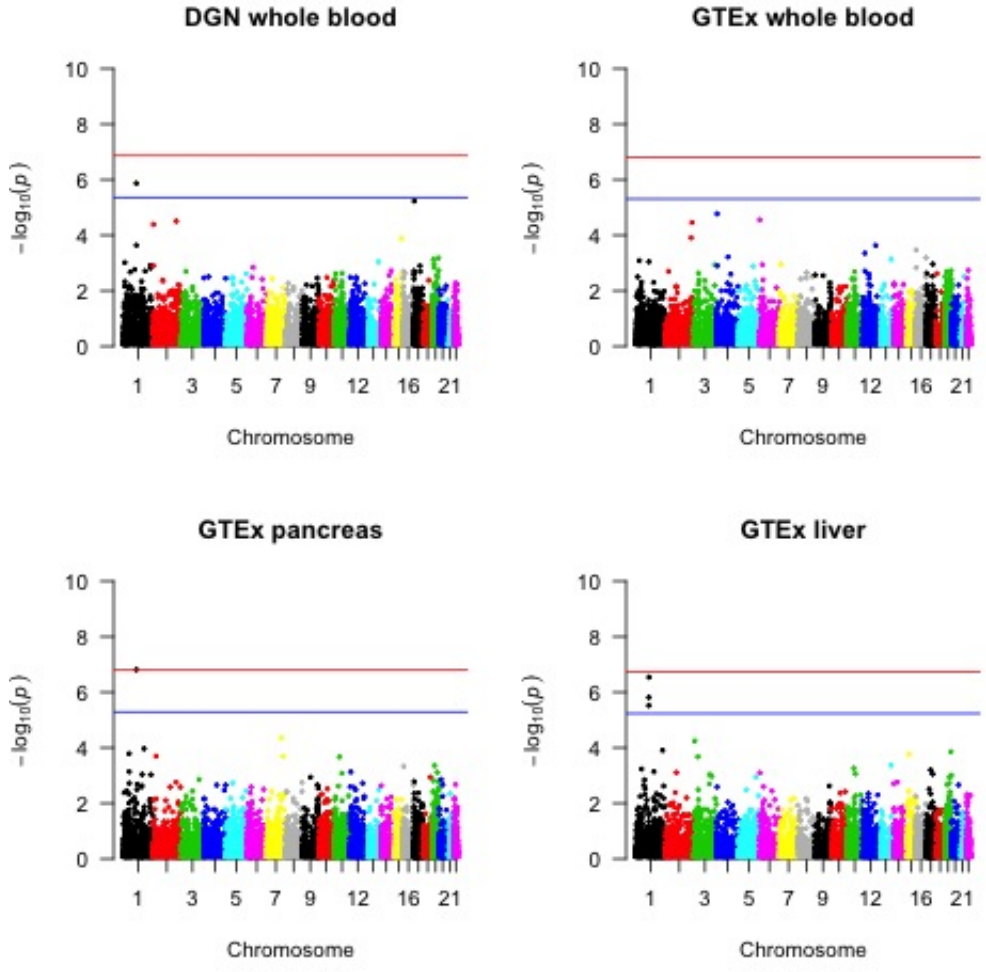

Figure 3.9: Manhattan plots for the mean LDL with the estimated GReX adjusted for covariates. The $-\log p$ value from associations using gene expression predictors from each of the training sets derived using elastic net are presented. The red line is significant and the blue line is suggestive.

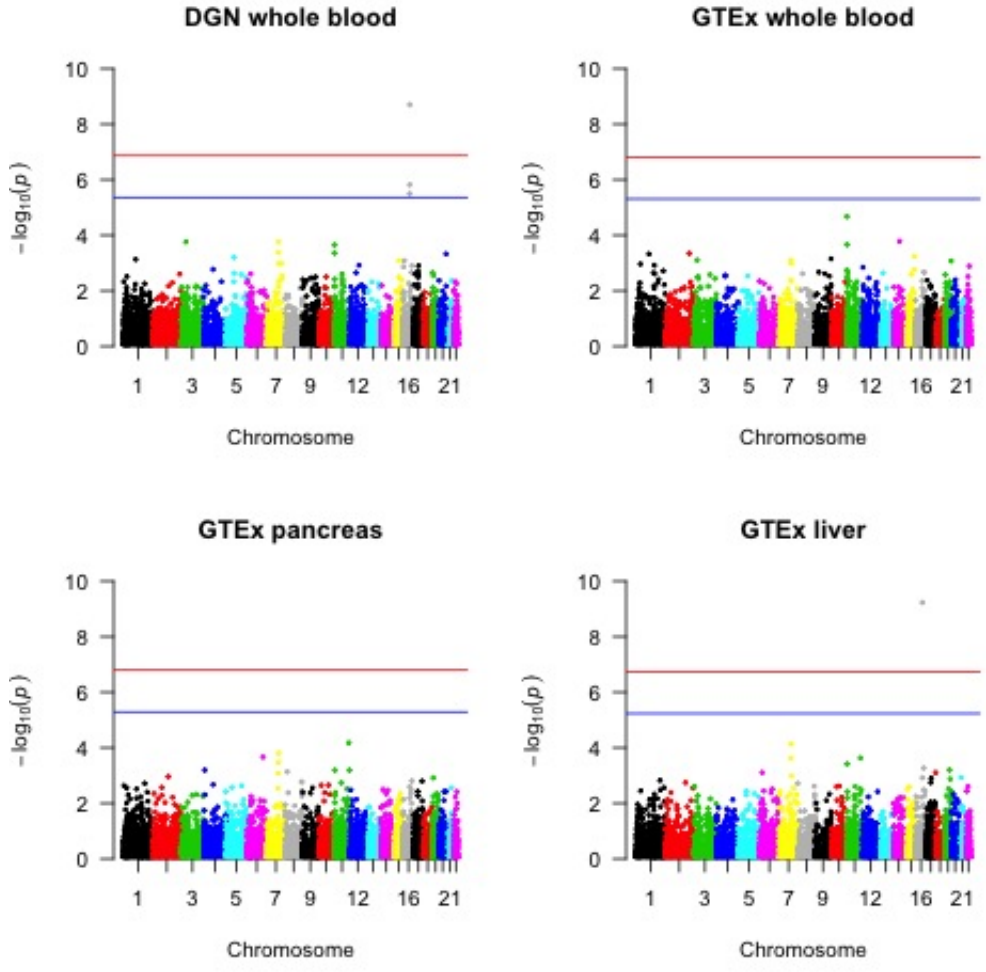

Figure 3.10: Manhattan plots for the mean $\sqrt{\text{HDL}}$ with the estimated GReX adjusted for covariates. The $-\log p$ value from associations using gene expression predictors from each of the training sets derived using elastic net are presented. The red line is significant and the blue line is suggestive.

The p values followed a uniform distribution and a few deviations from the null hypothesis were only apparent for the smallest p values (Figures 3.11-3.12). Values for λ_{gc} were slightly inflated but small enough to be consistent with nominal type 1 error for all associations from models 3.2 and 3.3. The fact that the sample of patients are related to one another by way of the disorder IDDM may explain the small inflation of p values (Table 3.5).

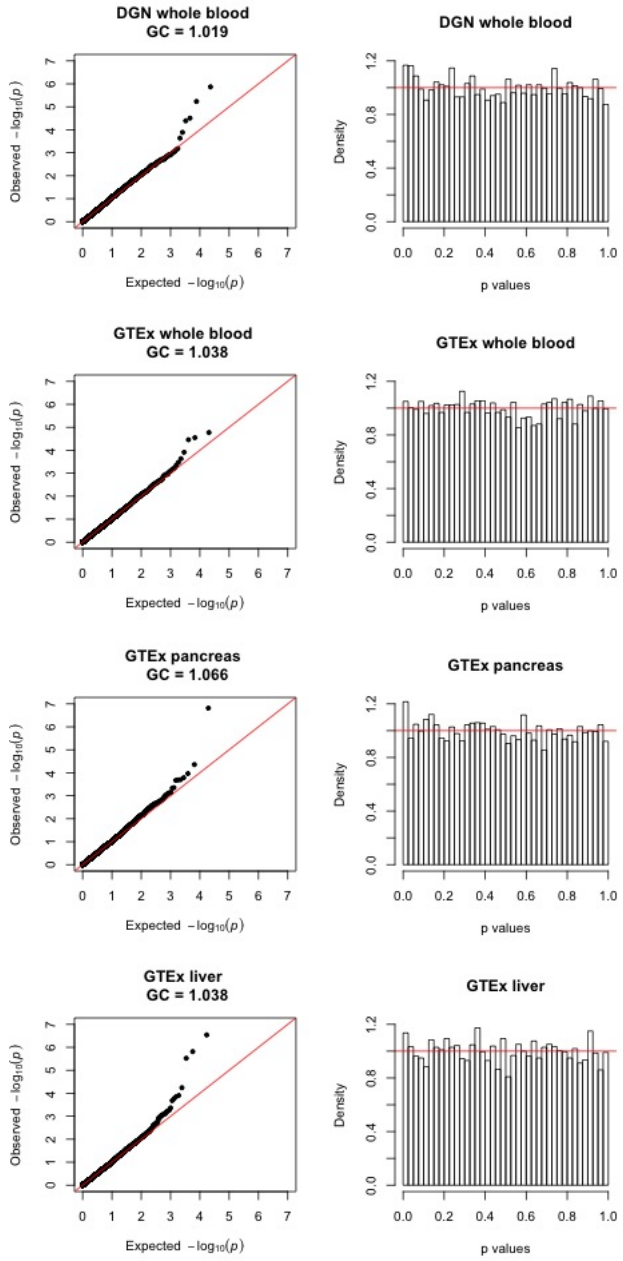

Figure 3.11: Quantile-quantile (Q-Q) and histogram p value plots for mean low-density lipoprotein cholesterol (LDL) with the estimated GReX adjusted for covariates. Left: Q-Q plots of the $-\log p$ values for associations across the genome versus the null (uniform) expectation using expression predictors derived using elastic net. GC, inflation factor for the distribution of p values. Right: histogram density plot of p values for associations across the genome. Red line marks density 1.

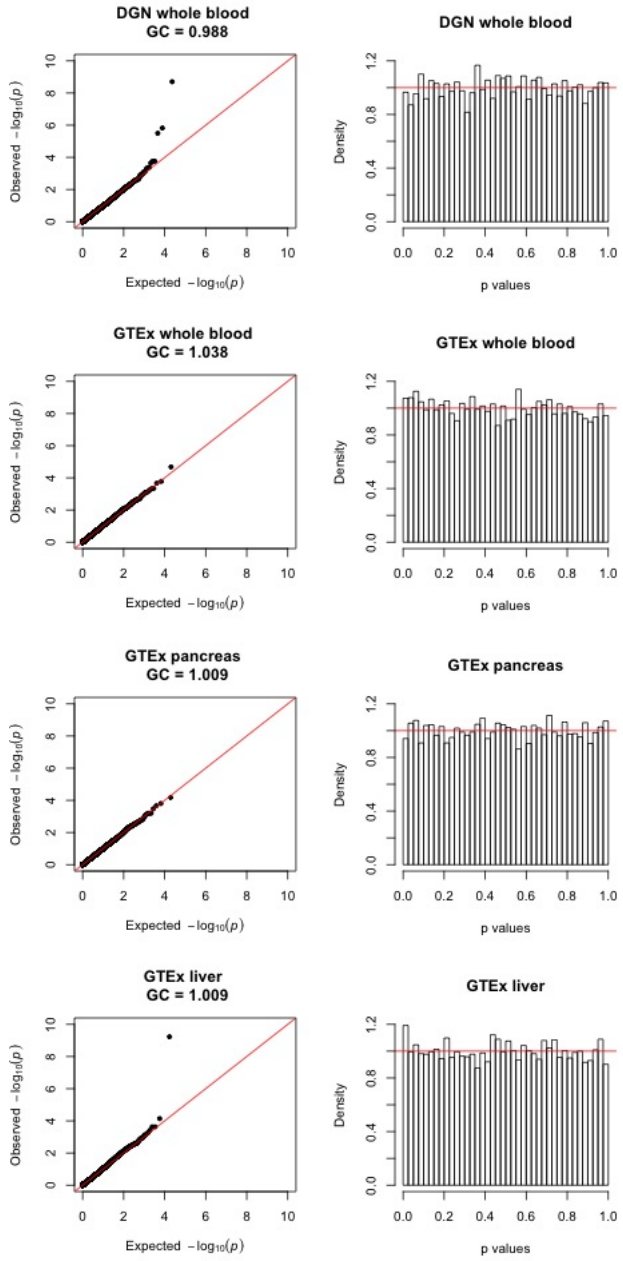

Figure 3.12: Quantile-quantile (Q-Q) and histogram p value plots for mean high-density lipoprotein cholesterol ($\sqrt{\text{HDL}}$) with the estimated GReX adjusted for covariates. Left: Q-Q plots of the $-\log p$ values for associations across the genome versus the null (uniform) expectation using expression predictors derived using elastic net. GC, inflation factor for the distribution of p values. Right: histogram density plot of p values for associations across the genome. Red line marks density 1.

Table 3.5: Inflation factor for the distribution of p values.

Lipid	Training Set	$M1_{EN}$	$M1_{LASSO}$	$M2_{EN}$	$M2_{LASSO}$
TC	DGN	1.019	1.016	1.012	1.006
	Blood	1.045	1.058	1.023	1.017
	Pancreas	1.032	1.051	1.046	1.034
	Liver	1.004	1.007	0.996	0.982
LDL	DGN	1.031	1.030	1.019	1.023
	Blood	1.026	1.023	1.038	1.031
	Pancreas	1.041	1.054	1.066	1.048
	Liver	1.004	1.021	1.038	1.018
$\sqrt{\text{HDL}}$	DGN	0.955	0.968	0.988	0.997
	Blood	1.028	1.032	1.038	1.041
	Pancreas	1.038	1.028	1.009	1.007
	Liver	1.023	1.018	1.009	1.015
logTG	DGN	1.035	1.039	1.030	1.025
	Blood	1.052	1.050	1.036	1.040
	Pancreas	1.021	1.034	1.017	1.018
	Liver	1.053	1.060	1.062	1.044

$M1_{EN}$, model without covariates from EN (elastic net); $M1_{LASSO}$, model without covariates from LASSO (least absolute shrinkage and selection operator); $M2_{EN}$, model with covariates from EN; $M2_{LASSO}$, model with covariates from LASSO.; TC; total cholesterol; LDL, low-density lipoprotein cholesterol; HDL, high-density lipoprotein cholesterol; TG, triglycerides.

The smallest p values for the associations using models 3.2 and 3.3 are summarized in Tables A.2 and 3.6, respectively.

Table 3.6: Top gene expression predictors with covariates.

Model	Training Set	Lipid	*Gene*	Chr.	$R^2_{prediction}$	SNPs	β_1	P	R^2
Elastic net	GTEx liver	$\sqrt{\text{HDL}}$	*NLRC5*	16	0.00	11	-0.76	5.89E-10	0.192
	DGN whole blood	$\sqrt{\text{HDL}}$	*SLC12A3*	16	0.01	42	-0.76	1.98E-09	0.191
	GTEx pancreas	LDL	*CELSR2**	1	0.06	2	-28.67	1.53E-07	0.064
	GTEx liver	LDL	*SORT1**	1	0.44	13	-7.51	2.88E-07	0.063
	DGN whole blood	LDL	*PSRC1**	1	0.25	11	-7.89	1.35E-06	0.061
	DGN whole blood	$\sqrt{\text{HDL}}$	*CETP**	16	0.02	30	-0.62	1.50E-06	0.183
	GTEx liver	LDL	*CELSR2**	1	0.37	23	-6.67	1.53E-06	0.061
	GTEx liver	LDL	*PSRC1**	1	0.39	40	-5.78	2.99E-06	0.060
	DGN whole blood	$\sqrt{\text{HDL}}$	*HERPUD1**	16	0.09	15	-0.43	3.17E-06	0.182
	GTEx pancreas	TC	*CELSR2**	1	0.06	2	-28.19	3.52E-06	0.069
	GTEx whole blood	TC	*TADA2B*	4	0.02	31	-23.09	3.94E-06	0.069
LASSO	GTEx liver	$\sqrt{\text{HDL}}$	*NLRC5*	16	0.00	9	-0.73	1.66E-09	0.191
	DGN whole blood	$\sqrt{\text{HDL}}$	*SLC12A3*	16	0.01	32	-0.74	2.06E-09	0.191
	GTEx pancreas	LDL	*CELSR2**	1	0.06	2	-26.13	1.13E-07	0.064
	GTEx liver	LDL	*SORT1**	1	0.48	5	-7.35	2.11E-07	0.063
	GTEx liver	LDL	*CELSR2**	1	0.40	10	-7.77	2.13E-07	0.063
	GTEx liver	LDL	*PSRC1**	1	0.38	25	-5.66	1.68E-06	0.061
	DGN whole blood	LDL	*PSRC1**	1	0.25	6	-7.69	2.01E-06	0.060
	DGN whole blood	$\sqrt{\text{HDL}}$	*CETP**	16	0.02	28	-0.60	2.02E-06	0.182
	GTEx pancreas	TC	*CELSR2**	1	0.06	2	-25.67	2.81E-06	0.070
	DGN whole blood	$\sqrt{\text{HDL}}$	*HERPUD1**	16	0.08	11	-0.42	3.99E-06	0.182
	GTEx whole blood	TC	*TADA2B*	4	0.02	28	-21.39	4.36E-06	0.069
	GTEx liver	TC	*SORT1**	1	0.48	5	-7.23	4.52E-06	0.069

Imputed gene expression-lipid associations with the smallest p values. PrediXcan models included covariates. Gene expression predictors above the dotted line were significant. Genes marked by an asterisk (*) are known quantitative trait loci for the lipid trait (NHGRI-EBI GWAS Catalog; MacArthur et al. (2017)). Gene, gene expression predictor; EN, elastic net; LASSO, least absolute shrinkage and selection operator; Chr., chromosome; $R^2_{prediction}$, 10-fold cross-validated R^2 for predictive performance; SNPs, the number of variants in the expression predictor; R^2, coefficient of determination for association; TC, total cholesterol; LDL, low-density lipoprotein cholesterol; HDL, high-density lipoprotein cholesterol.

Collectively, LDL was negatively associated with expression predictors for genes on Chromosome 1, TC was negatively associated with an expression predictor for a gene on Chromosome 4, and $\sqrt{\text{HDL}}$ was negatively associated with expression predictors for genes on Chromosome 16. Of the suggestive and significantly associated expression predictors with lipid traits, pairwise plots for Chromosome 1p13.3 haplotype members $CELSR2-PSRC1-SORT1$ from GTEx liver derived using elastic net showed strong correlation and similar plots for $CETP$, $SLC12A3$, and $HERPUD1$ from DGN whole blood derived using elastic net showed weak to moderate correlation (Figure 3.13).

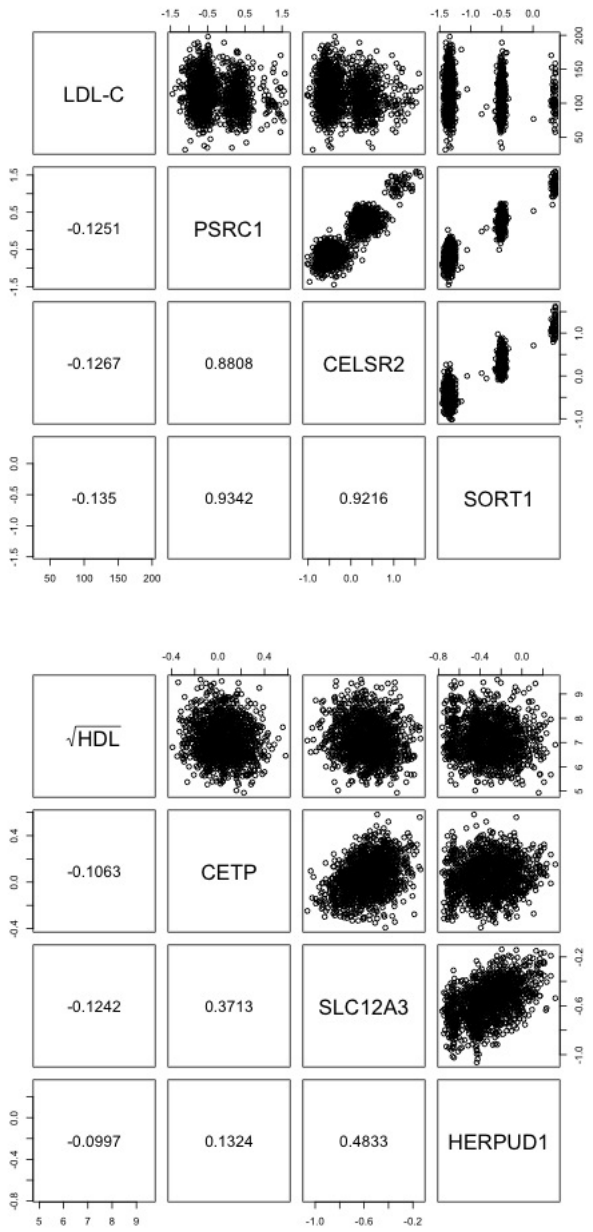

Figure 3.13: Pairwise plots of the estimated GReX for LDL (Top) and $\sqrt{\text{HDL}}$ (Bottom) associated genes. Correlations among the lipid traits and estimated GReX for genes are shown.

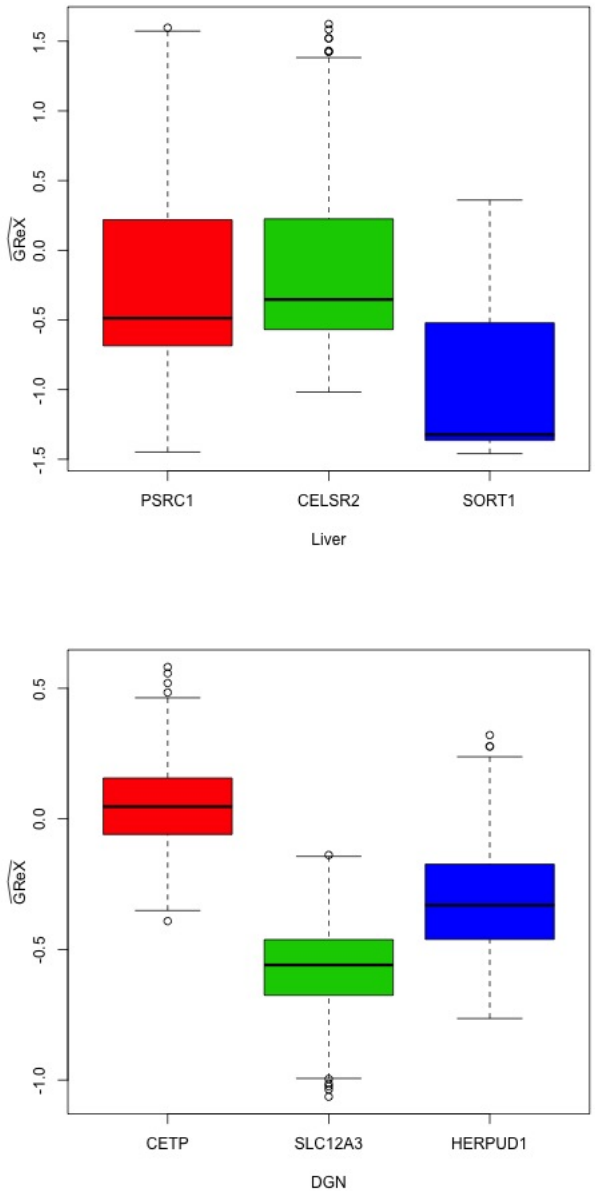

Figure 3.14: Distribution of the estimated GReX for LDL (Top) and $\sqrt{\text{HDL}}$ (Bottom) suggestive and significant genes.

The distribution of estimated GReXs for $PSRC1$ and $CELSR2$ were similar, consistent with their p values from associations with LDL; GReXs for SORT1 were more negatively distributed and association with LDL had a smaller p value than

the other haplotype block members. The distribution of predicted expressions for $CETP$, $SLC12A3$, and $HERPUD1$ were less similar (Figure 3.14) along with their p values from associations. Four SNPs located in the noncoding region of $CELSR2$ and $PSRC1$ are commonly cited in the literature: rs599839, rs629301, rs646776 and rs12740374 (Postmus et al., 2014; Schierding and O'Sullivan, 2015). In liver, rs646776 was highly associated with the expression of $CELSR2$, $PSRC1$ and $SORT1$, with $SORT1$ being most pronounce (Musunuru et al., 2010). While GTEx liver derived using LASSO contained rs12740374 for only $SORT1$, GTEx liver derived using elastic net contained rs646776 and rs12740374 for all three expression predictors, rs599839 for $PSRC1$ and rs629301 for $PSRC1$ and $SORT1$ (Table 3.7). The $CELSR2$ expression predictor from GTEx pancreas derived using elastic net had only two variants and these SNPs were also present in the other expression predictors that resulted in suggestive associations with LDL (Table 3.8). One of these two SNPs, rs12740374, was shown to be associated with LDL (MacArthur et al., 2017). Interestingly, the weights of these two SNPs had the largest absolute magnitude relative to the other SNPs in the expression predictor for GTEx training sets. With the exception of DGN whole blood, the weight for one variant was the negative of the other variant (Table 3.8 and A.9-A.12).

Table 3.8: Weights for two variants that were common to the gene expression predictors used for the LDL association models.

SNP	Position	GTEx liver			DGN blood	GTEx pancreas
		$PSRC1$	$CELSR2$	$SORT1$	$PSRC1$	$CELSR2$
rs12740374	109817590	-0.3366	-0.3870	-0.4015	0.0566	-0.1137
rs7528419	109817192	0.3442	0.3971	0.3996	0.0553	0.1128

SNP, the rsid number; position, the genomic position of the SNP in base pairs

Table 3.7: Previously reported quantitative trait loci for LDL from the NHGRI-EBI GWAS Catalog.

SNP	A1	A2	GTEx liver			DGN blood	GTEx pancreas
			PSRC1	CELSR2	SORT1	PSRC1	CELSR2
rs12740374	G	T	*	*	*	*	*
rs599839	G	A	*			*	
rs629301	G	T	*		*	*	
rs646776	C	T	*	*	*	*	
rs660240	T	C				*	

The asterisk (*) indicates the presence of the SNP in the gene expression predictor. SNP; represents known (NHGRI-EBI GWAS Catalog; www.ebi.ac.uk/gwas/) significantly associated SNPs with LDL; A1; effect allele from DCCT; A2, reference allele from DCCT.

Regression diagnostics for the suggestive and significant associations from models 3.2 and 3.3 showed that the standard regression assumptions were satisfied and the models adequately fit the data. The models were linear in the regression parameters and the standardized errors were independently and identically distributed normal random variables with mean zero and variance one. $\sqrt{\text{HDL}}$ association with the expression predictor for $NLRC5$ from GTEx liver derived using elastic net showed few residuals beyond 3 standard deviations but many high leverage points that were not influential at a threshold of $C_i = 0.69$ in model 3.2 and threshold $C_i = 0.92$ for model 3.3. The diagnostic plots for association of the expression predictor for $NLRC5$ with $\sqrt{\text{HDL}}$ from model 3.3 are shown in Figure 3.15.

The GTEx liver expression predictor for $NLRC5$ was most significantly associated with $\sqrt{\text{HDL}}$ in both models, but it had a negligible 10-fold cross-validated R^2 for explaining the variation in the expression of $NLRC5$. On the other hand, the 10-fold cross-validated R^2 for the suggestive gene expression predictors on Chromosome 1 were very strong for the models built with DGN whole blood and GTEx liver.

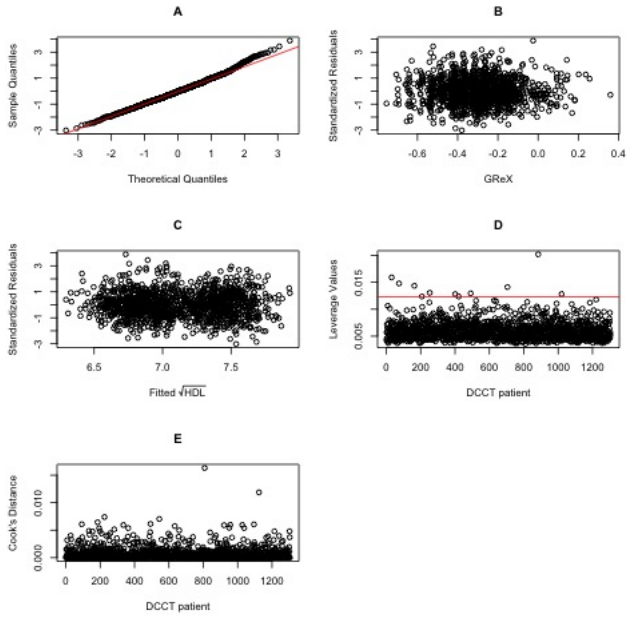

Figure 3.15: Diagnostic plots for the PrediXcan model with covariates relating $\sqrt{\text{HDL}}$ with the $NLRC5$ expression predictor from GTEx liver derived using elastic net. (A) Normal probability plot of the standardized residuals; (B) Scatter plot of the standardized residuals against the estimated GReX for $NLRC5$; (C) Scatter plot of the standardized residuals against the fitted values; (D) Leverage values; (E) Cook's distance. No points exceeded the threshold of $C_i = 0.92$ which is beyond the axis of the plot.

As noted in Chapter 2, the PredictDB training sets did not contain the same expression predictors. Expression predictors for two genes on Chromosome 1 ($CELSR2$ and $SORT1$) were absent from GTEx whole blood explaining in part the failure to observe significant associations on Chromosome 1 with this training set. $TADA2B$ (Chromosome 4) and $SLC12A3$ (Chromosome 16) expression predictors were not present in GTEx pancreas and GTEx liver, respectively, and an expression predictor for a gene on Chromosome 16 ($HERPUD1$) was only present in DGN whole blood.

DGN whole blood and GTEx liver contained expression predictors for $TADA2B$ but associations with TC and $TADA2B$ were not observed for these training sets. This finding substantiates the observations of Figure 3.6 that demonstrated that the

estimated GReX for a gene differs substantially across training sets. Such differences suggest that care should be taken when selecting training sets for the imputation of gene expression.

Chapter 4

Lipid associations using multiple linear regression

4.1 Methods

Lipid Associations. The mean across time for each of the four lipid traits: TC, LDL, $\sqrt{\text{HDL}}$ or $\log_{10}$ TG was associated with the expression predictors from the training sets using the multiple linear regression (MLR) model

$$Y_{lipid} = \beta^M_{0,g,T} + \beta^M_{1,g,T} X^M_{1,g,T} + \beta^M_{2,g,T} X^M_{2,g,T} + \cdots + \beta^M_{p,g,T} X^M_{p,g,T} + \epsilon^M_{g,T} \tag{4.1}$$

with model M: elastic net or LASSO, training set T: DGN whole blood, GTEx whole blood, GTEx pancreas, or GTEx liver, and $X_{i,g,T}$: the dosage of SNP i in a set for gene g for $i = 1, 2, ..., p$. The F-test with p and $n - p - 1$ degrees of freedom was used to test the null hypothesis that all of the SNP β coefficients were zero

$$H_0 : \quad \beta^M_{1,g,T} = \beta^M_{2,g,T} = \cdots = \beta^M_{p,g,T} = 0$$

versus the alternative that at least one SNP β coefficient was not zero; p denotes those SNPs from the set for a gene that varied among the 1304 patients.

The multiple linear regression model with expression predictors and covariates age (C_1), duration of IDDM (C_2), gender (C_3), cohort (C_4), treatment (C_5) and the interaction between cohort and treatment (C_6):

$$Y_{lipid} = \beta^M_{0,g,T} + \beta^M_{1,g,T} X^M_{1,g,T} + \beta^M_{2,g,T} X^M_{2,g,T} + \cdots + \beta^M_{p,g,T} X^M_{p,g,T} + \gamma_1 C_1 + \cdots + \gamma_6 C_6 + \epsilon^M_{g,T} \tag{4.2}$$

was tested against the model with only covariates:

$$Y_{lipid} = \gamma_0 + \gamma_1 C_1 + \cdots + \gamma_6 C_6 + \epsilon$$

using the F test with p and $n - p - 7$ degrees of freedom to see if the models with SNPs and covariates explained variation in the lipid trait better than the models with only covariates. Each model assumed that the ϵ's were independent and normally distributed with $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2$. The lm and anova functions from R version 3.3.0 (2016-05-03) were used to fit the multiple linear regression models. The p values for the associations and deviations from the null hypothesis were visualized graphically according to methods described in Chapter 3. Similarly, the linearity and normality assumptions of the models were evaluated as described in Chapter 3 with modifications.

4.2 Results

4.2.1 Associations

Multiple linear regression (MLR) and the F-test were used to model the dependence of lipids on the SNPs from the gene expression predictors. The number of F-tests were identical to the number of Student's t-tests in Chapter 3 except for the expression predictors from GTEx liver derived using LASSO, which had three fewer $\widehat{GReX}_g$ associations in Chapter 3 because the vector for three $\widehat{GReX}_g$ had the same dosage for every individual. The $-\log p$ value correlations for MLR associations for matched genes between elastic net and LASSO (Figure 4.1) and pairs of training sets (Figure 4.2) were weaker than those of Chapter 3. Hence, differences in the SNP sets either between elastic net and LASSO or the training sets had a larger impact on the p values using MLR. However, a comparison of the $-\log p$ values attained using model 4.1 (without covariates) and model 4.2 (with covariates) showed stronger correlations for associations than those observed in Chapter 3 indicating that the covariates in MLR models had a smaller effect on the p values (Figure 4.3). The correlations between models 4.1 and 4.2 for $\sqrt{\text{HDL}}$ were weaker than for the other lipids, as observed in Chapter 3.

Multiple linear regressions of the SNPs from expression predictors with TC did not reveal any suggestive or significant associations. Suggestive associations with LDL were observed on Chromosome 1 ($CELSR2$; GTEx pancreas and $SORT1$; GTEx liver) and Chromosome 19 ($ZNF222$; GTEx whole blood) irrespective of whether covariates were included (Figure 4.4).

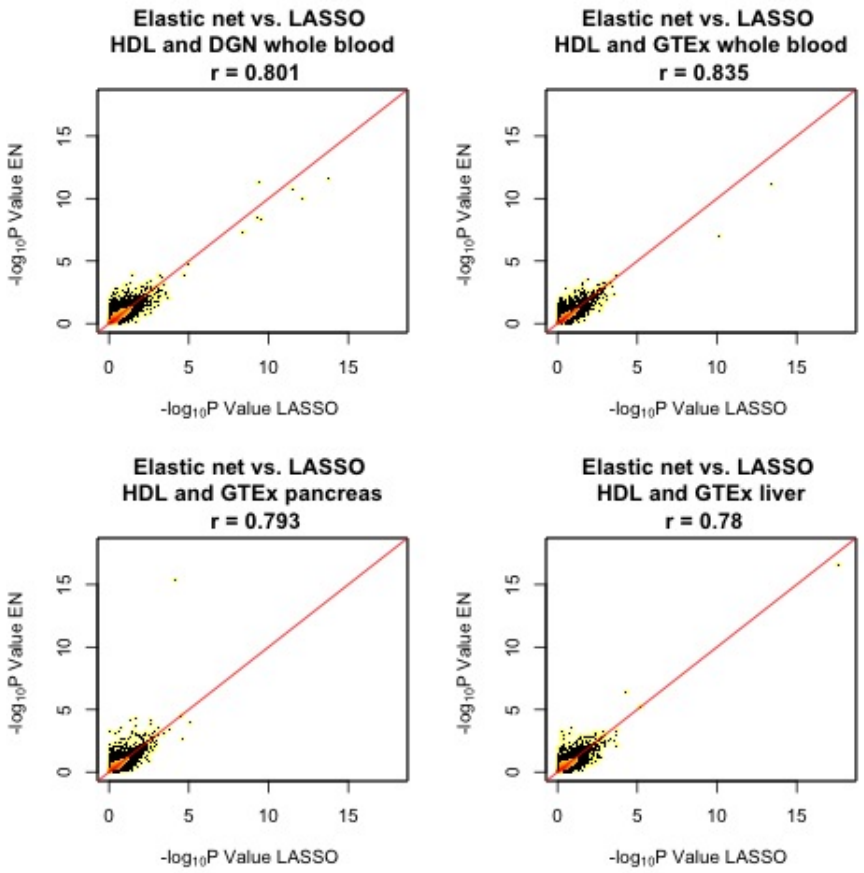

Figure 4.1: Elastic net versus LASSO model comparisons of the $-\log p$ values for associations with $\sqrt{\text{HDL}}$ across the genome for matched expression predictors. Comparisons for all training sets are shown. Red line is $y = x$ and Spearman's rank correlation coefficient (r).

Model 4.2 with covariates derived from DGN whole blood using LASSO showed a suggestive association with LDL on Chromosome 1 ($PSRC1$) and models 4.1 and 4.2 (without and with covariates) derived from DGN whole blood using LASSO showed suggestive associations with LDL on Chromosome 19 ($ZNF233$ and $KLC3$).

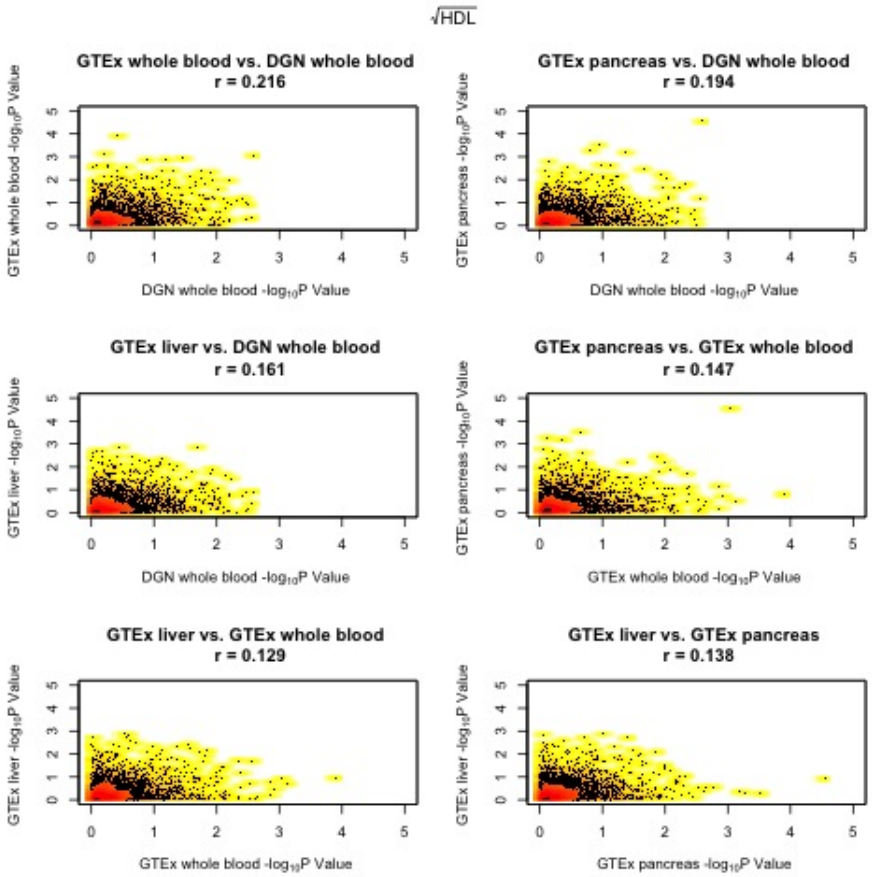

Figure 4.2: Training set comparisons of the $-\log p$ values from 2463 matched associations with $\sqrt{\text{HDL}}$. Training sets derived using elastic net are presented. Spearman's rank correlation coefficient (r) is shown.

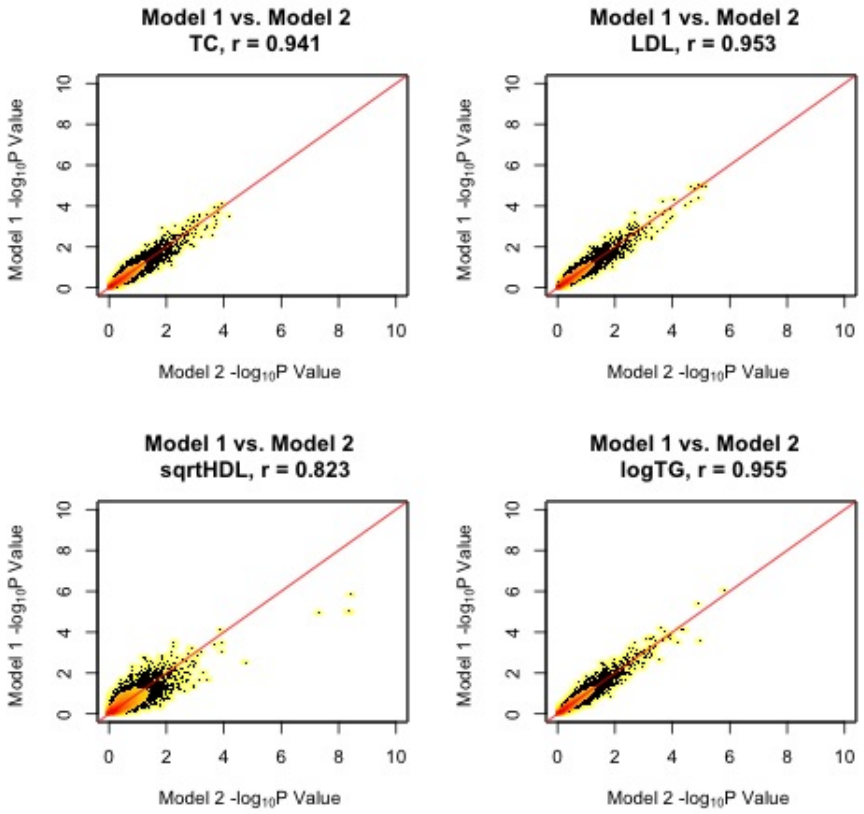

Figure 4.3: Comparison of $-\log p$ values across the genome for matched expression predictors between models without (Model 1) and with covariates (Model 2). Associations for all lipid traits with the expression predictors from DGN whole blood derived using elastic net are shown. Red line is $y = x$ and Spearman's rank correlation coefficient (r) is presented.

Multiple linear regressions for models 4.1 and 4.2 with $\sqrt{\text{HDL}}$ revealed significant associations on Chromosome 16 for all training sets from elastic net (Figure 4.5) and all training sets (except GTEx pancreas) from LASSO. A suggestive association with $\log_{10}$ TG on Chromosomes 19 ($SUGP2$) in models 4.1 and 4.2 occurred using DGN whole blood. One gene expression predictor on Chromosome 16 ($ZNRF1$) from DGN whole blood was also associated with $\log_{10}$ TG using model 4.1. The smallest p values for models 4.1 and 4.2 are summarized in Tables A.3 and 4.3, respectively.

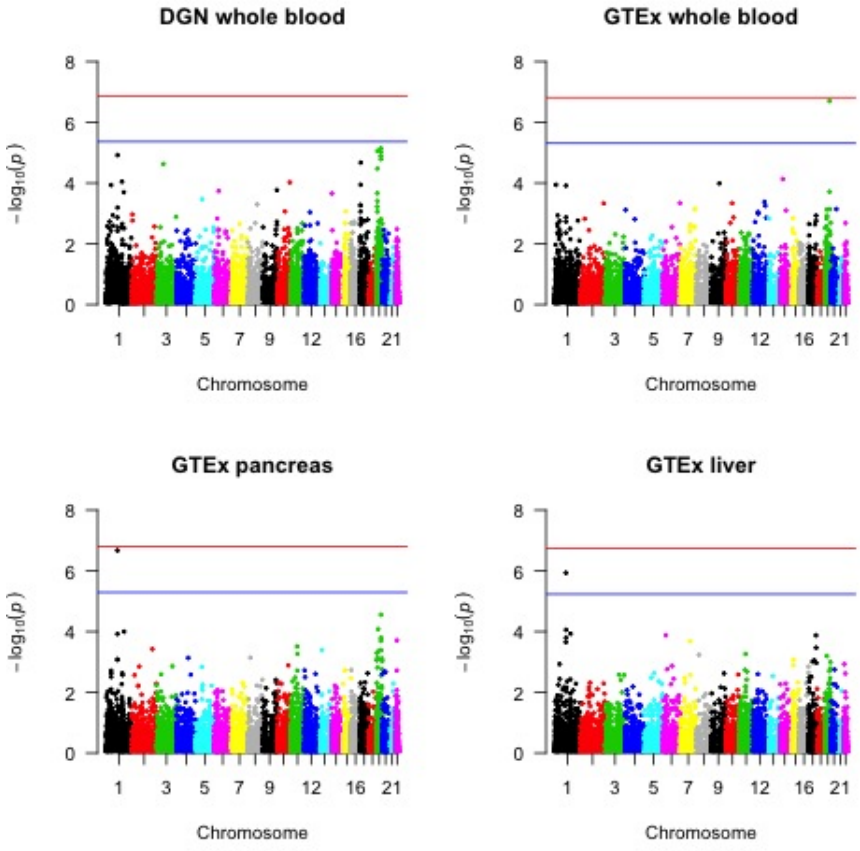

Figure 4.4: Manhattan plots for the mean LDL using MLR with covariates. The $-\log\ p$ value from associations using gene expression predictors from each of the training sets derived using elastic net are presented. The red line is significant and the blue line is suggestive.

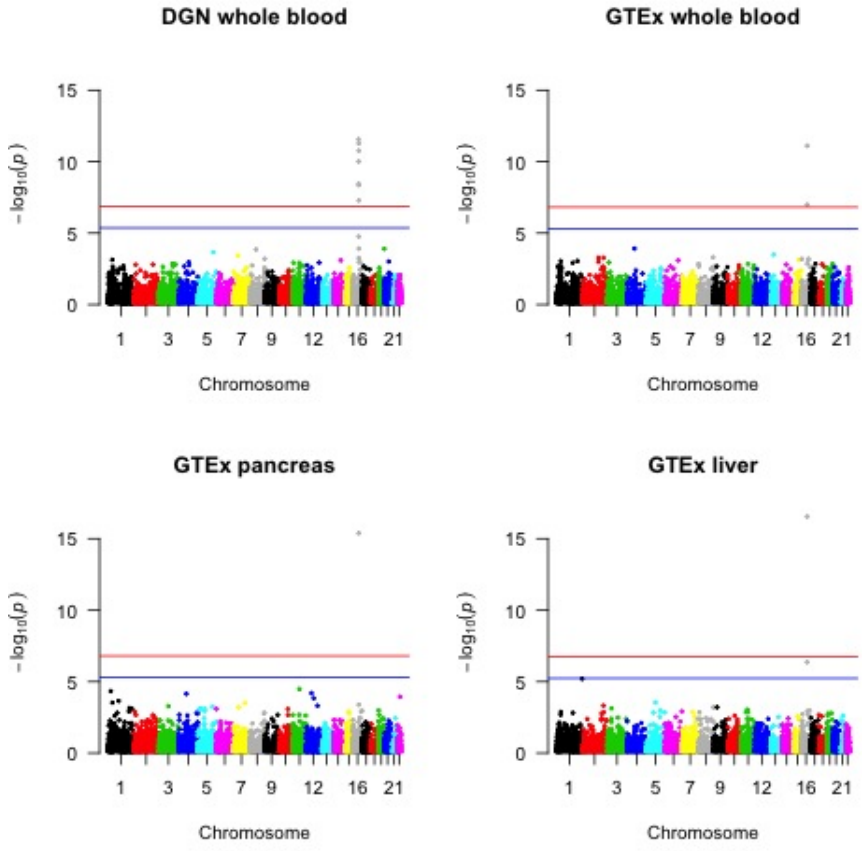

Figure 4.5: Manhattan plots for $\sqrt{\text{HDL}}$ using MLR with covariates. The $-\log p$ value from associations using gene expression predictors from each of the training sets derived from elastic net are presented. The red line is significant and the blue line is suggestive.

Q-Q plots showed that the $-\log p$ values followed the $y = x$ line for most of the range except for the smallest p values. Histograms of p values for each training set revealed more than were expected of the very smallest p values for a few of the training sets (Figures 4.6-4.7). Collectively, the λ_{gc} for MLR models tended to be larger than

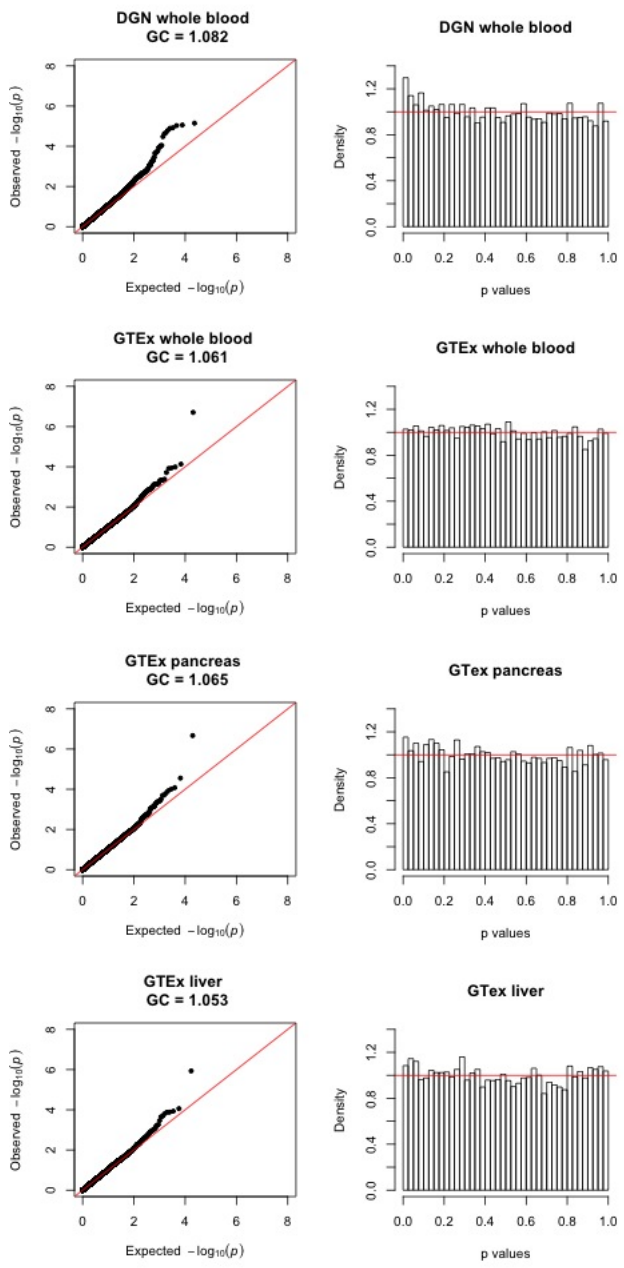

Figure 4.6: Quantile-quantile (Q-Q) and histogram p values for associations with mean LDL. MLR models with covariates and gene expression predictors derived using elastic net are presented. Left: $-\log p$ values from associations across the genome versus the null expectation. Theoretical quantiles are from a uniform distribution. GC, the genomic control inflation factor for p values is shown. Right: histograms of p values for associations across the genome. Red line marks density 1.

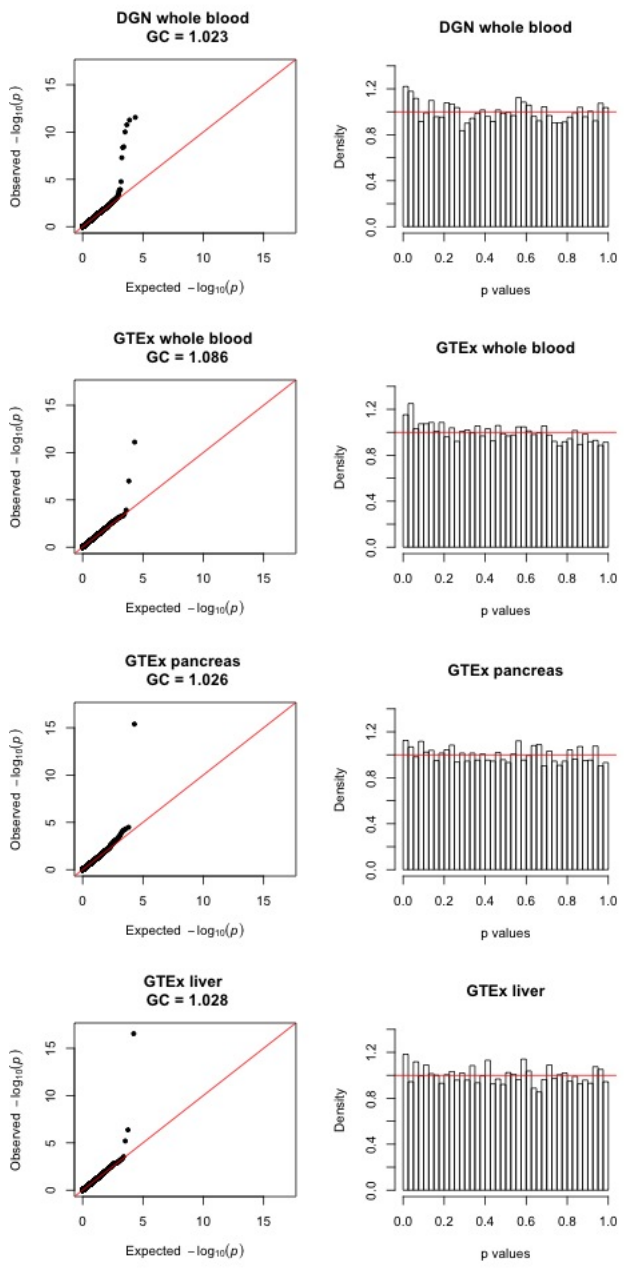

Figure 4.7: Quantile-quantile (Q-Q) and histogram p values for associations with mean $\sqrt{\text{HDL}}$. MLR models with covariates and gene expression predictors derived using elastic net are presented. Left: $-\log p$ values from associations across the genome versus the null expectation. Theoretical quantiles are from a uniform distribution. GC, the genomic control inflation factor for p values is shown. Right: histograms of p values for associations across the genome. Red line marks density 1.

Table 4.1: Genomic control inflation factor for p values for MLR models derived using elastic net (EN) and least absolute shrinkage operator (LASSO) without covariates (M1) and with covariates (M2).

Lipid	Training Set	M1 EN	M1 LASSO	M2 EN	M2 LASSO
TC	DGN whole blood	1.078	1.058	1.085	1.070
	GTEx whole blood	1.005	1.019	1.002	1.019
	GTEx pancreas	0.997	0.997	1.023	1.021
	GTEx liver	1.024	0.994	1.003	0.998
LDL	DGN whole blood	1.092	1.075	1.082	1.055
	GTEx whole blood	1.050	1.055	1.061	1.080
	GTEx pancreas	1.055	1.035	1.065	1.066
	GTEx liver	1.036	1.027	1.053	1.027
sqrtHDL	DGN whole blood	1.072	1.025	1.023	1.017
	GTEx whole blood	1.049	1.028	1.086	1.074
	GTEx pancreas	1.026	1.049	1.026	1.044
	GTEx liver	1.044	1.072	1.028	1.072
logTG	DGN whole blood	1.078	1.080	1.059	1.081
	GTEx whole blood	1.066	1.066	1.060	1.059
	GTEx pancreas	1.050	1.047	1.057	1.030
	GTEx liver	1.016	1.043	1.023	1.048

TC, Total cholesterol; LDL, low-density lipoprotein cholesterol; HDL, high-density lipoprotein cholesterol; TG, triglycerides.

those for PrediXcan models (Tables 4.1 and 4.2). Taken together, the significant and suggestive expression predictors were from Chromosomes 1, 16 and 19. Associations with expression predictors for genes on Chromosome 16 with $\sqrt{\text{HDL}}$ were most frequent and the genomic regions spanned by the SNPs in these expression predictors overlapped one another (Table 4.4).

Table 4.2: Comparison of genomic control inflation factor for p values for PrediXcan and MLR models.

Lipid	Training Set	PrediXcan	MLR
TC	DGN whole blood	1.012	1.085
	GTEx whole blood	1.023	1.002
	GTEx pancreas	1.046	1.023
	GTEx liver	0.996	1.003
LDL	DGN whole blood	1.019	1.082
	GTEx whole blood	1.038	1.061
	GTEx pancreas	1.066	1.065
	GTEx liver	1.038	1.053
sqrtHDL	DGN whole blood	0.988	1.023
	GTEx whole blood	1.038	1.086
	GTEx pancreas	1.009	1.026
	GTEx liver	1.009	1.028
logTG	DGN whole blood	1.030	1.059
	GTEx whole blood	1.036	1.060
	GTEx pancreas	1.017	1.057
	GTEx liver	1.062	1.023

TC, Total cholesterol; LDL, low-density lipoprotein cholesterol; HDL, high-density lipoprotein cholesterol; TG, triglycerides.

The expression predictors from DGN whole blood derived using LASSO were subsets of their matched expression predictors from DGN whole blood derived from elastic net for all significant Chromosome 16 genes with the exception of $CCDC135$ which had one distinct SNP (rs7199577) in DGN whole blood derived using LASSO. As discussed in Chapter 2, a single variant may be present in more than one expression predictor and this mostly occurred in DGN whole blood (Table 2.3). One variant (rs9989419) intersected six expression predictors—$CIAPIN1$, $CPNE2$, $GNAO1$, $HERPUD1$, $MT1X$ and $SLC12A3$—from DGN whole blood and this SNP was reported with the genes $CETP$ and $HERPUD1$ from associations with HDL on the NHGRI-EBI

GWAS Catalag (MacArthur et al., 2017).

Table 4.3: Top gene expression predictors using MLR models with covariates.

Model	Training Set	Lipid	*Gene*	Chr.	$R^2_{prediction}$	SNPs	N.df	F	P
EN	GTEx liver	$\sqrt{\text{HDL}}$	*NLRC5*	16	0.00	11	11	9.81	2.84E-17
	GTEx pancreas	$\sqrt{\text{HDL}}$	*NUP93**	16	0.05	60	59	3.45	4.09E-16
	DGN whole blood	$\sqrt{\text{HDL}}$	*GNAO1*	16	0.18	38	35	3.75	2.77E-12
	DGN whole blood	$\sqrt{\text{HDL}}$	*CETP**	16	0.02	30	30	3.99	5.36E-12
	GTEx whole blood	$\sqrt{\text{HDL}}$	*BBS2*	16	0.26	25	20	4.93	7.79E-12
	DGN whole blood	$\sqrt{\text{HDL}}$	*CCDC135*	16	0.25	54	54	2.96	1.71E-11
	DGN whole blood	$\sqrt{\text{HDL}}$	*SLC12A3*	16	0.01	42	42	3.18	9.68E-11
	DGN whole blood	$\sqrt{\text{HDL}}$	*MT1X*	16	0.32	36	34	3.18	3.60E-09
	DGN whole blood	$\sqrt{\text{HDL}}$	*HERPUD1**	16	0.09	15	15	4.78	4.34E-09
	DGN whole blood	$\sqrt{\text{HDL}}$	*CIAPIN1*	16	0.05	37	35	2.90	5.17E-08
	GTEx whole blood	$\sqrt{\text{HDL}}$	*CPNE2*	16	0.06	29	28	3.12	1.02E-07
	GTEx whole blood	LDL	*ZNF222*	19	0.01	32	32	2.88	1.97E-07
	GTEx pancreas	LDL	*CELSR2**	1	0.06	2	2	15.54	2.14E-07
	GTEx liver	$\sqrt{\text{HDL}}$	*MT3*	16	0.11	81	78	2.06	4.28E-07
	GTEx liver	LDL	*SORT1**	1	0.44	13	13	4.09	1.17E-06
	DGN whole blood	$log_{10}TG$	*SUGP2*	19	0.03	32	31	2.71	1.62E-06
LASSO	GTEx liver	$\sqrt{\text{HDL}}$	*NLRC5*	16	0.00	9	9	12.00	2.48E-18
	DGN whole blood	$\sqrt{\text{HDL}}$	*GNAO1*	16	0.18	15	15	6.85	1.74E-14
	GTEx whole blood	$\sqrt{\text{HDL}}$	*BBS2*	16	0.26	14	12	7.78	4.00E-14
	DGN whole blood	$\sqrt{\text{HDL}}$	*SLC12A3*	16	0.01	32	32	4.04	7.25E-13
	DGN whole blood	$\sqrt{\text{HDL}}$	*CCDC135*	16	0.25	44	44	3.36	3.12E-12
	GTEx whole blood	$\sqrt{\text{HDL}}$	*CPNE2*	16	0.06	14	14	5.71	6.94E-11
	DGN whole blood	$\sqrt{\text{HDL}}$	*HERPUD1**	16	0.08	11	11	6.34	2.72E-10
	DGN whole blood	$\sqrt{\text{HDL}}$	*CETP**	16	0.02	28	28	3.70	4.01E-10
	DGN whole blood	$\sqrt{\text{HDL}}$	*MT1X*	16	0.31	28	27	3.74	5.11E-10
	DGN whole blood	$\sqrt{\text{HDL}}$	*CIAPIN1*	16	0.05	30	29	3.41	3.82E-09
	GTEx liver	LDL	*SORT1**	1	0.48	5	5	7.99	1.97E-07
	GTEx pancreas	LDL	*CELSR2**	1	0.06	2	2	15.54	2.14E-07
	GTEx whole blood	LDL	*ZNF222*	19	0.01	28	28	3.02	2.69E-07
	DGN whole blood	LDL	*ZNF233*	19	0.01	9	9	5.27	4.43E-07
	DGN whole blood	$log_{10}TG$	*SUGP2*	19	0.03	23	23	3.19	6.36E-07
	DGN whole blood	LDL	*KLC3*	19	0.05	13	13	4.04	1.50E-06
	DGN whole blood	LDL	*PSRC1**	1	0.25	6	6	6.15	2.25E-06

Imputed gene expression-lipid associations with the smallest p values. MLR models included covariates. Gene expression predictors above the dotted line were significant. Genes marked by an asterisk (*) are known quantitative trait loci for the lipid trait (NHGRI-EBI GWAS Catalog; MacArthur et al. (2017)). Gene, gene expression predictor; EN, elastic net; LASSO, least absolute shrinkage and selection operator; Chr., chromosome; $R^2_{prediction}$, 10-fold cross-validated R^2 for predictive performance; SNPs, the number of variants in the expression predictor; N.df, numerator degrees of freedom; LDL, low-density lipoprotein cholesterol; HDL, high-density lipoprotein cholesterol; TG, triglycerides.

Rejection of the null hypothesis using the F-test indicated that at least one of the SNPs in the set for the expression predictor explained the variation in the lipid trait.

Table 4.4: Genomic regions of the gene expression predictors that associated with lipid traits.

Training Set	Lipid	*Gene*	Chr.	Position	
				Start	End
DGN whole blood	*LDL*	*PSRC1*	1	109369915	109838918
	$\sqrt{\text{HDL}}$	*GNAO1*	16	55465544	57276202
	$\sqrt{\text{HDL}}$	*MT1X*	16	55726462	57664105
	$\sqrt{\text{HDL}}$	*SLC12A3*	16	55926873	57922948
	$\sqrt{\text{HDL}}$	*CETP*	16	56006578	57973989
	$\sqrt{\text{HDL}}$	*CIAPIN1*	16	56578562	58385455
	$\sqrt{\text{HDL}}$	*CCDC135*	16	56729963	58672763
	$\sqrt{\text{HDL}}$	*HERPUD1*	16	56745758	57326269
	$\log_{10}$ TG	*SUGP2*	19	18107913	20056348
GTEx whole blood	TC	*TADA2B*	4	6055158	8018565
	$\sqrt{\text{HDL}}$	*BBS2*	16	55798370	57156366
	$\sqrt{\text{HDL}}$	*CPNE2*	16	56402544	58145585
	LDL	*ZNF222*	19	43548332	45520285
GTEx pancreas	*LDL*	*CELSR2*	1	109817192	109817590
	$\sqrt{\text{HDL}}$	*NUP93*	16	55845351	57861649
GTEx liver	*LDL*	*PSRC1*	1	108859314	110650082
	LDL	*CELSR2*	1	108879960	110799893
	LDL	*SORT1*	1	109366554	110715768
	$\sqrt{\text{HDL}}$	*MT3*	16	55631508	57523861
	$\sqrt{\text{HDL}}$	*NLRC5*	16	56240331	57762401

HDL, high-density lipoprotein cholesterol; LDL, low-density lipoprotein cholesterol; TG, triglycerides; Gene, gene expression predictor; Chr., chromosome; Start and End, range of positions in base pairs for the expression predictor.

Thus, the higher number of significant expression predictors than would be expected under the null hypothesis using the MLR model and DGN whole blood may be related to one or more SNPs that are present in multiple copies across expression predictors. For the significant $\sqrt{\text{HDL}}$ associations, no variant was present in all of the DGN whole blood expression predictors. In fact, analysis of the set of SNPs for the significant associations with the largest expression predictors from DGN whole blood derived from elastic net showed that most of SNPs for a given expression predictor were not shared by any other expression predictors (Figure 4.8).

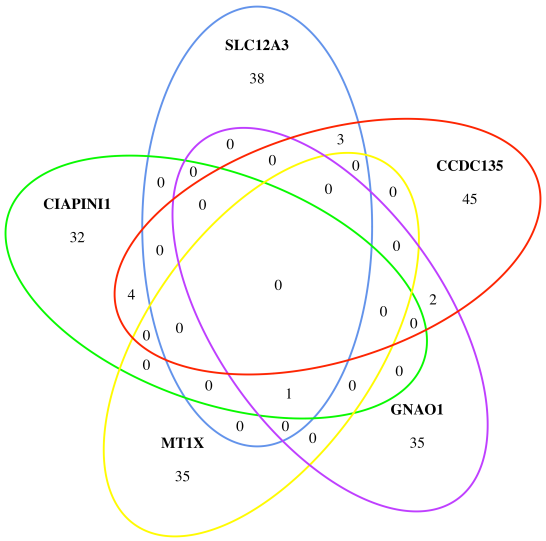

Figure 4.8: Intersection of SNPs from $\sqrt{HDL}$ associated DGN whole blood expression predictors derived using elastic net. No SNPs intersected all expression predictors.

Table 4.5: 10-fold cross-validated prediction R^2 of top gene expression predictors in the training sets derived using elastic net.

Gene	DGN whole blood	GTEx whole blood	GTEx pancreas	GTEx liver
NLRC5	0.261	0.067	0.001	0.000
SLC12A3	0.013	0.009	0.026	
CETP	0.022	0.001	0.006	0.050
HERPUD1	0.086			
CELSR2	0.160		0.056	0.372
SORT1	0.017		0.049	0.440
PSRC1	0.252	0.018	0.150	0.385

Missing values denote training sets without an expression predictor for the gene.

Despite the fact that the $NLRC5$ expression predictor from GTEx liver derived using elastic net had no SNPs in common with the DGN whole blood $CETP$ expression predictor derived using elastic net, four of the SNPs in the $NLRC5$ expression predictor identified with $CETP$ on the GeneCards Human Gene Database (Rebhan et al., 1997) and associations with HDL for two of these SNPs (rs11076175: $CETP$ and rs1800775: $CETP$, $HERPUD1$, $SLC12A3$, $NUP93$) were previously reported on the NHGRI-EBI GWAS Catalog.

A comparison of the 10-fold cross-validated R^2 for predictive performance ($R^2_{prediction}$) for significant expression predictors across training sets showed marked differences. With respect to the expression predictor for $NLRC5$, $R^2_{prediction}$ for DGN whole blood, GTEx whole blood, GTEx pancreas and GTEx liver were 0.26, 0.07, 0.001, 3.24×10^{-05}, respectively (Table 4.5) and none of the SNPs in the training sets with the lowest $R^2_{prediction}$ (GTEx pancreas and GTEx liver) intersected the training sets with the highest $R^2_{prediction}$ (DGN whole blood or GTEx whole blood) (Figure 4.9).

The negligible $R^2_{prediction}$ for $NLRC5$ expression is consistent with the fact that none of the SNPs for the $NLRC5$ expression predictor from the GTEx liver data set derived using elastic net were reported with $NLRC5$ in GeneCards and GWAS

catalogs, raising the possibility that the significantly associated gene with $\sqrt{\text{HDL}}$ is $CETP$ and not $NLRC5$. The significant associations with $\sqrt{\text{HDL}}$, however, leave little doubt about a relationship between HDL and one or more genes from the cytogenetic band 16q13 and provide support for the results of others (Shirali et al., 2016; Zhang et al., 2015).

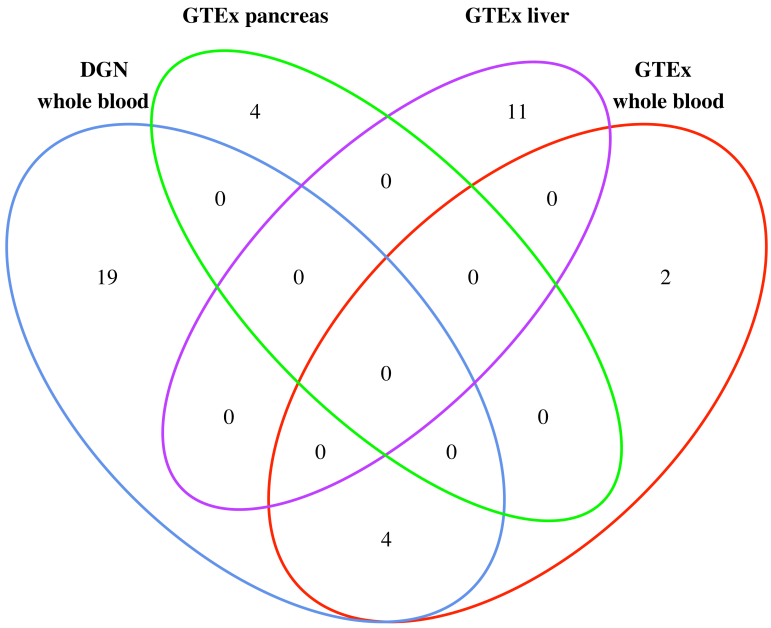

Figure 4.9: Training set intersection of SNPs for the $NLRC5$ expression predictor derived using elastic net. Only DGN whole blood and GTEx whole blood shared SNPs for $NLRC5$.

Similar to Chapter 3, regression diagnostics for models 4.1 and 4.2 did not reveal any model violations (Figure 4.10).

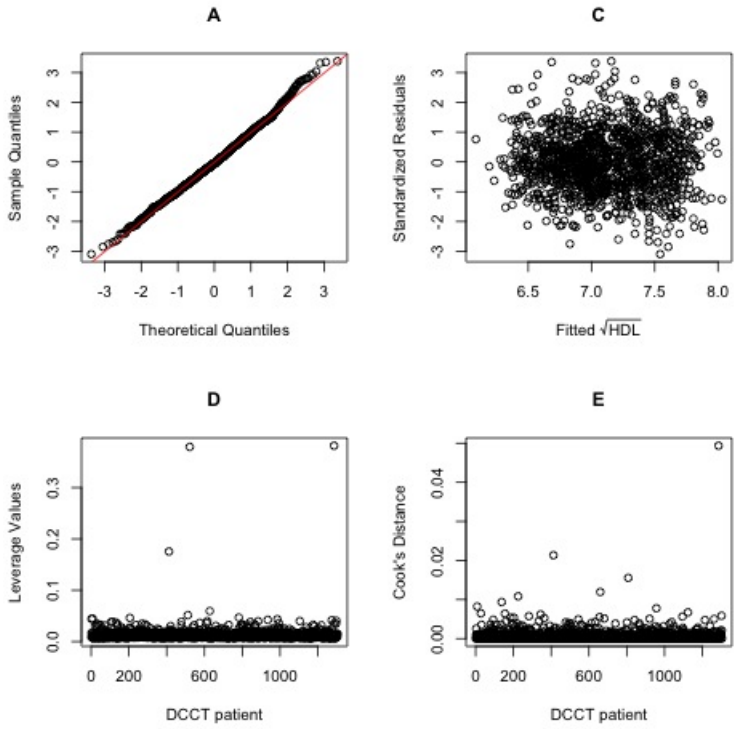

Figure 4.10: Diagnostic plots for the MLR model with covariates relating $\sqrt{\text{HDL}}$ with the $NLRC5$ expression predictor from GTEx liver derived using elastic net. (A) Normal probability plot of the standardized residuals; (B) Scatter plot of the standardized residuals against the fitted values; (C) Leverage values; (D) Cook's distance; no points exceeded the threshold of $C_i = 0.96$ which is beyond the axis of the plot.

A comparison of the matched associations from Chapter 3 and Chapter 4 showed weak $-\log p$ value correlations. The correlations were weakest for DGN whole blood and strongest for GTEx liver across lipids indicating that the outcomes of PrediXcan and MLR were predominately dissimilar even though the models were based on the same set of SNPs (Figure 4.11). The MLR models of Chapter 4 had more significant associations than the PrediXcan models of Chapter 3, and the smaller p values from the models of Chapter 4 were clearly seen with the significant $\sqrt{\text{HDL}}$ associations with SNPs and covariates (Figure 4.11).

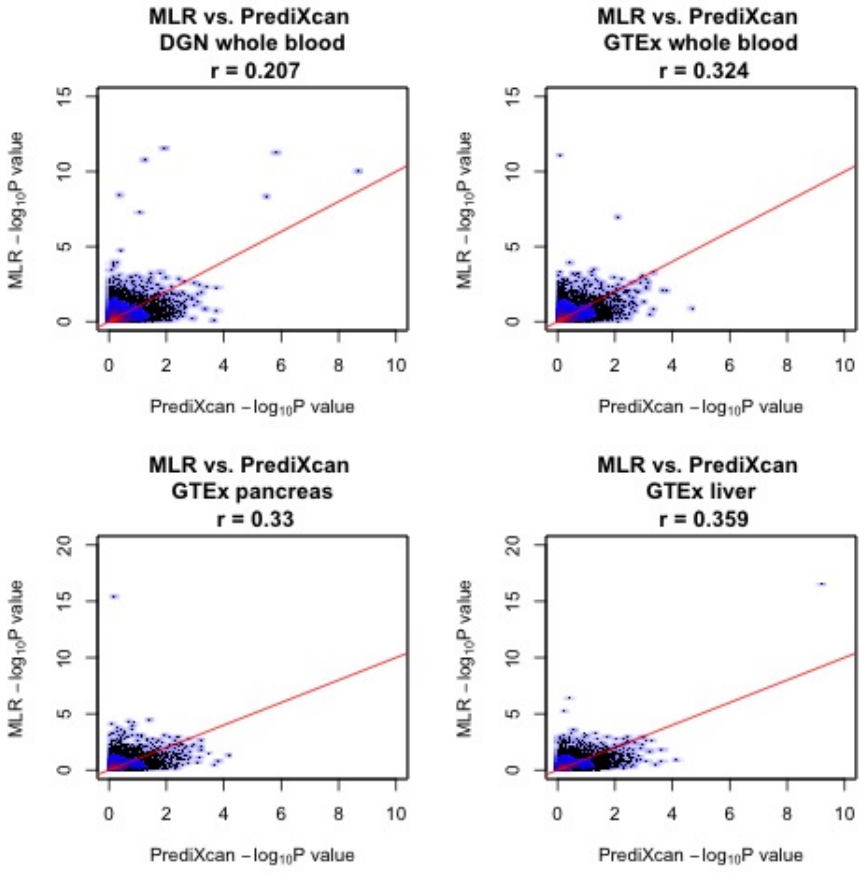

Figure 4.11: MLR versus PrediXcan $-\log p$ value comparison for $\sqrt{\text{HDL}}$ associations with SNPs and covariates. Comparisons for all training sets derived using elastic net are shown. Red line is $y = x$ and Spearman's rank correlation coefficient (r).

Chapter 5

Discussion

Genome-wide association studies (GWAS) have identified many common single nucleotide polymorphisms (SNPs) and genes associated with lipid traits, but they have facilitated an understanding of the mechanisms underlying the associations to a lesser degree (Ridker et al., 2009; Kathiresan et al., 2009; Teslovich et al., 2010; Willer et al., 2013; Zhang et al., 2015; Shirali et al., 2016; Nagy et al., 2017). PrediXcan is an imputed gene expression-trait association method that enables tests of association between the predicted transcriptome and a phenotype of interest (Gamazon et al., 2015). Herein, we applied the PrediXcan model to four lipid traits using genome wide data from 1304 patients of European ancestry from the Diabetes Control and Complications Trial (DCCT). Tests of association with one of total cholesterol (TC), low-density lipoprotein cholesterol (LDL), high-density lipoprotein cholesterol (HDL) and triglycerides (TG) and all of the gene expression predictors from the training sets: DGN whole blood, GTEx whole blood, GTEx pancreas, and GTEx liver were conducted using the Student's t to see if the PrediXcan model could verify known and identify novel eQTL-lipid associations. The dependence of the lipid traits on the set of SNPs from a gene expression predictor was also modelled using multiple linear

regression (MLR) and the F test.

Significant inverse relationships with LDL and gene expression predictors on Chromosome 1 and $\sqrt{\text{HDL}}$ and gene expression predictors on Chromosome 16 were identified at a very stringent Bonferroni correction threshold of 1.53×10^{-7}, for the collective number of tests from the two prediction models, four training sets, four lipids, and expression predictors in the training set (Figure 3.1). The use of gene expression predictors selected using elastic net or LASSO did not appreciably change the estimated GReX or p values when the PrediXcan model was applied; however, the p values did not coincide when pairwise comparisons of matched tests of association between training sets were made. Moreover, the responsible gene expression predictor for the association differed by training set for a given significant chromosomal region. The effect of the covariates (age, gender, duration of IDDM, cohort, treatment, and the interaction between cohort and treatment) on the p values was most pronounced for associations with $\sqrt{\text{HDL}}$.

Although significant associations with LDL and $\sqrt{\text{HDL}}$ and genes on Chromosomes 1 and 16, respectively, were detected in both PrediXcan and MLR models, the p values for similar tests of association were only weakly correlated. Furthermore, the PrediXcan model detected suggestive negative relationships with total cholesterol and expression predictors for genes on Chromosomes 1 and 4 that were not observed with MLR models, and an association with triglycerides and an expression predictor for a gene on Chromosome 19 was only apparent using MLR models. Associations with lipids and genomic regions on Chromosomes 1, 4, 16, and 19 are well documented in the literature (Dastani et al., 2010; Willer et al., 2013; Zhang et al., 2015; Kurano et al., 2016; Shirali et al., 2016; Gusev et al., 2016) and hence both the PrediXcan and MLR models verified known loci. More gene expression predictors were significantly associated with $\sqrt{\text{HDL}}$ using MLR models but the p values from MLR models did

not follow a uniform distribution as well as those from PrediXcan models. A recent publication by van Iterson et al. (2017) suggests that more phenotype associations are to be expected in TWAS and that the commonly practiced genomic control correction for the inflation factor for p values in GWAS is too conservative for TWAS. Hence, more of the very smallest p values from associations with DGN whole blood expression predictors than were expected under the null distribution may be interesting rather than concerning.

While the mean 10-fold cross-validated R^2 for predictive performance was highest for gene expression predictors from DGN whole blood, its gene expression predictors did not always have the strongest predictive performance for gene expression when compared with analogous gene expression predictors from the other training sets. The inverse associations with LDL and the expression predictors for $CELSR2$ and $PSRC1$ on Chromosome 1 in more than one training set with strong predictive performance provides support for the negative regulation of LDL by $CELSR2$ and $PSRC1$ expression in the liver. The same is true for $SORT1$, the third member of the three gene haplotype block on Chromosome 1p13.3, which was previously shown to be associated with decreased serum LDL in many independent GWAS (Willer et al., 2008; Wallace et al., 2008; Schadt et al., 2008; Kathiresan et al., 2009; Teslovich et al., 2010; Surakka et al., 2015; Kurano et al., 2016) in addition to lower risk for cardiovascular disease (Schunkert et al., 2011) and enhanced response of LDL to statin therapy (Postmus et al., 2014). Suggestive associations with LDL and the $SORT1$ expression predictors from GTEx liver using the PrediXcan and MLR models are consistent with the literature that suggests $SORT1$ is the main gene at the locus (Kjolby et al., 2010; Linsel-Nitschke et al., 2010; Musunuru et al., 2010) and that it has an important role in the liver (Folkersen et al., 2010). Moreover, expression levels for $SORT1$ and $CELSR2$ were negatively associated with LDL in human liver samples and proposed

as candidate genes for diabetes, obesity and atherosclerosis (Schadt et al., 2008; Breitling et al., 2015). More recent GWAS studies have also shown a relationship for TC and LDL with $CELSR2$ (Below et al., 2016). Consequently, the 1p13.3 locus is currently considered to be the most strongly associated with LDL in the genome and to primarily affect changes in very small, atherogenic LDL subclasses (Musunuru et al., 2010).

Associations using the PrediXcan model revealed suggestive relationships with the expression predictors for $CETP$, $HERPUD1$ using DGN whole blood, a significant association with the expression predictor for $SLC12A3$ using DGN whole blood, and a significant association with the expression predictor for $NLRC5$ using GTEx liver, and all associations were inversely related to $\sqrt{\text{HDL}}$. Cholesteryl ester transfer protein (CETP) is a well known modulator of HDL and CETP deficient individuals have high HDL levels (Brown et al., 1989; Inazu et al., 1994). A study by Ridker et al. (2009) showed many genome-wide significant associations with HDL within a 242 K base pair region of Chromosome 16 encompassing the $CETP$ locus. The authors noted that the majority of the genome-wide significant SNPs clustered around the $CETP$ gene, but three SNPs mapped to $NUP93$ (encoding a nuclear pore protein), and six mapped to $SLC12A3$ (encoding a sodium cotransporter) and $HERPUD1$ (encoding a endoplasmic reticulum stress inducible protein) and thus they proposed long range linkage disequilibrium at the locus with $CETP$ as a causal mechanism for association. Mutant mouse models for $HERPUD1$ and $SLC12A3$ suggest that these genes are not candidates for the regulation of HDL levels, but a role for $HERPUD1$ in glucose tolerance was demonstrated (Eura et al., 2012). Furthermore, human diseases associated with $SLC12A3$ such as Gitelman syndrome present with aberrant electrolyte homeostasis rather than dyslipidemia (Glaudemans et al., 2012).

The 10-fold cross-validated R^2 for the expression predictors on Chromosome 16

were much smaller than those for the expression predictors that associated with LDL raising the possibly that many more genes influence the expression of HDL. Associations using the MLR model revealed more significant relationships with $\sqrt{\text{HDL}}$ and five of the significant associations from expression predictors $BBS2$, $MT1X$, $NUP93$, $HERPUD1$, and $NLRC5$ were also significantly associated with HDL in a recent imputed TWAS (Gusev et al., 2016). Bardet-Biedl Syndrome (BBS) is an autosomal recessive disorder and both humans and mice deficient for BBS2 present with obesity (Benzinou et al., 2006; Rahmouni et al., 2008) suggesting that a relationship with HDL is plausible. Conversely, neither $MT1X$ (Metallothionein 1X) nor $NLRC5$ (NOD-Like Receptor CARD domain containing 5) are obvious candidates for HDL regulation since the former fosters cellular response to the zinc ion and the latter functions within the immune system.

While the similar findings of multiple investigative groups supports the use of the PrediXcan gene expression predictors for lipid associations, the methods presented in this thesis should be replicated using an independent GWAS data set from patients with type 1 diabetes such as the Wisconsin Epidemiologic Study of Diabetic Retinopathy (WESDR), the Coronary Artery Calcification in Type 1 Diabetes (CACTI), and the Renin-Angiotensin System Study (RASS). Replication using the PrediXcan model enables the direction and magnitude of effect to be observed (Gamazon et al., 2015) from associations with low genetic control inflation for p values and thus may be more informative than replication studies that use the MLR model. The PrediXcan model also enables polygenic models using multi-gene expression predictors and lipid traits to be tested in a cost-effective manner. Such models could help to elucidate the genetic determinants of dyslipidemia in IDDM and further the field of personalized medicine.

Appendix A

Supplementary Tables

Table A.1: PredictDB training sets.

Tissue	Model	Training Set	Version Date	Notes	Name
Liver	EN	GTEx	2015-11-12	run c	predb00000030
	LASSO	GTEx	2015-11-12	run c	predb00000002
Pancreas	EN	GTEx	2015-11-12	run c	predb00000072
	LASSO	GTEx	2015-11-12	run c	predb00000064
Blood	EN	GTEx	2015-11-12	run c	predb00000019
		DGN	2015-11-12	run c, unscaled	predb00000061
	LASSO	GTEx	2015-11-12	run c	predb00000058
		DGN	2015-11-12	run c	predb00000059

http://hakyimlab.org/predictdb/

Table A.2: Top gene expression predictors using PrediXcan.

Model	Training Set	Lipid	*Gene*	Chr.	$R^2_{prediction}$	SNPs	β_1	P	R^2
Elastic net	GTEx liver	$\sqrt{\text{HDL}}$	*NLRC5*	16	0.00	11	-0.73	7.10E-08	0.022
	GTEx pancreas	LDL	*CELSR2**	1	0.06	2	-27.81	5.45E-07	0.019
	GTEx liver	LDL	*SORT1**	1	0.44	13	-7.28	9.97E-07	0.018
	DGN whole blood	LDL	*PSRC1**	1	0.25	11	-7.65	4.07E-06	0.016
	GTEx liver	LDL	*CELSR2**	1	0.37	23	-6.47	4.44E-06	0.016
LASSO	GTEx liver	$\sqrt{\text{HDL}}$	*NLRC5*	16	0.00	9	-0.68	2.31E-07	0.020
	GTEx pancreas	LDL	*CELSR2**	1	0.06	2	-25.32	4.22E-07	0.019
	GTEx liver	LDL	*SORT1**	1	0.48	5	-7.13	7.21E-07	0.019
	GTEx liver	LDL	*CELSR2**	1	0.40	10	-7.51	8.02E-07	0.019
	GTEx liver	LDL	*PSRC1**	1	0.38	25	-5.58	3.52E-06	0.016

Imputed gene expression-lipid associations with the smallest p values. Gene expression predictors above the dotted line were significant. Genes marked by an asterisk (*) are known quantitative trait loci for the lipid trait (NHGRI-EBI GWAS Catalog; MacArthur et al. (2017)). Gene, gene expression predictor; EN, elastic net; LASSO, least absolute shrinkage and selection operator; Chr., chromosome; $R^2_{prediction}$, 10-fold cross-validated R^2 for predictive performance; SNPs, the number of variants in the expression predictor; LDL, low-density lipoprotein cholesterol; HDL, high-density lipoprotein cholesterol.

Table A.3: Top gene expression predictors using MLR.

Model	Training Set	Lipid	*Gene*	Chr.	$R^2_{prediction}$	SNPs	N.df	D.df	F	P
EN	GTEx pancreas	$\sqrt{\text{HDL}}$	*NUP93**	16	0.05	60	59	1244	2.98	1.83E-12
	GTEx liver	$\sqrt{\text{HDL}}$	*NLRC5*	16	0.00	11	11	1292	7.36	2.58E-12
	GTEx whole blood	$\sqrt{\text{HDL}}$	*BBS2*	16	0.26	25	20	1283	3.89	1.77E-08
	DGN whole blood	$\sqrt{\text{HDL}}$	*CETP**	16	0.02	30	30	1273	3.03	9.71E-08
	DGN whole blood	$\sqrt{\text{HDL}}$	*SLC12A3*	16	0.01	42	42	1261	2.55	3.02E-07
	DGN whole blood	$\sqrt{\text{HDL}}$	*CCDC135*	16	0.25	54	54	1249	2.32	3.76E-07
	DGN whole blood	$\sqrt{\text{HDL}}$	*GNAO1*	16	0.18	38	35	1268	2.70	4.60E-07
	GTEx whole blood	LDL	*ZNF222*	19	0.01	32	32	1271	2.79	5.02E-07
	DGN whole blood	$log_{10}TG$	*SUGP2*	19	0.03	32	31	1272	2.77	8.55E-07
	GTEx pancreas	LDL	*CELSR2**	1	0.06	2	2	1301	13.89	1.08E-06
	DGN whole blood	$\sqrt{\text{HDL}}$	*MT1X*	16	0.32	36	34	1269	2.62	1.44E-06
	DGN whole blood	$log_{10}TG$	*ZNRF1*	16	0.25	19	19	1284	3.20	4.08E-06
	GTEx liver	LDL	*SORT1**	1	0.44	13	13	1290	3.80	4.90E-06
LASSO	GTEx liver	$\sqrt{\text{HDL}}$	*NLRC5*	16	0.00	9	9	1294	8.86	4.85E-13
	GTEx whole blood	$\sqrt{\text{HDL}}$	*BBS2*	16	0.26	14	12	1291	5.82	6.83E-10
	DGN whole blood	$\sqrt{\text{HDL}}$	*GNAO1*	16	0.18	15	15	1288	4.80	3.82E-09
	DGN whole blood	$\sqrt{\text{HDL}}$	*SLC12A3*	16	0.01	32	32	1271	3.24	4.89E-09
	DGN whole blood	$\sqrt{\text{HDL}}$	*CCDC135*	16	0.25	44	44	1259	2.64	5.63E-08
	GTEx whole blood	$\sqrt{\text{HDL}}$	*CPNE2*	16	0.06	14	14	1289	4.30	1.67E-07
	DGN whole blood	LDL	*ZNF233*	19	0.01	9	9	1294	5.46	2.21E-07
	DGN whole blood	$\sqrt{\text{HDL}}$	*MT1X*	16	0.31	28	27	1276	3.08	2.33E-07
	DGN whole blood	$log_{10}TG$	*SUGP2*	19	0.03	23	23	1280	3.28	2.99E-07
	GTEx liver	LDL	*SORT1**	1	0.48	5	5	1298	7.50	5.91E-07
	DGN whole blood	$\sqrt{\text{HDL}}$	*CETP**	16	0.02	28	28	1275	2.91	7.03E-07
	GTEx whole blood	LDL	*ZNF222*	19	0.01	28	28	1275	2.90	8.00E-07
	GTEx pancreas	LDL	*CELSR2**	1	0.06	2	2	1301	13.89	1.08E-06
	DGN whole blood	$\sqrt{\text{HDL}}$	*CIAPIN1*	16	0.05	30	29	1274	2.82	1.10E-06
	DGN whole blood	LDL	*KLC3*	19	0.05	13	13	1290	3.97	2.19E-06
	DGN whole blood	$\sqrt{\text{HDL}}$	*HERPUD1**	16	0.08	11	11	1292	4.24	3.24E-06
	DGN whole blood	$log_{10}TG$	*ZNRF1*	16	0.24	16	16	1287	3.50	3.35E-06

Imputed gene expression-lipid associations with the smallest p values. Gene expression predictors above the dotted line were significant. Genes marked by an asterisk (*) are known quantitative trait loci for the lipid trait (NHGRI-EBI GWAS Catalog; MacArthur et al. (2017)). Gene, gene expression predictor; EN, elastic net; LASSO, least absolute shrinkage and selection operator; Chr., chromosome; $R^2_{prediction}$, 10-fold cross-validated R^2 for predictive performance; SNPs, the number of variants in the expression predictor; N.df, numerator degrees of freedom; D.df, denominator degrees of freedom; LDL, low-density lipoprotein cholesterol; HDL, high-density lipoprotein cholesterol; TG, triglycerides.

Table A.4: Model comparison for significant $\sqrt{\text{HDL}}$ regressions with gene expression predictors from DGN whole blood.

Model	*Gene*	N.df	P	R^2	R_a^2	AIC	BIC	S
Elastic net	*CCDC135*	54	1.71E-11	0.263	0.227	2801.75	3122.49	0.692
LASSO		44	3.12E-12	0.256	0.226	2794.36	3063.36	0.693
Elastic net	*CETP*	30	5.36E-12	0.240	0.218	2793.89	2990.47	0.696
LASSO		28	4.01E-10	0.231	0.210	2805.29	2991.53	0.700
Elastic net	*CIAPIN1*	35	5.17E-08	0.230	0.205	2820.75	3043.20	0.702
LASSO		29	3.82E-09	0.228	0.207	2811.68	3003.08	0.701
Elastic net	*GNAO1*	35	2.77E-12	0.246	0.222	2792.61	3015.05	0.694
LASSO		15	1.74E-14	0.230	0.217	2781.11	2900.10	0.696
Elastic net	*HERPUD1*	15	4.34E-09	0.212	0.199	2810.74	2929.72	0.704
LASSO		11	2.72E-10	0.211	0.200	2804.77	2903.06	0.704
Elastic net	*MT1X*	34	3.60E-09	0.234	0.209	2812.60	3029.87	0.700
LASSO		27	5.11E-10	0.229	0.209	2805.82	2986.88	0.700
Elastic net	*SLC12A3*	42	9.68E-11	0.248	0.219	2803.94	3062.59	0.696
LASSO		32	7.25E-13	0.245	0.223	2788.70	2995.62	0.694

N.df, numerator degrees of freedom; P, p value, R^2, coefficient of determination; R_a^2, adjusted coefficient of determination; AIC, Akaike's Information Criterion; BIC, Bayesian Information Criterion; S, residual standard error.

Table A.5: SNPs with zero variance in DCCT. SNPs from GTEx whole blood derived using elastic net were used for extraction.

Chr.	*Gene*	$\widehat{GReX}$	SNP	Weight	MAF	Dose
1	*TNNI3K*	-0.07	rs11801227	-0.04	0	0
2	*WDFY1*	-0.44	rs16866359	-0.22	0	0
2	*SNX17*	0.00	rs7579203	-0.02	0	0
2	*SNX17*	0.00	rs6749203	-0.02	0	0
2	*CXCR4*	0.14	rs4988176	0.07	0	0
3	*RNF123*	0.11	rs636168	0.06	1	2
3	*RNF123*	0.11	rs1208923	-0.05	1	2
3	*RNF123*	0.11	rs4450808	-0.06	1	2
3	*CD200R1L*	0.00	rs774750	0.36	0	0
7	*ZNF789*	0.00	rs10266141	-0.15	0	0
8	*CPA6*	-0.19	rs11985230	-0.09	0	0
10	*USMG5*	0.00	rs4288700	-0.31	1	2
15	*WDR72*	-0.09	rs16966558	-0.02	0	0
15	*WDR72*	-0.09	rs16966567	-0.02	0	0
16	*NPIPA7*	0.00	rs11861014	0.09	0	0
17	*SPATA22*	-0.12	rs9902438	-0.03	0	0
17	*SPATA22*	-0.12	rs9894685	-0.03	0	0
17	*SPATA22*	-0.12	rs7210926	0.03	0	0
17	*HN1*	-0.19	rs11872088	-0.10	0	0
20	*ABHD12*	0.10	rs7265940	0.02	0	0
20	*ABHD12*	0.10	rs8126155	0.03	0	0
20	*ABHD12*	0.10	rs7264886	-0.02	0	0

Table A.6: SNPs with zero variance in DCCT. SNPs from GTEx pancreas derived using elastic net were used for extraction.

Chr.	*Gene*	$\widehat{GReX}$	SNP	Weight	MAF	Dose
1	*PHTF1*	-0.09	rs7554422	-0.04	0	0
1	*PHTF1*	-0.09	rs7545980	0.05	0	0
1	*TARDBP*	0.32	rs2506894	-0.09	1	2
1	*TARDBP*	0.32	rs2506903	0.08	1	2
1	*TARDBP*	0.32	rs6540936	0.08	1	2
1	*GPATCH3*	-0.55	rs12083830	-0.28	0	0
3	*USP4*	0.00	rs6776029	-0.19	0	0
3	*C3orf33*	0.24	rs11927323	0.12	0	0
3	*SCN10A*	-1.25	rs10514702	-0.09	0	0
3	*SCN10A*	-1.25	rs17039230	-0.10	0	0
3	*SCN10A*	-1.25	rs1877554	-0.06	0	0
3	*SCN10A*	-1.25	rs17039124	-0.09	0	0
3	*SCN10A*	-1.25	rs7653831	-0.10	0	0
3	*SCN10A*	-1.25	rs1890516	-0.08	0	0
3	*SCN10A*	-1.25	rs17039169	0.10	0	0
3	*SCN10A*	-1.25	rs7612859	-0.10	0	0
7	*ANKMY2*	0.00	rs10228129	-0.06	0	0
7	*SLC29A4*	0.00	rs10215856	0.13	0	0
8	*VDAC3*	0.00	rs11995024	0.64	0	0
10	*ANK3*	0.17	rs7069890	-0.04	0	0
10	*ANK3*	0.17	rs10994325	0.04	0	0
10	*ANK3*	0.17	rs12246937	0.04	0	0
10	*DUPD1*	-1.19	rs11001442	-0.30	0	0
10	*DUPD1*	-1.19	rs12164797	-0.29	0	0
13	*MTIF3*	0.00	rs8000938	-0.11	0	0
14	*APOPT1*	-0.12	rs11160717	-0.06	0	0
19	*GAMT*	0.00	rs10411834	0.09	0	0
19	*DAND5*	-0.13	rs10410429	-0.06	0	0
19	*KIR3DL2*	-0.87	rs7256388	-0.43	0	0
21	*SCAF4*	0.19	rs8128009	0.10	0	0
21	*KRTAP10-8*	-0.23	rs4818719	-0.12	0	0

Table A.7: SNPs with zero variance in DCCT. SNPs from GTEx liver derived using elastic net were used for extraction.

Chr.	*Gene*	$\widehat{GReX}$	SNP	Weight	MAF	Dose
1	*LCE6A*	-0.26	rs16835548	-0.13	0	0
4	*KIAA1430*	-0.20	rs2696041	0.10	1	2
4	*KIAA1430*	-0.20	rs2696042	0.10	1	2
4	*KIAA1430*	-0.20	rs2705887	-0.10	1	2
7	*EIF2AK1*	0.00	rs10215856	18.43	0	0
19	*NLRP8*	-1.02	rs890868	-0.51	1	2

Table A.8: DGN whole blood *PSRC*1 derived using elastic net.

Chr.	SNP	Weight	RA	EA	Position	A1	A2	MAF
1	rs599839	-0.19	G	A	109822166	G	A	0.79
1	rs583104	-0.16	G	T	109821307	G	T	0.79
1	rs602633	-0.15	T	G	109821511	T	G	0.79
1	rs672569	-0.10	A	G	109827253	A	G	0.84
1	rs629301	-0.05	G	T	109818306	G	T	0.79
1	rs646776	-0.04	C	T	109818530	C	T	0.79
1	rs629001	-0.02	C	T	109838918	C	T	0.93
1	rs660240	-0.01	T	C	109817838	T	C	0.80
1	rs7551421	-0.01	G	T	109369915	G	T	0.61
1	rs7528419	0.06	A	G	109817192	A	G	0.21
1	rs12740374	0.06	G	T	109817590	G	T	0.21

Chr., chromosome; RA, reference allele from PredictDB; EA, effect allele from PredictDB; A1, effect allele from DCCT; A2, reference allele from DCCT; MAF, minor allele frequency

Table A.9: GTEx liver *SORT*1 derived using elastic net.

Chr.	SNP	Weight	RA	EA	Position	A1	A2	MAF
1	rs12740374	-0.40	T	G	109817590	G	T	0.21
1	rs17038491	-0.20	T	C	109659958	C	T	0.00
1	rs602265	-0.09	A	G	109781581	A	G	1.00
1	rs12063647	-0.03	A	G	110318221	A	G	0.24
1	rs17038458	-0.01	A	G	109672215	A	G	0.06
1	rs12116787	-0.01	A	C	110715768	A	C	0.06
1	rs4970834	-0.00	T	C	109814880	C	T	0.18
1	rs646776	0.00	T	C	109818530	C	T	0.79
1	rs629301	0.00	T	G	109818306	G	T	0.79
1	rs7529592	0.00	T	C	109366554	T	C	0.10
1	rs17035630	0.03	A	G	109810981	G	A	0.11
1	rs611917	0.04	A	G	109815252	A	G	0.32
1	rs7528419	0.40	A	G	109817192	A	G	0.21

Chr., chromosome; RA, reference allele from PredictDB; EA, effect allele from PredictDB; A1, effect allele from DCCT; A2, reference allele from DCCT; MAF, minor allele frequency

Table A.10: GTEx liver *PSRC*1 derived using elastic net.

Chr.	SNP	Weight	RA	EA	Position	A1	A2	MAF
1	rs12740374	-0.34	T	G	109817590	G	T	0.21
1	rs17038458	-0.20	A	G	109672215	A	G	0.06
1	rs17616480	-0.16	T	C	109100633	T	C	0.11
1	rs369741	-0.11	A	G	110427992	A	G	0.03
1	rs4970834	-0.10	T	C	109814880	C	T	0.18
1	rs4484951	-0.07	T	C	109270563	T	C	0.01
1	rs17646731	-0.07	A	G	109919525	G	A	0.05
1	rs1277205	-0.05	T	G	109393437	G	T	0.89
1	rs12402346	-0.04	A	G	110038019	G	A	0.01
1	rs12036884	-0.04	A	G	109727529	G	A	0.06
1	rs7517648	-0.03	T	G	109318228	G	T	0.05
1	rs579035	-0.03	A	C	110346885	C	A	0.52
1	rs12032606	-0.02	A	C	110645614	C	A	0.07
1	rs12124705	-0.01	A	G	108872823	G	A	0.18
1	rs12406978	-0.01	A	C	110043187	C	A	0.01
1	rs4839135	-0.00	T	G	110650082	G	T	0.07
1	rs1417300	0.00	T	C	108871245	T	C	0.33
1	rs673792	0.00	A	G	110646283	A	G	0.49
1	rs11185315	0.00	A	G	108870525	A	G	0.33
1	rs4970821	0.00	T	C	108859314	T	C	0.33
1	rs617477	0.01	T	C	110649693	T	C	0.49
1	rs7529976	0.01	T	G	108866862	G	T	0.07
1	rs617126	0.01	T	G	110649713	T	G	0.49
1	rs583104	0.01	T	G	109821307	G	T	0.79
1	rs629301	0.01	T	G	109818306	G	T	0.79
1	rs646776	0.01	T	C	109818530	C	T	0.79
1	rs11576956	0.01	A	G	110646858	A	G	0.36
1	rs3768490	0.02	T	G	110259016	G	T	0.33
1	rs3818562	0.02	A	G	110300441	G	A	0.49
1	rs1149144	0.03	T	C	109391027	T	C	0.89
1	rs599839	0.05	A	G	109822166	G	A	0.79
1	rs17035415	0.05	A	C	109787493	C	A	0.19
1	rs585362	0.06	T	C	109789795	C	T	0.86
1	rs17647543	0.06	T	C	109964605	T	C	0.05
1	rs17586966	0.07	T	C	109955569	T	C	0.05
1	rs12142041	0.08	T	C	110505038	T	C	0.09
1	rs17035630	0.11	A	G	109810981	G	A	0.11
1	rs12403287	0.13	A	G	108873582	A	G	0.33
1	rs504316	0.13	A	G	109663420	G	A	0.07
1	rs7528419	0.34	A	G	109817192	A	G	0.21

Chr., chromosome; RA, reference allele from PredictDB; EA, effect allele from PredictDB; A1, effect allele from DCCT; A2, reference allele from DCCT; MAF, minor allele frequency

Table A.11: GTEx liver $CELSR2$ derived using elastic net.

Chr.	SNP	Weight	RA	EA	Position	A1	A2	MAF
1	rs12740374	-0.39	T	G	109817590	G	T	0.21
1	rs2275123	-0.08	A	G	110458234	G	A	0.17
1	rs1504405	-0.07	T	C	110482486	C	T	0.19
1	rs17025216	-0.07	T	C	110486839	C	T	0.18
1	rs9662782	-0.04	T	G	109445706	T	G	0.09
1	rs6695237	-0.03	A	G	110415626	G	A	0.47
1	rs11102632	-0.03	T	C	109453129	T	C	0.14
1	rs12120692	-0.03	T	C	109436391	T	C	0.14
1	rs7522260	-0.01	T	C	109366370	C	T	0.69
1	rs3738759	-0.00	T	C	110599796	C	T	0.07
1	rs28553535	-0.00	A	C	109612484	C	A	0.07
1	rs17574954	0.00	A	G	108879960	G	A	0.08
1	rs12059276	0.02	T	C	110273541	C	T	0.07
1	rs453577	0.02	A	G	110426778	G	A	0.65
1	rs3093037	0.03	T	C	110471906	C	T	0.19
1	rs1105803	0.05	T	C	110428878	T	C	0.37
1	rs11102072	0.06	A	C	110799893	C	A	0.31
1	rs756325	0.07	T	C	110478064	T	C	0.19
1	rs504316	0.10	A	G	109663420	G	A	0.07
1	rs839551	0.11	T	C	109474581	T	C	0.05
1	rs12131828	0.12	A	G	110195152	G	A	0.05
1	rs518076	0.13	A	G	110109039	A	G	0.11
1	rs7528419	0.40	A	G	109817192	A	G	0.21

Chr., chromosome; RA, reference allele from PredictDB; EA, effect allele from PredictDB; A1, effect allele from DCCT; A2, reference allele from DCCT; MAF, minor allele frequency

Table A.12: GTEx pancreas $CELSR2$ derived using elastic net.

Chr.	SNP	Weight	RA	EA	Position	A1	A2	MAF
1	rs12740374	-0.11	T	G	109817590	G	T	0.21
1	rs7528419	0.11	A	G	109817192	A	G	0.21

Chr., chromosome; RA, reference allele from PredictDB; EA, effect allele from PredictDB; A1, effect allele from DCCT; A2, reference allele from DCCT; MAF, minor allele frequency

Table A.13: DGN whole blood $SLC12A3$ derived using elastic net.

Chr.	SNP	Weight	RA	EA	position	A1	A2	MAF
16	rs247616	-0.06	C	T	56989590	C	T	0.31
16	rs7198642	-0.06	T	G	57032461	T	G	0.21
16	rs4329913	-0.06	T	C	56905432	T	C	0.76
16	rs8060037	-0.05	C	T	56484820	C	T	0.25
16	rs17372800	-0.05	C	T	57185761	C	T	0.09
16	rs12448377	-0.05	G	T	57082366	G	T	0.13
16	rs976977	-0.05	A	G	57173674	A	G	0.92
16	rs2923131	-0.04	C	T	57729576	C	T	0.20
16	rs247617	-0.04	C	A	56990716	C	A	0.31
16	rs289748	-0.03	A	G	57025063	A	G	0.52
16	rs4784727	-0.03	T	C	56787107	T	C	0.52
16	rs1466293	-0.03	T	C	57172185	T	C	0.92
16	rs4783999	-0.03	C	T	57651985	C	T	0.48
16	rs4784650	-0.02	A	G	56306640	A	G	0.35
16	rs17282194	-0.02	C	T	56365355	C	T	0.37
16	rs4784842	-0.02	T	C	57708483	T	C	0.51
16	rs1561140	-0.02	C	T	56864398	C	T	0.52
16	rs889558	-0.02	T	C	57172629	T	C	0.92
16	rs13339005	-0.02	G	A	55934159	G	A	0.16
16	rs2587881	-0.01	G	A	56324697	G	A	0.53
16	rs9989419	-0.01	A	G	56985139	A	G	0.59
16	rs2399622	-0.01	A	G	57173080	A	G	0.92
16	rs3751710	-0.01	C	T	57095775	C	T	0.16
16	rs7184439	-0.01	C	T	56867804	C	T	0.82
16	rs9889080	-0.01	G	A	55926873	G	A	0.16
16	rs4784651	-0.01	G	A	56311774	G	A	0.35
16	rs11643815	-0.01	G	A	56602798	G	A	0.14
16	rs7188495	-0.01	G	A	57175105	G	A	0.92
16	rs247040	-0.00	C	T	57877310	C	T	0.76
16	rs1561141	-0.00	T	C	56869430	T	C	0.82
16	rs289703	0.00	C	T	57048118	C	T	0.30
16	rs8058898	0.00	T	C	57631371	T	C	0.16
16	rs154044	0.00	C	T	57114982	C	T	0.58
16	rs289717	0.00	G	A	57009388	G	A	0.35
16	rs7184983	0.00	G	A	56554709	G	A	0.12
16	rs12927110	0.01	C	T	57722866	C	T	0.23
16	rs7198661	0.01	T	C	56090428	T	C	0.47
16	rs247037	0.01	C	T	57922948	C	T	0.83
16	rs8049632	0.01	C	A	57631279	C	A	0.16
16	rs955513	0.02	C	T	56946072	C	T	0.55
16	rs12708967	0.02	T	C	56993211	T	C	0.21
16	rs43216	0.03	A	G	57116819	A	G	0.61

Chr., chromosome; RA, reference allele from PredictDB; EA, effect allele from PredictDB; A1, effect allele from DCCT; A2, reference allele from DCCT; MAF, minor allele frequency

Table A.14: DGN whole blood *CETP* derived using elastic net.

Chr.	SNP	Weight	RA	EA	position	A1	A2	MAF
16	rs7187427	-0.05	C	T	57070372	C	T	0.07
16	rs12708974	-0.05	C	T	57005550	C	T	0.10
16	rs935743	-0.04	G	A	57677885	G	A	0.15
16	rs1532624	-0.04	C	A	57005479	C	A	0.41
16	rs1532625	-0.03	C	T	57005301	C	T	0.41
16	rs506829	-0.02	C	T	57383759	C	T	0.77
16	rs4404068	-0.02	C	T	56733941	C	T	0.09
16	rs13330423	-0.02	T	C	57051501	T	C	0.56
16	rs1273580	-0.02	A	G	57352124	A	G	0.60
16	rs223869	-0.02	C	A	57494992	C	A	0.07
16	rs12599065	-0.01	T	C	56896036	T	C	0.41
16	rs948705	-0.01	G	T	57349346	G	T	0.86
16	rs11508026	-0.01	C	T	56999328	C	T	0.40
16	rs282225	-0.01	A	G	56583042	A	G	0.92
16	rs4784750	-0.01	G	T	57056064	G	T	0.29
16	rs739813	-0.01	C	T	57513076	C	T	0.09
16	rs11643127	-0.00	C	T	57919034	C	T	0.69
16	rs821469	0.00	T	G	57085547	T	G	0.05
16	rs422804	0.00	A	G	57910233	A	G	0.10
16	rs486356	0.00	G	T	57907095	G	T	0.10
16	rs2303779	0.02	G	A	57973989	G	A	0.42
16	rs17373793	0.02	T	C	57276202	T	C	0.07
16	rs1684575	0.02	G	T	57057619	G	T	0.57
16	rs289754	0.03	C	T	57065556	C	T	0.32
16	rs4784745	0.03	A	G	57014875	A	G	0.34
16	rs1894947	0.04	T	C	56006578	T	C	0.82
16	rs291040	0.04	T	C	57061189	T	C	0.34
16	rs12597002	0.05	C	A	57002404	C	A	0.30
16	rs4783968	0.06	G	A	57072004	G	A	0.69
16	rs12926631	0.07	G	A	57321362	G	A	0.06

Chr., chromosome; RA, reference allele from PredictDB; EA, effect allele from PredictDB; A1, effect allele from DCCT; A2, reference allele from DCCT; MAF, minor allele frequency

Table A.15: DGN whole blood $HERPUD1$ derived using elastic net.

Chr.	SNP	Weight	RA	EA	position	A1	A2	MAF
16	rs12920659	-0.12	C	T	56895873	C	T	0.81
16	rs1561140	-0.07	C	T	56864398	C	T	0.52
16	rs8044753	-0.05	G	A	56883438	G	A	0.52
16	rs4784724	-0.05	T	G	56745758	T	G	0.23
16	rs4784727	-0.04	T	C	56787107	T	C	0.52
16	rs9989419	-0.04	A	G	56985139	A	G	0.59
16	rs12599065	-0.03	T	C	56896036	T	C	0.41
16	rs28438857	-0.02	T	C	57060353	T	C	0.10
16	rs3829502	-0.01	G	A	56896730	G	A	0.40
16	rs247615	0.01	A	G	56984763	A	G	0.24
16	rs4353475	0.02	A	G	57326269	A	G	0.29
16	rs12447924	0.03	C	T	56994192	C	T	0.77
16	rs2217332	0.03	G	A	56969148	G	A	0.15
16	rs9921780	0.06	A	G	56952098	A	G	0.44
16	rs952439	0.07	A	C	56975277	A	C	0.15

Chr., chromosome; RA, reference allele from PredictDB; EA, effect allele from PredictDB; A1, effect allele from DCCT; A2, reference allele from DCCT; MAF, minor allele frequency

Table A.16: GTEx liver $NLRC5$ derived using elastic net.

Chr.	SNP	Weight	RA	EA	position	A1	A2	MAF
16	rs7205824	-0.23	T	C	56240331	C	T	0.05
16	rs17546208	-0.14	A	G	57539511	G	A	0.06
16	rs7198642	-0.11	T	G	57032461	T	G	0.21
16	rs2923147	-0.01	T	C	57762401	T	C	0.19
16	rs11076175	0.01	A	G	57006378	A	G	0.19
16	rs7203984	0.02	A	C	56999258	A	C	0.21
16	rs16956194	0.03	A	G	56273532	G	A	0.16
16	rs16956168	0.03	A	G	56249774	G	A	0.16
16	rs2399594	0.05	A	G	56946197	A	G	0.40
16	rs1800775	0.10	A	C	56995236	C	A	0.47
16	rs12720898	0.12	T	C	57011243	C	T	0.07

Chr., chromosome; RA, reference allele from PredictDB; EA, effect allele from PredictDB; A1, effect allele from DCCT; A2, reference allele from DCCT; MAF, minor allele frequency

Appendix B

Scripts

B.1 SNP Extraction

Script 1

```
#$ -S /bin/bash
#! /bin/bash
#export HOME=/hpf/largeprojects/andrew/adp/francis/GTeX_pan

let a=1 b=22
while [ $a -le $b ]
do

qsub  -o ~/queue -e ~/queue -v chr=$a anal2_GTeX_pan.sh
let a=$a+1
done
```

Script 2

```
#!/bin/bash
#PBS -l vmem=8g
#PBS -l nodes=1:ppn=1

cd $PBS_O_WORKDIR

chr=$chr

fgrep -w -f snps_GTeX_pan.txt /hpf/largeprojects/andrew/
hswong/dcct_1000genome_imputation/
dcct_imputation_result_folder/out/dcct_1000genome_impute
```

```
_chr${chr}_*.out > extracted${chr}.out

module load R/3.3.0

/hpf/tools/centos6/R/3.3.0/bin/Rscript impute_to_dosage_jf.r
--args ${chr} "/hpf/largeprojects/andrew/hswong/dcct_1000genome_
imputation/dcct_imputation_result_folder/
out/dcct_1000genome_impute_chr1_1.out_samples"
```

B.2 Impute to Dosages

```
# Created 13 Nov 2014
## last edited 24 June 2015
## created by: Mohsen Hosseini
########################################
## adapted for thesis July 27, 2016

### this script takes a .sample file and a .out file from
impute output and
###transforms it into a file with dosages of a1 (2nd allele)

Args <- commandArgs(TRUE)
#out.file <- Args[2]
chr <- Args[2]
sample.file <- Args[3]
out.file <- paste("extracted",chr,".out",sep="")
### reading impute file

dose0 <- read.table(out.file,header=F,comment.char="",
stringsAsFactors=F, sep=" ")
dose <- subset(dose0,!V5 %in% c("-"))

### reading sample file
samp <- read.table(sample.file, header=T, comment.char="",
stringsAsFactors=F)
samp <- samp[-1,]
samp <- subset(samp, select=c(1,2))

ssize<-nrow(samp)
n<-nrow(dose)

gt.mx<-matrix(nrow=n,ncol=ssize)
```

```
### calculating additive dosage (0 to 2) from
posterior probabilities
### calculates dosage for the 2nd allele a1 (vs a0)
for(i in 1:ssize)
{
j <- 7+(i-1)*3
### a0 is the effect allele
gt.mx[,i] <- 0*dose[,j-1]+1*dose[,j]+2*dose[,j+1]
}

### assigning missing to the SNPs with three
possibilites EQ 0
for (x in 1:n)
{
for (y in 1:ssize)
{
z=6+(y-1)*3
if (dose[x,z]==0 & dose[x,z+1]==0 & dose[x,z+2]==0)
{gt.mx[x,y] <- NA}
}
}

gtmx <- as.data.frame(gt.mx)
names(gtmx) <- as.character(samp[,2])

output <- data.frame(chromosome=chr, rsid = dose$V2,
 position = dose$V3, allele1 = dose$V4, allele2 = dose$V5,
MAF =  rowMeans(gtmx, na.rm=TRUE)/2)

output2 <- data.frame(samp)

write.table(output, paste("chr", chr,".dosage.txt",sep=""),
 quote=F, sep="\t", col.names=F, row.names=F)

write.table(output2, "samples.txt", quote=F, sep="\t",
col.names=F, row.names=F)
```

B.3 Estimate GReX

GTEx pancreas Example

```
#!/bin/bash
#PBS -l vmem=8g
#PBS -l nodes=1:ppn=1

cd $PBS_O_WORKDIR

module load PrediXcan/1.0

/hpf/tools/centos6/PrediXcan/1.0/bin/PrediXcan.py --predict
--dosages /hpf/largeprojects/andrew/adp/francis/GTeX_pan
--dosages_prefix chr --samples samples.txt
--weights GTeX_pan_predb72.db --output_dir output
```

B.4 Lipid Association with GReX

LDL Example using PrediXcan.py

```
#!/bin/bash
#PBS -l vmem=8g
#PBS -l nodes=1:ppn=1

cd $PBS_O_WORKDIR

module load PrediXcan/1.0

module load R/3.3.0

/hpf/tools/centos6/PrediXcan/1.0/bin/PrediXcan.py --assoc
--pheno win_mean_ldl --pred_exp output/predicted_expression.txt
--linear --output_dir output/ldl_win_mean
```

Bibliography

1000 Genomes Project Consortium et al. (2010) A map of human genome variation from population scale sequencing. *Nature* **467**(7319), 1061–1073.

Albert, F. W. and Kruglyak, L. (2015) The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics* **16**(4), 197–212.

Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P. and Zondervan, K. T. (2010) Data quality control in genetic case-control association studies. *Nature Protocols* **5**(9), 1564.

Aulchenko, Y. S., Ripke, S., Isaacs, A. and Van Duijn, C. M. (2007) Genabel: An r library for genome-wide association analysis. *Bioinformatics* **23**(10), 1294–1296.

Balding, D. J. (2006) A tutorial on statistical methods for population association studies. *Nature Reviews Genetics* **7**(10), 781–791.

Battle, A., Mostafavi, S., Zhu, X., Potash, J. B., Weissman, M. M., McCormick, C., Haudenschild, C. D., Beckman, K. B., Shi, J., Mei, R. et al. (2014) Characterizing the genetic basis of transcriptome diversity through rna-sequencing of 922 individuals. *Genome Research* **24**(1), 14–24.

Below, J. E., Parra, E. J., Gamazon, E. R., Torres, J., Krithika, S., Candille, S., Lu, Y., Manichakul, A., Peralta-Romero, J., Duan, Q. et al. (2016) Meta-analysis of lipid-traits in hispanics identifies novel loci, population-specific effects, and tissue-specific enrichment of eqtls. *Scientific Reports* **6**.

Benzinou, M., Walley, A., Lobbens, S., Charles, M.-A., Jouret, B., Fumeron, F., Balkau, B., Meyre, D. and Froguel, P. (2006) Bardet-biedl syndrome gene variants are associated with both childhood and adult common obesity in french caucasians. *Diabetes* **55**(10), 2876–2882.

Boudina, S. and Abel, E. D. (2007) Diabetic cardiomyopathy revisited. *Circulation* **115**(25), 3213–3223.

Breitling, C., Gross, A., Büttner, P., Weise, S., Schleinitz, D., Kiess, W., Scholz, M., Kovacs, P. and Körner, A. (2015) Genetic contribution of variants near sort1

and apoe on ldl cholesterol independent of obesity in children. *PloS one* **10**(9), e0138064.

Brown, M. L., Inazu, A., Hesler, C. B., Agellon, L. B., Mann, C., Whitlock, M. E., Marcel, Y. L., Milne, R. W., Koizumi, J., Mabuchi, H. et al. (1989) Molecular basis of lipid transfer protein deficiency in a family with increased high-density lipoproteins. *Nature* **342**(6248), 448–451.

Browning, B. L. and Browning, S. R. (2007) Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. *Genetic Epidemiology* **31**(5), 365–375.

Carlson, C. S., Eberle, M. A., Rieder, M. J., Smith, J. D., Kruglyak, L. and Nickerson, D. A. (2003) Additional snps and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nature Genetics* **33**(4), 518–521.

Chatterjee, S. and Hadi, A. S. (2006) *Regression analysis by example*. Fourth edition. Hoboken, New Jersey: John Wiley Sons, Inc.

Chen, H. and Boutros, P. C. (2011) Venndiagram: A package for the generation of highly-customizable venn and euler diagrams in r. *BMC Bioinformatics* **35**(1), 35.

Dastani, Z., Pajukanta, P., Marcil, M., Rudzicz, N., Ruel, I., Bailey, S. D., Lee, J. C., Lemire, M., Faith, J., Platko, J. et al. (2010) Fine mapping and association studies of a high-density lipoprotein cholesterol linkage region on chromosome 16 in french-canadian subjects. *European Journal of Human Genetics* **18**(3), 342–347.

Davis, C., Williams, D., Oganov, R., Tao, S.-C., Rywik, S., Stein, Y. and Little, J. (1996) Sex difference in high density lipoprotein cholesterol in six countries. *American Journal of Epidemiology* **143**(11), 1100–1106.

DCCT Research Group et al. (1986) The diabetes control and complications trial (dcct): design and methodologic considerations for the feasibility phase. *Diabetes* **35**(5), 530–545.

DCCT Research Group et al. (1993) The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *N Engl J Med* **329**(14), 977–986.

Despres, J.-P., Lemieux, I., Dagenais, G.-R., Cantin, B. and Lamarche, B. (2000) Hdl-cholesterol as a marker of coronary heart disease risk: the quebec cardiovascular study. *Atherosclerosis* **153**(2), 263–272.

Devlin, B. and Roeder, K. (1999) Genomic control for association studies. *Biometrics* **55**(4), 997–1004.

Edwards, A. O., Ritter, R., Abel, K. J., Manning, A., Panhuysen, C. and Farrer, L. A. (2005) Complement factor h polymorphism and age-related macular degeneration. *Science* **308**(5720), 421–424.

Eura, Y., Yanamoto, H., Arai, Y., Okuda, T., Miyata, T. and Kokame, K. (2012) Derlin-1 deficiency is embryonic lethal, derlin-3 deficiency appears normal, and herp deficiency is intolerant to glucose load and ischemia in mice. *PloS One* **7**(3), e34298.

Ezkurdia, I., Juan, D., Rodriguez, J. M., Frankish, A., Diekhans, M., Harrow, J., Vazquez, J., Valencia, A. and Tress, M. L. (2014) Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Human Molecular Genetics* **23**(22), 5866–5878.

Folkersen, L., van't Hooft, F., Chernogubova, E., Agardh, H. E., Hansson, G. K., Hedin, U., Liska, J., Syvanen, A.-C., Paulssson-Berne, G., Franco-Cereceda, A. et al. (2010) Association of genetic risk variants with expression of proximal genes identifies novel susceptibility genes for cardiovascular disease. *Circulation: Cardiovascular Genetics* pp. CIRCGENETICS–110.

Friedewald, W. T., Levy, R. I. and Fredrickson, D. S. (1972) Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clinical Chemistry* **18**(6), 499–502.

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., Eyler, A. E., Denny, J. C., Nicolae, D. L., Cox, N. J. et al. (2015) A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics* **47**(9), 1091–1098.

Gilad, Y., Rifkin, S. A. and Pritchard, J. K. (2008) Revealing the architecture of gene regulation: the promise of eqtl studies. *Trends in Genetics* **24**(8), 408–415.

Glaudemans, B., Yntema, H. G., San-Cristobal, P., Schoots, J., Pfundt, R., Kamsteeg, E.-J., Bindels, R. J., Knoers, N. V., Hoenderop, J. G. and Hoefsloot, L. H. (2012) Novel ncc mutants and functional analysis in a new cohort of patients with gitelman syndrome. *European Journal of Human Genetic* **20**(3), 263–270.

GTEx Consortium et al. (2015) The genotype-tissue expression (gtex) pilot analysis: Multitissue gene regulation in humans. *Science* **348**(6235), 648–660.

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chong, W., Penninx, B. W., Jansen, R., de Geus, E. J., Boomsma, D. I., Wright, F. A. et al. (2016) Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics* **48**(3), 245–255.

Guy, J., Ogden, L., Wadwa, R. P., Hamman, R. F., Mayer-Davis, E. J., Liese, A. D., D'agostino, R., Marcovina, S. and Dabelea, D. (2009) Lipid and lipoprotein profiles in youth with and without type 1 diabetes. *Diabetes Care* **32**(3), 416–420.

Howie, B. N., Donnelly, P. and Marchini, J. (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**(6), e1000529.

Inazu, A., Jiang, X.-C., Haraki, T., Yagi, K., Kamon, N., Koizumi, J., Mabuchi, H., Takeda, R., Takata, K., Moriyama, Y. et al. (1994) Genetic cholesteryl ester transfer protein deficiency caused by two prevalent mutations as a major determinant of increased levels of high density lipoprotein cholesterol. *Journal of Clinical Investigation* **94**(5), 1872.

International HapMap 3 Consortium et al. (2010) Integrating common and rare genetic variation in diverse human populations. *Nature* **467**(7311), 52.

Iribarren, C., Karter, A. J., Go, A. S., Ferrara, A., Liu, J. Y., Sidney, S. and Selby, J. V. (2001) Glycemic control and heart failure among adult patients with diabetes. *Circulation* **103**(22), 2668–2673.

van Iterson, M., van Zwet, E. W. and Heijmans, B. T. (2017) Controlling bias and inflation in epigenome-and transcriptome-wide association studies using the empirical null distribution. *Genome Biology* **18**(1), 19.

Johnson, R. C., Nelson, G. W., Troyer, J. L., Lautenberger, J. A., Kessing, B. D., Winkler, C. A. and O'Brien, S. J. (2010) Accounting for multiple comparisons in a genome-wide association study (gwas). *BMC Genomics* **11**(1), 724.

Kathiresan, S., Willer, C. J., Peloso, G. M., Demissie, S., Musunuru, K., Schadt, E. E., Kaplan, L., Bennett, D., Li, Y., Tanaka, T. et al. (2009) Common variants at 30 loci contribute to polygenic dyslipidemia. *Nature Genetics* **41**(1), 56–65.

Kjolby, M., Andersen, O. M., Breiderhoff, T., Fjorback, A. W., Pedersen, K. M., Madsen, P., Jansen, P., Heeren, J., Willnow, T. E. and Nykjaer, A. (2010) Sort1, encoded by the cardiovascular risk locus 1p13. 3, is a regulator of hepatic lipoprotein export. *Cell Metabolism* **12**(3), 213–223.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T. et al. (2005) Complement factor h polymorphism in age-related macular degeneration. *Science* **308**(5720), 385–389.

Kurano, M., Tsukamoto, K., Kamitsuji, S., Kamatani, N., Hara, M., Ishikawa, T., Kim, B.-J., Moon, S., Kim, Y. J. and Teramoto, T. (2016) Genome-wide association study of serum lipids confirms previously reported associations as well as new

associations of common snps within pcsk7 gene with triglyceride. *Journal of Human Genetics* .

Lappalainen, T., Sammeth, M., Friedländer, M. R., AC't Hoen, P., Monlong, J., Rivas, M. A., Gonzalez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G. et al. (2013) Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**(7468), 506–511.

Lettre, G., Lange, C. and Hirschhorn, J. N. (2007) Genetic model testing and statistical power in population-based association studies of quantitative traits. *Genetic Epidemiology* **31**(4), 358–362.

Li, M., Li, C. and Guan, W. (2008) Evaluation of coverage variation of snp chips for genome-wide association studies. *European Journal of Human Genetics* **16**(5), 635–643.

Li, Y., Willer, C., Sanna, S. and Abecasis, G. (2009) Genotype imputation. *Annual Review of Genomics and Human Genetics* **10**, 387.

Linsel-Nitschke, P., Heeren, J., Aherrahrou, Z., Bruse, P., Gieger, C., Illig, T., Prokisch, H., Heim, K., Doering, A., Peters, A. et al. (2010) Genetic variation at chromosome 1p13. 3 affects sortilin mrna expression, cellular ldl-uptake and serum ldl levels which translates to the risk of coronary artery disease. *Atherosclerosis* **208**(1), 183–189.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N. et al. (2013) The genotype-tissue expression (gtex) project. *Nature Genetics* **45**(6), 580–585.

Ma, L., Yang, J., Runesha, H. B., Tanaka, T., Ferrucci, L., Bandinelli, S. and Da, Y. (2010) Genome-wide association analysis of total cholesterol and high-density lipoprotein cholesterol levels using the framingham heart study data. *BMC Medical Genetics* **11**(1), 55.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J. et al. (2017) The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). *Nucleic Acids Research* **45**(D1), D896–D901.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A. et al. (2009) Finding the missing heritability of complex diseases. *Nature* **461**(7265), 747–753.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. and Hirschhorn, J. N. (2008) Genome-wide association studies for complex

traits: consensus, uncertainty and challenges. *Nature Reviews Genetics* **9**(5), 356–369.

Monir, M. M. and Zhu, J. (2017) Comparing gwas results of complex traits using full genetic model and additive models for revealing genetic architecture. *Scientific Reports* **77**.

Mooney, M. A. and Wilmot, B. (2015) Gene set analysis: A step-by-step guide. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* **168**(7), 517–527.

Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N. E., Ahfeldt, T., Sachs, K. V., Li, X., Li, H., Kuperwasser, N., Ruda, V. M. et al. (2010) From noncoding variant to phenotype via sort1 at the 1p13 cholesterol locus. *Nature* **466**(7307), 714–719.

Nagy, R., Boutin, T. S., Marten, J., Huffman, J. E., Kerr, S. M., Campbell, A., Evenden, L., Gibson, J., Amador, C., Howard, D. M. et al. (2017) Exploration of haplotype research consortium imputation for genome-wide association studies in 20,032 generation scotland participants. *Genome Medicine* **9**(1), 23.

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E. and Cox, N. J. (2010) Trait-associated snps are more likely to be eqtls: annotation to enhance discovery from gwas. *PLoS Genetics* **6**(4), e1000888.

Paterson, A. D., Waggott, D., Boright, A. P., Hosseini, S. M., Shen, E., Sylvestre, M.-P., Wong, I., Bharaj, B., Cleary, P. A., Lachin, J. M. et al. (2010) A genome-wide association study identifies a novel major locus for glycemic control in type 1 diabetes, as measured by both a1c and glucose. *Diabetes* **59**(2), 539–549.

Postmus, I., Trompet, S., Deshmukh, H. A., Barnes, M. R., Li, X., Warren, H. R., Chasman, D. I., Zhou, K., Arsenault, B. J., Donnelly, L. A. et al. (2014) Pharmacogenetic meta-analysis of genome-wide association studies of ldl cholesterol response to statins. *Nature Communications* **5**.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. and Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**(8), 904.

Prospective Studies Collaboration et al. (2007) Blood cholesterol and vascular mortality by age, sex, and blood pressure: a meta-analysis of individual data from 61 prospective studies with 55 000 vascular deaths. *The Lancet* **370**(9602), 1829–1839.

Rahmouni, K., Fath, M. A., Seo, S., Thedens, D. R., Berry, C. J., Weiss, R., Nishimura, D. Y. and Sheffield, V. C. (2008) Leptin resistance contributes to obesity and hypertension in mouse models of bardet-biedl syndrome. *The Journal of clinical investigation* **118**(4), 1458–1467.

Rebhan, M., Chalifa-Caspi, V., Prilusky, J. and Lancet, D. (1997) Genecards: integrating information about genes, proteins and diseases. *Trends in Genetics* **13**(4), 163.

Ridker, P. M., Paré, G., Parker, A. N., Zee, R. Y., Miletich, J. P. and Chasman, D. I. (2009) Polymorphism in the cetp gene region, hdl cholesterol, and risk of future myocardial infarction. *Circulation: Cardiovascular Genetics* **2**(1), 26–33.

Risch, N., Merikangas, K. et al. (1996) The future of genetic studies of complex human diseases. *Science* **273**(5281), 1516–1517.

Ritchie, R. H., Zerenturk, E. J., Prakoso, D. and Calkin, A. C. (2017) Lipid metabolism and its implications for type 1 diabetes-associated cardiomyopathy. *Journal of Molecular Endocrinology* **58**(4), R225–R240.

Schadt, E. E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P. Y., Kasarskis, A., Zhang, B., Wang, S., Suver, C. et al. (2008) Mapping the genetic architecture of gene expression in human liver. *PLoS Biol* **6**(5), e107.

Schierding, W. and O'Sullivan, J. M. (2015) Connecting snps in diabetes: a spatial analysis of meta-gwas loci. *Frontiers in Endocrinology* **6**.

Schunkert, H., König, I. R., Kathiresan, S., Reilly, M. P., Assimes, T. L., Holm, H., Preuss, M., Stewart, A. F., Barbalic, M., Gieger, C. et al. (2011) Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature Genetics* **43**(4), 333–338.

Shaffer, J. P. (1995) Multiple hypothesis testing. *Annual Review of Psychology* **46**(1), 561–584.

Shirali, M., Pong-Wong, R., Navarro, P., Knott, S., Hayward, C., Vitart, V., Rudan, I., Campbell, H., Hastie, N. D., Wright, A. F. et al. (2016) Regional heritability mapping method helps explain missing heritability of blood lipid traits in isolated populations. *Heredity* **116**(333–338).

Siebel, A. L., Heywood, S. E. and Kingwell, B. A. (2015) Hdl and glucose metabolism: current evidence and therapeutic potential. *Frontiers in Pharmacology* **6**, 258.

Spencer, C. C., Su, Z., Donnelly, P. and Marchini, J. (2009) Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet* **5**(5), e1000477.

Surakka, I., Horikoshi, M., Mägi, R., Sarin, A.-P., Mahajan, A., Lagou, V., Marullo, L., Ferreira, T., Miraglio, B., Timonen, S. et al. (2015) The impact of low-frequency and rare variants on lipid levels. *Nature Genetics* **47**(6), 589–597.

Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., Pirruccello, J. P., Ripatti, S., Chasman, D. I., Willer, C. J. et al. (2010) Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**(7307), 707–713.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 267–288.

Tregout, D.-A., Konig, I. R., Erdmann, J., Munteanu, A., Braund, P. S., Hall, A. S., Grosshennig, A., Linsel-Nitschke, P., Perret, C., DeSuremain, M. et al. (2009) Genome-wide haplotype association study identifies the slc22a3-lpal2-lpa gene cluster as a risk locus for coronary artery disease. *Nature Genetics* **41**(3), 283–285.

Turner, S. D. (2014) qqman: an r package for visualizing gwas results using qq and manhattan plots. *BioRxiv* p. 005165.

Vergs, B. (2009) Lipid disorders in type 1 diabetes. *Diabetes and Metabolism* **35**(5), 353–360.

Wallace, C., Newhouse, S. J., Braund, P., Zhang, F., Tobin, M., Falchi, M., Ahmadi, K., Dobson, R. J., Marçano, A. C. B., Hajat, C. et al. (2008) Genome-wide association study identifies genes for biomarkers of cardiovascular disease: serum urate and dyslipidemia. *The American Journal of Human genetics* **82**(1), 139–149.

Willer, C. J. et al. (2013) Discovery and refinement of loci associated with lipid levels. *Nature Genetics* **45**(11), 1274–1283.

Willer, C. J., Sanna, S., Jackson, A. U., Scuteri, A., Bonnycastle, L. L., Clarke, R., Heath, S. C., Timpson, N. J., Najjar, S. S., Stringham, H. M. et al. (2008) Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nature Genetics* **40**(2), 161–169.

Wittkopp, P. J. and Kalay, G. (2012) Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics* **13**(1), 59.

Wray, N. R., Yang, J., Hayes, B. J., Price, A. L., Goddard, M. E. and Visscher, P. M. (2013) Pitfalls of predicting complex traits from snps. *Nature Reviews Genetics* **14**(7), 507–515.

Zeng, P., Zhao, Y., Qian, C., Zhang, L., Zhang, R., Gou, J., Liu, J., Liu, L. and Chen, F. (2015) Statistical analysis for genome-wide association study. *Journal of Biomedical Research* **29**(4), 285–297.

Zhang, Q. S., Browning, B. L. and Browning, S. R. (2015) Genome-wide haplotypic testing in a finnish cohort identifies a novel association with low-density lipoprotein cholesterol. *European Journal of Human Genetics* **23**(5), 672.

Zheng, G., Freidlin, B. and Gastwirth, J. L. (2006) Robust genomic control for association studies. *The American Journal of Human Genetics* **78**(2), 350–356.

Zou, H. and Hastie, T. (2005) *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(2), 301–320.