Association of imputed gene expression with glycated haemoglobin (HbA1c) levels in people with type 1 diabetes

by

Sara V. Good

A Thesis submitted to the Faculty of Graduate Studies of
The University of Manitoba
in partial fulfilment of the requirements of the degree of
MASTER OF SCIENCE

Department of Community Health Science
University of Manitoba
Winnipeg

Abstract

Background: Type 1 diabetes (T1D) is a common complex metabolic disease characterized by the autoimmune destruction of insulin-producing β cells of the pancreatic islets. Due to a drastic loss of β cells, resulting in no to low insulin production, persons with T1D require life-long insulin injections. Genetic factors play a significant role in the aetiology of T1D, and the heritability of T1D is estimated to be between 85-88%. Many of the known susceptibility genes are involved in immune processes, which is consistent with its aetiology as an autoimmune disorder in which auto-antigens attack the pancreatic islets cells. However, the main causes of morbidity and mortality today are the complications that arise from even mild hyperglycaemia in T1D cases. Glycemic control is often measured using an assay called haemoglobin A1c (HbA1c), that provides an estimate of average blood glucose in individuals over an ~ 3month period. **Objectives:** In this thesis, I used phenotype and genetic data from the Diabetes Complications and Control Trial (DCCT), a large (n=1441) randomized clinical trial designed to assess the impact of two different (conventional and intensive) approaches to glycemic control, to examine the relationship between *imputed* gene expression and mean HbA1c. **Hypothesis**: I hypothesis that variation in levels of imputed gene expression in blood and/or other tissues will be associated variation in glycemic control. Secondarily, I hypothesis that this association may differ between treatment arms, because of the effect of treatment on HbA1c. **Methods:** Gene expression or transcriptome imputation is a relatively new method in genetic epidemiology in which training sets from different tissues are used to identify the set of single nucleotide polymorphisms (SNPs) within a given window (typically 2Mb) surrounding a gene that best predict RNA abundance in that tissue using regression regularization or a similar statistical approach. Then, the set of SNPs that predict gene expression in each tissue can be used to impute the transcriptome in an independent genotyped dataset. Here, I use the open-source PredictDB data repository, the DCCT genotype data set and the PrediXcan workflow described by Gamazon *et al* to test for associations between imputed gene expression and mean HbA1c in ten different training sets (tissues). **Results:** I find weakly suggestive associations between imputed gene expression and mean HbA1c for nine genes, and I find a treatment*gene interaction for two of these genes. Three of the associated genes mapped to chromosomal regions with known genome wide association study (GWAS) hits, and one of the genes (*MAPKI81*) has been associated with glycemic control in individuals with Type 2 diabetes. **Conclusions and future work:** Collectively, this suggests that employing transcriptome imputation may help identify non-coding variants that influence (disease) phenotypes by dysregulating gene expression. In future work, this workflow will be applied to multiple datasets and then a meta-analysis will be performed, in the hopes that we wil be powered to find statistically significant associations between HbA1c and predicted gene expression. The ultimate goal of this research is to tease apart the genetic determinants of hyperglycemia in T1D and discover new therapeutic approaches to controlling blood sugar.

Acknowledgements

Undertaking this degree in Community Health Sciences has been rewarding and challenging since I completed the degree while also a faculty member. I would like to acknowledge the excellence of the courses offered in CHS at UofM; I think that the program prepared me for work in public health and epidemiology.

Next, I would like to acknowledge the wonderful support and mentorship I received from my two co-supervisors Marissa Becker and Xiaoqing (Michelle) Liu. Marissa was very understanding of my (multiple) changing thesis plans, and used grace and humour to try to keep me on track of the various deadlines and requirements of the degree, something that ironically gets more challenging once one is already a professor. When I started the program in CHS, I had hoped to complete the thesis portion of the program in genetic epidemiology, but was unaware of any faculty member working in the field at UofM. Serendipitously, I discovered that Michelle (Xiaoqing) Liu was a newly hired genetic epidemiologist, and I started to work with her on Autism-related project in 2016. Finding her to work with prolonged my degree (because I had to switch projects) but made it the experience I had hoped for. Although my thesis was ultimately written based on work I performed at Sick Kids Hospital under the direction of Andrew Paterson, I hope to maintain my collaborations with Michelle on ASD and any other projects in genetic epidemiology; it is really wonderful to work with her.

Also serendipitously, a student at UofW told me about the CIHR funded STAGE program (Strategic Training in Advanced Genetic Epidemiology) hosted by the Dala Lana school of Public Health at the University of Toronto. After reviewing the possible mentors to apply to this program, the work of Andrew D. Paterson stood out to me as the most interesting because it focused on a complex disease (Type 1 diabetes) and studied it from equally strong foundations in genetics and statistics, my two favourite areas. As a result, I undertook a one-year position as a research fellow at SickKids working with Dr. Paterson's group; this thesis stems from that work. It has been beyond fantastic to work with Andrew and to learn so much about human genomics from him and the other people at SickKids and the StatGen journal club. Truly, working with him and Dr. Liu has been the highlight of my professional life in the last few years, so a big thank you to them.

I would also like to thank my two committee members Lisa Lix and Louise Simard, who were wonderful to meet and inspiring, my two stage mentors Michael Wilson and Angelo Canty, who were equally inspiring and encouraging, and my awesome "lab mates" at Sick Kid: Delnaz Roshandel, Sareh Keshavari, Jingjing Cao, Joanna Francis for sharing her code; a special note of appreciation to Shelley Bull at the Lunenfeld-Tanenbaum Research Institute for being an important role model for me while in Toronto (and many other women I believe). Heartfelt thanks to my MSc students Brett Vahkal, Nisha Ajmani and Hend for being so supportive of their supervisor while she was a student, and to my sons Julian Joseph and Manuel Isaac for constant support, inquisitiveness and love.

Lastly, special thanks to my partner Sergey Yegorov and our newly born son who were and continue to be a constant source of inspiration for me, forever encouraging me to try new things; I hope that completing this degree will open new avenues for us to explore together. This thesis is dedicated to our son Alyosha who was such a good baby that I managed to finish the thesis during the first three months of his life.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

BSLMM	Bayesian Sparse Linear Mixed Model
CONV	Conventional Therapy
CSII	continuous subcutaneous insulin infusion
CVD	Cardiovascular Disease
DCCT	Diabetes Control and Complications Trial
DGN	Depression Genes and Networks
DHA	docosahexaenoic acid
DHS	Dnase Hypersensitivity Site
DNA	deoxyribonucleic acid
EDIC	Epidemiology of Diabetes Interventions and Complications
ENCODE	The Encyclopedia of DNA Elements
FG	Fasting Glucose
FP	Fasting ProInsulin
FWER	Family Wide Error Rate
GAD	Glutamic acid decarboxylase
GReXg	Genotically Controlled Component of Gene Expression
GTEx	Genotype-Tissue Expression
GWAS	Genome Wide Association Study
HBA	haemoglobin A
HbA1c	Haemoglobin A1c
HBB	haemoglobin B
HLA	Human Leukocyte Antigen
HMM	hidden markov modeling
HWE	Hardy Weinberg Equilibrium
IDDM	Insulin-dependent diabetes mellitus
INS	Insulin
INT	Intensive Therapy
LASSO	Least Absolute Shrinkage and selection operator
LD	Linkage Disequilibrium
LOF	loss of function
LRT	likelihood ratio test
MAF	minor allele frequency
MDI	multiple daily injections
mRAN	messenger RNA
NCBI	National Centre for Biotechnology Information
NHGRI-EBI	National Human Genome Research Institute - European Bioinformatics Institute
NIH	National Institute of Health

OLS	Ordinary Least Squares
ORF	Open Reading Frame
PCA	Principal Component Analysis
PEER	Probabilistic Estimation of Expression Residuals
pLI	probability of being Intolerant to mutation
QC	Quality Control
RNA	Ribonucleic Acid
RSS	Residual Sum of Squares
SMBG	self-monitoring of blood glucose
SNP	Single Nucleotide Polymorphism
TFBS	Transcription Factor Binding Sites
TSS	Transcription Start Site
TWAS	Transcriptome Wide Association Study
UTR	Untranslated Region
WHO	World Health Organisation
ZnT8	Zinc transporter 8

Chapter 1: Introduction

1.1 Overview

In this thesis, I performed a transcriptome-wide association study (TWAS) using a workflow outlined by Gamazon *et al.* (2015) to investigate the associations between imputed gene expression and mean glycated haemoglobin levels in a cohort of individuals with type 1 diabetes. To establish the background for the thesis, I first provided a literature review of the genetic and environmental risk factors for type 1 diabetes, and described the DCCT trial from which the genotype and phenotype data employed in this thesis were obtained. Then to contextualize the TWAS analyses, I provided an overview of the structure of genes, gene regulation, and the nature of genetic variants in the human genome, and followed this with an overview of the statistical and genetic methodology of expression quantitative trait locus (eQTL) study, particularly that developed by Gamazon *et al.*, (2015).

1.2 Type 1 Diabetes

Insulin-dependent diabetes mellitus (IDDM) or type 1 diabetes (T1D) is a complex metabolic disease characterized by the autoimmune destruction of insulin-producing β cells of the pancreatic islets (VAN BELLE *et al.* 2011). Due to a drastic loss of β cells resulting in zero to low insulin production, individuals with T1D require life-long insulin injections. Although, non-autoimmune factors contribute to the destruction of pancreatic β –cells in a small proportion of individuals, the prevailing view is that T1D arises due to an environmentally triggered autoimmune destruction of pancreatic β cells against a background of genetic risk, a process that usually occurs during early childhood

(MANNERING *et al.* 2016). A central role for the immune system in the aetiology of T1D is supported by the involvement of multiple immune-related factors including: the human leukocyte antigen (HLA) (*HLA-DQ2-DQ8*), Insulin (*INS*) and immune system genes (PTPN2, IL2 and others); specificity of the disease in destroying beta cells in the pancreatic islets, positive impact of immune-suppressive therapies as well as the presence of circulating autoantibodies to β-cell proteins (POCIOT AND LERNMARK 2016). It is important to note that although T1D is predominantly an autoimmune disorder strongly influenced by genetic factors, there is considerable heterogeneity among individuals in the rate of progression and severity of the disease, notably in the rate of β-cell loss, and therefore insulin production, responsiveness to immunotherapies and even islet pathology (e.g. insulitis) (ARIF *et al.* 2014; CHO AND FELDMAN 2015; LEETE *et al.* 2016).

The heritability of T1D has been estimated to be around 85-88% (REDONDO *et al.* 2001; HYTTINEN *et al.* 2003; REDONDO *et al.* 2008), meaning that up to 88% of the phenotypic liability in T1D risk is due to genetic factors, and 12% to environmental factors. The heritability of traits can be measured by the difference in concordance rates for identical (monozygotic) versus non-identical (dizygotic) twins, in which two times this difference provides an estimate of heritability. Twin studies in Finland, Great Britain and the United States have all found that concordance rates for T1D among monozygotic twins is less than 100% but the degree of concordance is significantly influenced by age at diagnosis, with the highest concordance rates among twins diagnosed with T1D prior to the age of fifteen years (REDONDO *et al.* 2001; HYTTINEN *et al.* 2003). A survival analysis of twins in Great Britain and England found that non-diabetic identical twins of probands diagnosed with T1D under the age of 25 years had a 38% probability of developing diabetes compared with only 6% for twins of probands diagnosed later (REDONDO *et al.* 2001). In the same study, for twins that eventually became concordant for T1D, the median time of discordance was 4.2 years, but for probands diagnosed after 25 years of age, 23% of the twins were discordant for more than 15 years. This indicates that the heritability of T1D decreases with age of diagnosis, and that environmental and gene x environment factors also influence risk.

The influence of environmental factors on T1D risk is supported by fact that the incidence of T1D has increased significantly in the past 30 years indicating that changes in environment or lifestyle also play a role. For example, Scandinavian countries have some of the highest rates of T1D in the world, yet immigrant families to both Sweden and Finland adopt a similar (higher) risk for acquiring T1D than children from the immigrant family's country of origin (OILINKI *et al.* 2012; SODERSTROM *et al.* 2012). Similarly, in Europe, the risk of developing T1D differs significantly among individuals who are closely related but separated by international borders that mirror socioeconomic disparities (KONDRASHOVA *et al.* 2005).

Prior to the introduction of insulin therapies in 1922, T1D was a uniformly fatal disease, but was subsequently transformed into a chronic one (NATHAN AND GROUP 2014). In the following decades, the primary causes of morbidity and mortality for individuals with T1D on insulin therapy have been chronic complications of the eyes, kidneys, peripheral and autonomic nervous systems (microvascular complications leading to retinopathy, nephropathy and neuropathy) as well as a substantially increased risk of cardiovascular disease (CVD, macrovascular complications) (JACOBSON *et al.* 2013; DIABETES *et al.* 2016). In the following sections, I will briefly summarize the genetic risk factors for both acquiring T1D and developing the long-term complications, and then discuss possible environmental triggers for developing T1D.

1.3.1 Genes associated with type 1 Diabetes

The genes associated with T1D are divided into those that contribute to the aetiological onset of the disease, characterised by β-cell autoimmunity and the development of β-cell autoantibodies, and those that contribute to the progression and clinical manifestation of the disease (i.e. the pathogenesis and complications associated with the disease) (POCIOT AND LERNMARK 2016).

Although T1D is diagnosed after the onset of overt hyperglycaemia, there is now strong evidence that autoantibodies to pancreatic proteins is the first stage of the disease (INSEL *et al.* 2015). In the early, even pre-symptomatic, stages, children typically have β-cell autoantibodies to insulin (INS), glutamic acid decarboxylase (GAD), islet antigen-2, or Zinc transporter 8 (ZnT8), such that when patients are diagnosed with T1D, autoantibodies to one or more of these four proteins is detected (ZIEGLER *et al.* 2013). Two HLA haplotypes, *HLA-DR3-DQ2* and *HLA-DR4-DQ8*, individually or together, are known to be significant risk factors for the development of T1D (LENZ *et al.* 2015; POCIOT AND LERNMARK 2016; JERRAM AND LESLIE 2017). Nearly 90% of Scandinavian children diagnosed with T1D have one or both of these HLA haplotypes (POCIOT AND LERNMARK 2016). These two HLA haplotypes are associated with the development of specific autoantibody subsets: for example, *HLA-DR3-DQ2* is associated with GAD autoantibodies

(GAD), and *HLA-DR4-DQ8* with INS autoantibodies. Apart from the HLA-haplotypes (Lenz *et al.* 2015; Jerram and Leslie 2017), 58 other regions in the human genome have been shown to have associations with T1D and the development of autoantibodies (Pociot and Lernmark 2016). Most of the associated genes in these regions are in immune system pathways, and some of them greatly increase the odds of having T1D; for example, two risk variants in the gene PTPN22 have an odds ratio of 1.89 for T1D (Pociot and Lernmark 2016).

There is less clear evidence about which genetic loci are associated with the long-term complications of T1D, probably for several reasons (Pociot and Lernmark 2016). Firstly, the micro- and macro-vascular phenotypes associated with the long-term complications of T1D are less well-defined and can develop over decades. Secondly, the sample sizes of studies examining the genetic risk factors for complications of T1D are small compared to those studying T1D, reducing statistical power. Thirdly, it is not clear whether the loci contributing to the long-term complications associated with T1D are different from those associated with type 2 diabetes (T2D) (Pociot and Lernmark 2016). Nevertheless, some loci have been found to be specifically associated with the long-term complications of T1D (Ahlqvist *et al.* 2015). The identification of genetic variants robustly associated with the long-term complications of T1D would be very beneficial for improving the clinical outcomes of patients. For example, if variants associated with susceptibility to the micro and macro-vascular complications of T1D could be identified, preventive therapies could be pursued prior to the development of complications.

1.3.2 Environmental risk factors for the development of T1D

Many factors have been proposed to be potential environmental triggers for both the development of islet autoimmunity and the progression from autoimmunity to overt T1D, but only a few have found consistent experimental support. A recent paper by Rewers *et al.* (2016) provides a comprehensive review of the subject. They divide the environmental factors impacting T1D development into those that may have adverse or protective effects on the development of islet autoimmunity during both pre-natal and post-natal periods and then those factors that are thought to accelerate or slow the progression from islet autoimmunity to clinical onset of T1D (summarized in Figure 1.1). During the pre-natal period, associations with subsequent development of islet autoantibodies have been found for maternal enteroviral infections and older maternal age while a protective effect of higher vitamin D intake or serum levels during late pregnancy or a child's infancy has found some, but inconsistent support. For example, a meta-analysis of retrospective studies found support for a relationship between higher vitamin D levels during pregnancy or early infancy and a lower risk of developing T1D (ZIPITIS AND AKOBENG 2008; MIETTINEN *et al.* 2012), but randomized clinical trials that were inspired by these studies found no support for the protective effects of Vitamin D in prospective cohort designs (BIZZARRI *et al.* 2010; WALTER *et al.* 2010).

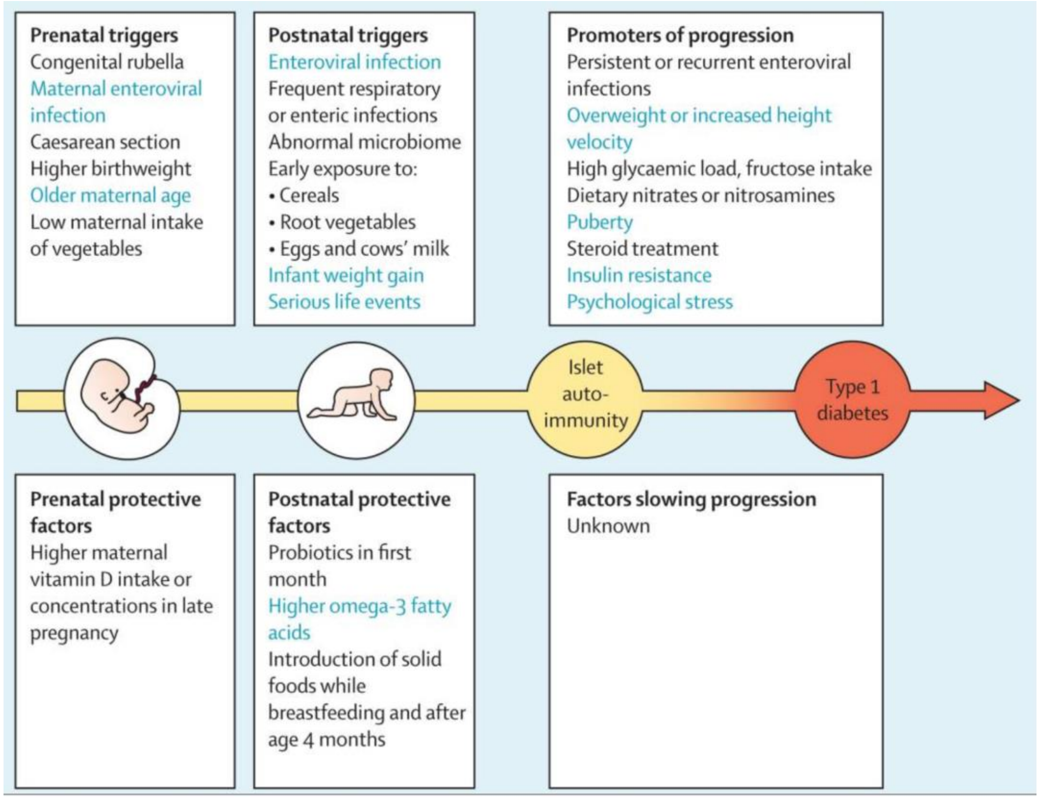

Prenatal triggers
Congenital rubella
Maternal enteroviral infection
Caesarean section
Higher birthweight
Older maternal age
Low maternal intake of vegetables
Postnatal triggers
Enteroviral infection
Frequent respiratory or enteric infections
Abnormal microbiome
Early exposure to:
• Cereals
• Root vegetables
• Eggs and cows' milk
Infant weight gain
Serious life events
Promoters of progression
Persistent or recurrent enteroviral infections
Overweight or increased height velocity
High glycaemic load, fructose intake
Dietary nitrates or nitrosamines
Puberty
Steroid treatment
Insulin resistance
Psychological stress
Islet auto-immunity
Type 1 diabetes
Prenatal protective factors
Higher maternal vitamin D intake or concentrations in late pregnancy
Postnatal protective factors
Probiotics in first month
Higher omega-3 fatty acids
Introduction of solid foods while breastfeeding and after age 4 months
Factors slowing progression
Unknown

Figure 1.1 Summary of environmental risk factors for the development of islet autoimmunity during pre-natal and post-natal periods and the progression from autoimmunity to clinical type 1 diabetes. Factors in light blue are those for which there is stronger and/or more consistent evidence (from Rewers and Ludvigsson, 2016).

During the post-natal period a variety of dietary factors have found limited support as risk or protective factors in the development of islet autoimmunity (summarized in Figure 1.1), but there is stronger support for the hypothesis that enteroviral infection or rapid weight gain in the child, or serious life events such as divorce or a death in the family that influence both mother and child are all risk factors (REWERS AND LUDVIGSSON 2016). Support for a protective effect of a diet higher in omega-3 fatty acids has been found in children that are at genetic risk for T1D (NORRIS *et al.* 2007), although research studying whether mothers that had diets rich in omega-3 fatty acids and thus have higher levels of docosahexaenoic acid (DHA), and eicosapentaenoic acid has been equivocal (SORENSEN *et al.* 2012).

As described above, once islet autoimmunity to one or more of the four autoantibodies associated with the development of T1D is present, the progression to T1D becomes almost inevitable, but environmental factors associated with the rate of progression include: being overweight or having rapid growth during adolescence, puberty, insulin resistance, and psychological stress (summarized in Figure 1.1). On the other hand, no protective factors have been identified (REWERS AND LUDVIGSSON 2016). Collectively, these studies suggest that more research into the environmental triggers of T1D is warranted, and that even for individuals with high genetic risk, interventions into environmental or lifestyle factors could reduce risk of the development of T1D or its rate or severity of progression.

At the individual level, the environmental triggers for the development of T1D as summarized above are only weakly understood, but at a population level, some risk factors include age, sex, race, geographic location, and seasonality (MAAHS *et al.* 2010). The DIAMOND registry, an initiative of WHO to track the incidence, prevalence and risk factors for the development of T1D in youth, report large differences in the incidence T1D among children from over 50 countries ranging from over 60 to 0.5 per 100,000 children < 15 years of age, with the highest rates reported in Scandinavian populations and Sardinia, and lower incidence rates in Asia and Latin America (GROUP 2006). Importantly, the incidence of T1D has been increasing by about 3% annually over the last few decades (GALE 2002; DABELEA 2009). In the US, the SEARCH study found a mean prevalence of T1D of 2.28/1000 in youth < 20 years or 5,399 cases in a population of ~3.5 million (GROUP 2006). A more recent review of epidemiological data from the US estimates that 1 in 300 youth < 18 years of age has T1D in the US (MAAHS *et al.* 2010).

1.3.3 HbA1c and the long-term complications of T1D

Once T1D is diagnosed, the average time to develop micro and macro vascular complications of the disease is 15-20 years (JACOBSON *et al.* 2013). The epidemiological risk factors associated with the development of complications include: duration of diabetes, glycemic control (typically as measured from haemoglobin A1c (HbA1c)), age, weight and waist-hip ratio, smoking status, cholesterol, triglycerides, and blood pressure. Of these, HbA1c is particularly important because the results of a large randomized control trial in the US from 1983-1993 entitled the Diabetes Control and Complications Trial (DCCT) showed that the major risk factor for both micro- and macrovascular complications of T1D is long-term hyperglycaemia (DE BOER *et al.* 2011; JACOBSON *et al.* 2013; GENUTH *et al.* 2015)

Glycemic control of T1D can be measured by frequent monitoring of glucose in the blood or urine or by the percent of glycated haemoglobin in the blood, so called-haemoglobin A1c (HbA1c). An early trial in a cohort of T1D patients found that monitoring of blood glucose via HbA1c led to improved treatment of metabolic control compared to monitoring only blood or urine glucose (LARSEN *et al.* 1990).

Adult haemoglobin, HbA, is the oxygen carrying protein found in red blood cells (RBCs). HbA is a tetrameric protein composed of two chains of haemoglobin A (HBA) and two chains of haemoglobin B (HBB). Circulating plasma glucose molecules may become attached to amino acids in HbA in a non-enzymatic glycation reaction during which sugar moities are attached to the N-terminal valines of HBB. Since the average lifespan for RBCs is 120 days and the glycation reaction is irreversible, levels of HbA1c in the blood provide an estimate of the average of the plasma glucose concentration over the preceding 2-3 months (average life span of the RBCs). Normal levels of HbA1c are lower than 6%, while a HbA1c value of 6.5% or more can be diagnostic of hyperglycemia.

HbA1c levels have been shown to be heritable, with heritability estimates ranging from 47% to 59% (MEIGS *et al.* 2002; PILIA *et al.* 2006). A genome-wide association study identified 11 single nucleotide polymorphisms (SNPs) in 10 different genetic loci associated with HbA1c levels (SORANZO *et al.* 2010). Collectively, they explained approximately 2.4% of the variance in HbA1c levels and 5% of heritability of HbA1c. However, environmental factors also influence variation in HbA1c; age and levels of C-reactive protein (a measure of stress) were found to be the strongest determinants of variation in HbA1c levels among Finnish men without diabetes (FIZELOVA *et al.* 2015).

Nevertheless, HbA1c is a marker of long-term glycemia. Prolonged hyperglycemia causes a number of micro and macro-vascular complications tissues, and leads to the accumulation of advanced glycation end products (AGEs), which are diverse protein or lipids that become glycated because of exposure to sugar moieties. AGE's can accumulate in most cell types and tissues in the body, and play a causative role in the micro-vascular complications in T1D, including retinopathy, nephropathy and neuropathy (GENUTH *et al.* 2015). High HbA1c can also cause an increase in the number of reactive oxygen species inside RBCs, which can alter the cell membrane properties and lead to blood cell aggregation and impaired blood flow, or lead to inflammation that can result in atheroscelerotic plaques, both of which become risk factors for CVD (SALEH 2015).

1.3.4 Diabetes Control and Complications Trial (DCCT)

As mentioned, in the early 1980s there was some debate about whether the long-term complications of T1D were caused primarily by hyperglycemia. The Diabetes Control and Complications Trial (DCCT) was a multicenter randomized clinical trial designed to compare intensive control with conventional control of glycemia for diabetes therapy with respect to their effectiveness in preventing development and progression of the early vascular and neurologic complications of T1D. A total of 1,441 participants with T1D aged 13-39 years were recruited from 29 medical centers in the United States and Canada between 1983 and 1989 and were followed up until 1993. The participants were randomly assigned to one of two groups. The standard or conventional diabetes therapy (CONV) group received one or two insulin injections daily, using a quantity of insulin that was based upon daily urine or self-monitoring of blood glucose (SMBG). The clinical goals of CONV therapy included absence of hyperglycemia and avoidance of severe or frequent episodes of hypoglycemia, a potentially fatal consequence of insulin therapy. On the other hand, individuals assigned to the experimental or intensive therapy group (INT), had their blood sugar levels monitored frequently (numerous times/day) and they received insulin injections at least 3 times per day (multiple daily injections, MDI) or treatment with continuous subcutaneous insulin infusion (CSII) with dose adjustments guided by four or more SMBG tests per day, meal size and content, and anticipated exercise. The overall goal of the INT therapy was to achieve and maintain HbA_{1c} levels less than 6.05%. In addition to patients being divided between CONV and INT therapies, patients were recruited as two cohorts to study two related questions: 1) Will intensive therapy prevent the development of diabetic retinopathy in patients with no retinopathy (primary prevention) and 2) will intensive therapy affect the progression of patients with early retinopathy (secondary intervention). Although the primary study outcome was retinopathy, the effect of CONV and INT therapies on multiple renal, neurologic, cardiovascular and neuropsychological outcomes was also measured.

The 1441 individuals enrolled in the study were followed up for a mean of 6.5 years and over this time period, the mean HbA1c of those in the CONV treatment was 9% while that of those in the INT treatment was ~7% (DIABETES *et al.* 1993). This reduction in HbA1c in the INT treatment group was associated with a 35–76% reduction in the early stages of microvascular diseases. Even though individuals in INT therapy also had a threefold increased risk of hypoglycemia, their hypoglycaemia events were not associated with a decline in cognitive function or quality of life (NATHAN AND GROUP 2014). Following the original DCCT, an observational follow-up study entitled the Epidemiology of Diabetes Interventions and Complications (EDIC) continues to follow all participants in the study not lost to follow-up or mortality to monitor many descriptors of their health including HbA1c, micro and macro-vascular complications, quality of life, etc. (JACOBSON *et al.* 2013). In addition to on-going protection against microvascular diseases, the EDIC study has found significant reductions in CVD risk afforded by INT therapy (WHITE 2015).

1.3 Primer on Genes and Gene Regulation in the Human Genome

The genome is the entire collection of genetic material in a cell and consists of long chains of deoxyribonucleic acid (DNA) that are tightly packaged into subcellular structures called chromosomes in the cell nuclei with an additional small independent genome in the mitochondria. In humans, there are normally 46 chromosomes arranged in 23 pairs in diploid, somatic cells, and 23 chromosomes in germ cells (sperm and egg cells). Twenty-two pairs of the chromosomes are called autosomes, and one pair are the sex chromosomes designated as X and Y, with males being the heterogametic sex (22 autosomes + XY) and females the homogametic sex (22 autosomes + XX). The germ cells contain ~ 3.2 billion base pairs of DNA. To store this vast amount of nucleic acids in each cell, short stretches of DNA are wrapped around groups of histone proteins into individual units called nucleosomes. The organization of the genome into nucleosomes is critical for storage but it also offers a high-level of control over gene regulation since winding or unwinding of the DNA in response to epigenetic signals silences or enables gene expression respectively (discussed below). On the other hand, the mitochondrial genome with only ~ 16,500 base pairs does not require the same level of compaction or regulatory control and lacks histones. Its 37 genes are primarily involved in performing or regulating cellular respiration, the primary function of the mitochondria. Although historically the mitochondrial genome was not considered an important factor in disease, this view is changing but will not be discussed further in this thesis (SCHARFE *et al.* 2009; HERST *et al.* 2017).

DNA consists of two long anti-parallel strands of nucleotides, each of which is made up of a deoxyribose sugar, a phosphate (attached to the 5' carbon of the ribose sugar) and one of four bases: adenine (A), guanine (G), cytosine (C) or thymine (T). Strings of nucleotides are connected via a phosphodiester bond linking the phosphate group, attached to the 5' carbon of a deoxyribose sugar, to a hydroxyl group attached to the 3' carbon of an adjacent sugar: thus DNA is said to have a 5' to 3' directionality. In the three-dimensional structure of DNA, the bases on the two antiparallel strands project inwards and are connected via hydrogen bonds based on the following strict base-pairing rules, A pairs with T, and C with G such that the two anti-parallel strands always have complementary base pairs.

1.4.1 Gene Expression

Genes are made up of DNA segments that can be transcribed into ribonucleic acids (RNA); the resulting RNA transcripts are classified into groups based on their function. One group is messenger RNA's (mRNA's) which are transported from the nucleus to the cytoplasm where they will bind to ribosomes, the site of protein synthesis, and be translated into proteins. The other classes of RNA's are divided into two groups, long non-coding RNA's (lncRNA's) which are longer than 200 nucleotides, and small non-coding RNA's. LncRNA's are emerging as an important part of the non-coding genome and play diverse functions in the regulation of gene transcription, post-transcriptional, and epigenetic modifications. On the other hand, small non-coding RNA's include molecules such as microRNA's (miRNA's), small-interfering RNA's (siRNA's), and small nucleolar RNA's (snRNA's), and are involved in modifying mRNA's or regulating their transport from the nucleus to the cytoplasm. In the current annotation of the human genome (GRCh38, version 21, Ensembl77; www.ensembl.org), there are approximately 60,000 genes; ~20,000 of these are protein coding genes, ~16,000 lncRNA's and ~7500 small non-coding RNA's. Perhaps surprisingly, the remaining ~ 15,000 genes are pseudogenes, segments of DNA which are similar to functional genes but have lost their functionality relative to a complete gene.

Following transcription, mRNA's are translated into proteins using a three-letter code in which groups of three RNA nucleotides, called codons, are translated into amino acids forming polypeptide chains that will ultimately be folded into archetypical 3-dimensional protein structures. With four possible nucleotides (A, C, G and Uracil (which replaces thymine) in RNA), there are 4^3=64 possible codons; 61 codons encode for one of the twenty essential amino acids while 3 are "stop" codons and initiate termination of translation. Since there are 61 codons but only 20 amino acids, multiple triplet codons may code for the same amino acid, a phenomenon called codon degeneracy.

Codons are located in exons, which are the coding segments of protein-coding genes, and genes may contain a few to hundreds of exons separated by intervening non-coding segments called introns. For those genes that contain introns, the introns are removed following transcription in a process called splicing, which, for genes containing two or more introns, allows different mRNA transcripts to be produced from the same gene, a process called alternative splicing, which generates multiple transcript isoforms per gene (Figure 1.2). In addition to introns, pre-mRNA molecules also contain segments at the beginning and ending of the transcripts that are not translated – named the 5' and 3' Untranslated Regions (5' and 3'UTRs) (Figure 1.2). The average length of the human 5'UTR is 210 base pairs (bp), while the average length of the 3' UTR is 1027 bp (MIGNONE *et al.* 2002). Modifications to the 5'UTR typically affects translation efficiency of mRNA (e.g. binding of proteins to the 5' UTR that block translation), while modifications to the 3'UTR tend to affect mRNA stability, (e.g. binding of miRNA's to the 3' UTR that initiate transcript degradation in a process called RNA interference (MIGNONE *et al.* 2002)). Post-transcriptional modifications to the 5' and 3' UTR regions of mRNA's can greatly reduce the number of mRNA transcripts that are translated, which explains part of the reason why there is a notoriously poor correlation between mRNA and protein abundance in cells (MAIER *et al.* 2009).

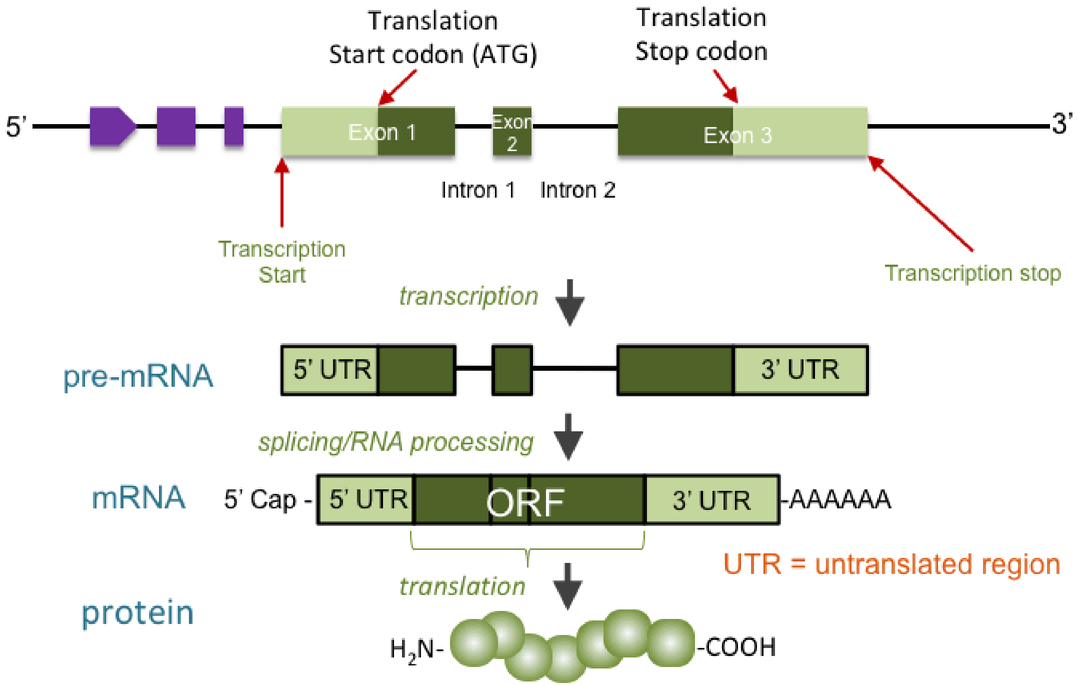


Figure 1.2 Structure of a gene and an overview of gene expression. Purple rectangles and arrows refer to regulatory regions (discussed below and in Figure 1.3). ORF – open reading frame, UTR- Untranslated Region.

In order for transcription to occur, a variety of molecular signals must first be initiated in a complex process known as gene regulation. In the last 10-15 years, tremendous developments in genomic technologies, such as RNA-Seq, CHIP-seq, and capture Hi-C (cHi-C), have made it possible to both identify regulatory elements and to examine the genetic and environmental factors influencing gene regulation (ROMERO *et al.* 2012).

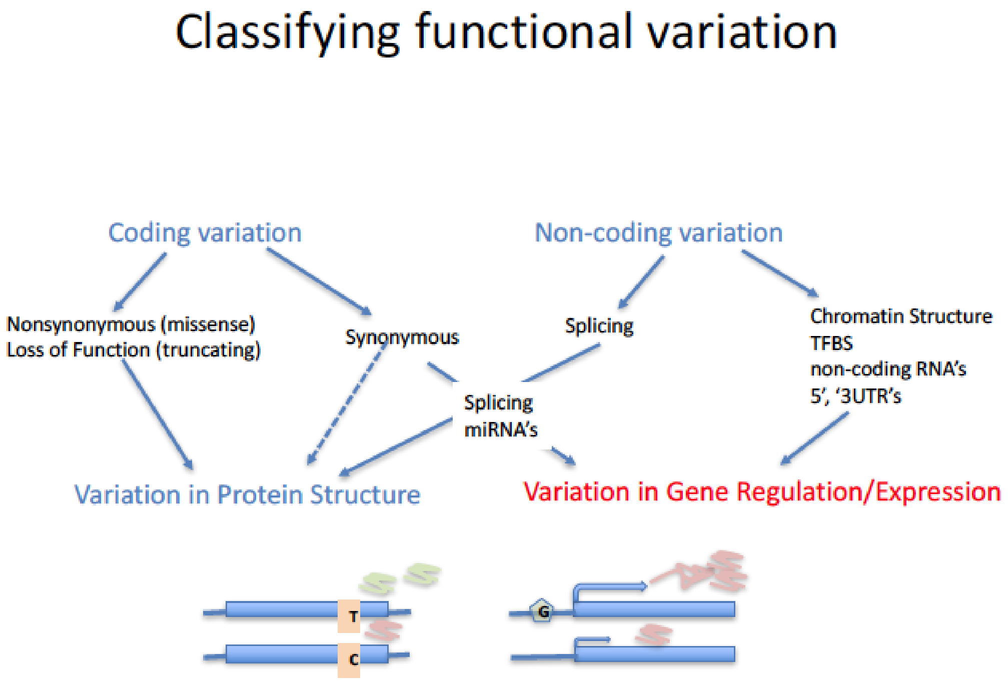


Figure 1.3. Classifying functional variation in coding and non-coding regions of the genome. miRNA-microRNA's, TFBS – transcription factor binding sites, UTR-Untranslated Regions (see text for discussion).

1.4.2 Gene Regulation

The first level of control of gene expression is regulated by epigenetic changes to histone proteins that influence the degree of compaction of chromatin, the complex of DNA and proteins that make up chromosomes. In its most closed form, DNA is tightly wound around histone proteins in units called nucleosomes and is inaccessible to the proteins that initiate transcription. Changes in chromatin remodeling occur via suites of epigenetic modifications, which include histone methylation, acetylation and sumoylation. Changes in the chemical groups (methyl, acetyl) attached to histone tails embedded in nucleosomes influences chromatin compaction. In addition, DNA methylation in promoter regions is associated with gene silencing. Collectively, these epigenetic modifications are known to have major influences on gene regulation, and are primary factors dictating differences in gene expression between tissues and among individuals (CHEN *et al.* 2017).

Assuming DNA is accessible, transcription is initiated when proteins called transcription factors bind to idiosyncratic DNA motifs, called transcription factor binding sites (TFBS) located just upstream of the regulatory and core promoters of a gene (Figure 1.4). In response, RNA polymerase binds to the core promoter and begins transcribing a gene at the transcription start site (TSS; typically located in the first exon). Core and regulatory promoters typically occur within the first few hundred (core) or thousand (regulatory) base pairs upstream of the TSS. Collectively, the proteins that bind to DNA and affect levels of DNA transcription are called trans-factors or trans-regulatory elements, while the DNA motifs that they recognize and bind to are called *cis*-regulatory elements. Binding of trans-acting proteins to *cis*-elements occurs via characteristic DNA binding domains located in the trans factors, which recognize cognate motifs and their corresponding 3-D structures in the DNA.

In addition to the proximal promoter (composed of the core and regulatory promoters), binding of trans-acting factors to other more-distally situated *cis*-acting elements, called enhancers and silencers, can increase or decrease rates of transcription respectively. The trans-acting factors that bind to enhancers are called activators while those that bind to the silencers are called inhibitors or repressors (Figure 1.4). Enhancers and silencers affect the rate of transcription because when activators or inhibitors are bound to them, they form complexes that loop over and create cross-bridges to the transcription factors bound to the core and regulator promoter and thereby enhance or weaken the strength of the binding associations and rate of transcription. MIFSUD *et al.* (2015) used capture Hi-C (Chi-C) to examine the long-range interactions of almost 22,000 promoters in 2 human blood cell types. They found that the average distance of long-range interactions between *cis*-elements was 118 Kb, while loops could range up to 1.6Mb, a window of 1Mb (on either side of the gene) would capture 98% of interactions for *cis*-acting SNPs (MIFSUD *et al.* 2015). A window size of 1Mb is often used for gene-based association studies (see below), although some studies suggest that a better window size should be up to 2Mb (BRODIE *et al.* 2016). Nevertheless, collectively, this underscores that genetic variants in intronic, proximal, and distal (up to 1-2Mb away) *cis* regulatory elements or in the trans-acting factors that bind to them could all affect rates of transcription, thereby acting as eQTL's.

1.4.3 The Encode Project and identification of Conserved Non-Coding Elements.

When fully sequenced genomes first became available in the early 2000s, regulatory sequences were inferred by performing DNA alignments between species and looking for conserved and ultra-conserved regions in non-coding parts of the genome. For example, the divergence time of human and mouse lineages occurred approximately 70-90 million years ago, and comparisons of their genomes revealed that when non-coding regions exhibited 70% similarity or greater, the region could be said to be highly conserved and may contain regulatory elements (ZHANG AND GERSTEIN 2003). Once these conserved non-coding elements were identified, a bioinformatics approach known as phylogenetic footprinting was used to search for short motifs, i.e. probable TFBS, using hidden markov modeling (HMM) to identify putative regulatory sites (BLANCHETTE AND

TOMPA 2002). This approach could provide *ab initio* evidence for TFBS, but experimental verification was still needed to confirm that it was a regulatory element.

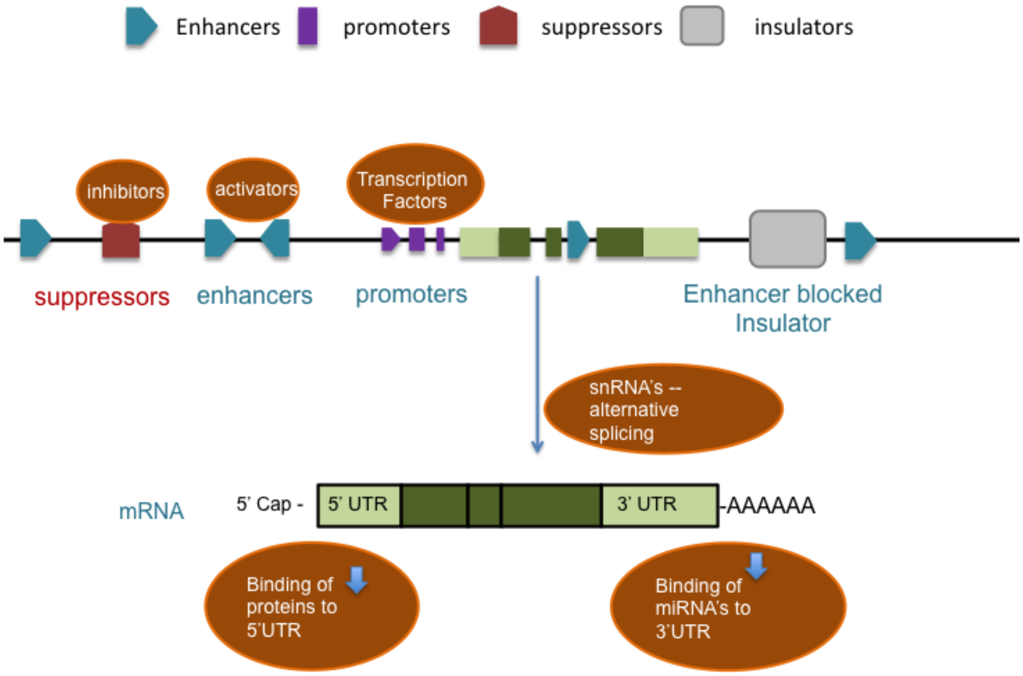


Figure 1.4. Major players in gene regulation.
Transcription begins when transcription factors bind to the core and regulatory promoters just 5'upstream of the TSS. The rate of transcription is enhanced by the binding of regulatory elements to the regulatory promoter (not shown) and by the binding of activators to cis-acting enhancer elements which may be located upstream or downstream of a gene or in introns. The binding of inhibitors to cis-acting suppressor elements will decrease the rate of transcription. The presence of insulators in between an enhancer and gene can also suppress transcription of the gene. Additionally, the amount of protein produced can be down-regulated by the binding of small proteins to the 5' UTR which block translation or by degradation of the mRNA via processes such as binding of miRNA's to the 3'UTR. A variety of other proteins and short RNA's (such as snRNA's) can influence the production of alternative transcripts.

Since 2012, an international consortium known as the The Encyclopedia of DNA Elements (ENCODE) has been systematically mapping and integrating information about candidate regulatory elements (cRE) in the human and mouse genomes, including elements involved in transcription, transcription factor association, chromatin structure and histone modification (CONSORTIUM 2012). The project has now identified over 1.3 million cRE in humans and over 500,000 cRE in mice, and estimates that while 1.5% of the human genome is composed of protein coding exons, anywhere from 20-40% of the human genome is made up of regulatory elements that are essential for controlling chromatin remodelling and gene expression (www.encodeproject.org). This has led to a radical change in our understanding of the function of the vast regions of non-coding DNA in the genome.

1.4.4 Genetic Variation and Gene Expression and Regulation

Heritable genetic variations are ultimately introduced into populations by mutations in the germline which are passed from parent to offspring (inherited mutations) or that arise *de novo* in the germline of a given individual (*de novo* mutations). Genetic variants are broadly divided into two classes: point and large-scale variants. Point variants include bi-allelic single-nucleotide polymorphism (SNPs), multi-allelic nucleotide polymorphisms (tri and tetra-allelic SNPs), small-scale (1-8 base pair) insertions and deletions (indels), and variation in the number of short tandem repeats (microsatellites). Large-scale variations include inversions and copy number variations. SNPs occur when there is a single base-pair substitution at a site leading to the presence of two nucleotides segregating at a specific position in the genome. The two different nucleotides are called alleles of a SNP. When an individual has two copies of the same allele, the individual is said to be homozygous for the allele, while individuals with two different alleles are heterozygous.

For gene expression, DNA substitutions that do not lead to a change in amino acid (i.e. use a different codon for the same amino acid) are called synonymous changes, substitutions that lead to a change in the resulting amino acid are called nonsynonymous or missense changes, and substitutions that introduce a premature stop codon are called loss of function (LOF) or nonsense substitutions. The functional impact of these three different types of substitutions is usually very different: LOF changes are, in general, the most deleterious, synonymous changes are often neutral and the effect of missense changes depends on where they occur within the three-dimensional structure of the protein they encode and whether the new amino acid results in a conservative or a radical change in chemical properties.

Despite these broad differences in the functional impact of substitution types, genes also differ dramatically in how tolerant they are to mutation, in particular to LOF and to non-synonymous changes. Comparative evolutionary analysis has found that genes that exhibit few amino acid changes over evolutionary time, are often essential for survival and are intolerant to LOF mutations, meaning that embryos containing such mutations die prior to birth and individuals with LOF mutations are rarely observed in populations (SAMOCHA *et al.* 2014). Building upon this principal, a number of statistics have been developed to use the number, kind and position of changes observed in a gene to predict the probability that a given mutation will be deleterious and potentially contribute to a disease state (e.g. PolyPhen, SIFT) (ADZHUBEI *et al.* 2010; SIM *et al.* 2012). One such elegant model was developed by SAMOCHA *et al.* (2014) who developed a sequence-context based selection-neutral model to compare the observed number of rare variants per gene to the expected number, and quantified the deviation from expectation with a Z-score. Based on these Z-scores, human protein-coding genes were categorized into groups (tolerant, recessive or haploinsufficient) based on how intolerant they are predicted to be to LOF mutations, and additionally given a metric, pLI to describe their intolerance to mutation. Subsequent analyses of exome-wide protein-coding variation in 60,706 individuals, led to the identification of >3,200 human genes that are highly intolerant to changes, rendering them good candidates for disease–associated phenotypes (LEK *et al.* 2016).

For gene regulation, since the influence of the 5' and 3' UTR's on gene expression depends largely on the binding of small RNA's and proteins to these regions, genetic variants (e.g. SNPs) in the UTR's can have direct impacts on RNA and protein abundance in the cell. Thus SNPs in these regions may directly impact gene expression although classifying the functional effect of variants in coding versus non-coding regions can be difficult (Figure 1.3).

Genetic variants in *cis*- and *trans*- regulatory elements may affect more than just the abundance of RNA or protein produced by a gene. In the current release of the human genome, the ~60,000 genes encode ~ 204,000 transcripts; while the ~20,000 protein coding genes encode ~ 82,000 transcripts, ~56,000 of which produce full-length functional proteins and ~25,000 produce partial length proteins (www.genecodegenes.org). Different transcript isoforms are generated by multiple mechanisms such as alternative TSS, alternative splicing, and/or differential termination of transcript length at the 3' end. The number of transcript isoforms produced per gene is highly variable: all one and two exon genes will produce a single transcript, the average number of transcript isoforms in humans is five, and a few genes have as many as 80 different transcript isoforms (FLOOR AND DOUDNA 2016). Incorrect transcript isoforms may be produced by changes in *cis*-regulatory elements, core components of the splicing machinery and/or in the trans-acting regulatory factors that mediate transcription. A recent review suggests that changes associated with RNA mis-splicing causes a large array of human diseases due to genetic variants in both germ-line (hereditary) and somatic cells (SCOTTI AND SWANSON 2016).

In light of the fundamental role of gene regulation, the finding that ~ 90% of GWAS associated SNPs fall into non-coding regions of the genome (HINDORFF *et al.* 2009) is less surprising. As of 15th of April 2018, the GWAS catalogue contains almost 60,000 SNP-trait associations identified from over 3,300 studies (www.ebi.ac.uk/gwas). MAURANO *et al.* (2012) performed genome-wide DNase I hypersensitive assays on 85 cell types to quantify the enrichment of genome-wide significant associated genetic variants in *cis*-regulatory elements. DNase I hypersensitive sites (DHSs) are sensitive and precise markers of *cis*-regulatory elements because they identify sites of TF binding (i.e. TFBS) or sites that alter allelic chromatin states (i.e. sites that affect the opening and closing of chromatin). They found 40% enrichment of GWAS SNPs in DHSs, and 76.6% of the 5,654 non-coding GWAS SNPs examined were found either within a DHS (57.1%, 2,931 SNPs) or in complete linkage disequilibrium (LD) with SNPs in a nearby DHS (19.5%, 999). A smaller percentage of the non-coding GWAS SNPs were predicted to disrupt or create micro-RNA (miRNA) that bind to the 3′-UTR and alter splicing isoforms or mRNA stability (MAURANO *et al.* 2012). In addition to a significant role for GWAS SNPs to be found in TFBS or micro-RNA's, there is growing interest in the possibility that disease associated variants are in lncRNA's. LncRNA's were discovered relatively recently (HARROW *et al.* 2012), but since lncRNA's are primarily involved in regulating the expression of protein-coding genes, which are dysregulated in many diseases and have been shown to play a role in cancer cells, there is increasing interest in examining their association with disease phenotypes (HRDLICKOVA *et al.* 2014).

The finding that genetic variants in non-coding regions are associated with disease is consistent with evolutionary studies that have consistently shown that divergence in gene regulation occurs more rapidly than protein sequence during speciation and likely plays an important role in adaptation and phenotypic evolution. Indeed in 1975, King and Wilson famously predicted that the vast phenotypic differences between humans and chimpanzees could be more driven by differences in gene regulation than protein evolution (KING AND WILSON 1975) and comparative experimental work has shown that TFBS and epigenetic marks change rapidly and underlie many phenotypic differences among species (ROMERO *et al.* 2012). In one of the first comparative functional genome-wide studies of transcription factor binding, CHIP-seq was performed on liver samples from five vertebrates, including humans, to examine the evolution of TFBS for two transcription factors. Of the ~16,000-30,000 binding sites identified in each species, only 35 were shared across all five species, and only 344 were shared by the three mammalian species (humans, mice and dogs) (SCHMIDT *et al.* 2010). Another study comparing binding sites of RNA polymerase II in human and chimpanzee found that 32% of the binding sites in immortalized B cell lines differed between human and chimpanzee and 25% different among human individuals. This suggests that the evolutionary turnover of TFBS is rapid and on a genome-wide scale differences in SNPs embedded in TFBS could even contribute to phenotypic differences among humans (KASOWSKI *et al.* 2010).

Similarly, comparative studies have found that a substantial, though smaller, fraction of differences in gene expression among closely related species are associated with epigenetic changes (PRABHAKAR *et al.* 2008). For instance, experimental work using H3K4me3 - Chip-seq, which identifies histone marks associated with active transcription, in immortalized B cells from humans, chimpanzees, and rhesus macaques, identified large differences in patterns of histone modification among species except near TSSs, where H3K4me3 is most likely to be most functional (CAIN *et al.* 2011). Another study comparing gene expression levels with promoter DNA methylation status in three tissues from humans and chimpanzees, found, as expected, greater methylation differences among tissues than species, but still estimated that 12-18% of differences in gene expression between human and chimpanzee could be explained by changes in DNA promoter methylation alone (Pai *et al.,* 2011). Collectively, these comparative studies underscore that changes in gene regulation evolve rapidly, and are probably a primary driver of phenotypic evolution.

1.4.5 Single Nucleotide Polymorphisms and the HapMap and 1000 Genomes Projects

Compared to other types of genetic variants, SNPs are the most commonly used in GWAS/TWAS for associations with human diseases/traits. SNPs do not occur homogeneously across the human genome, but rather their frequency differ between genes, genomic regions and regions of a gene reflecting differences in mutation and recombination rate across the genome as well as the degree of selective constraint. In 2001, collaboration between the International SNP consortium and the International Human Genome Sequencing Consortiums led to the publication of the first genome-wide map of SNPs, which identified 1.42M SNPs distributed at an average density of 1 every 1.9kb throughout the genome (SACHIDANANDAM *et al.* 2001). These SNPs became the foundation for the first genotyping arrays used by Illumina and Affymetrix. The National Centre for Biotechnology Information (NCBI) hosts a free public archive of all registered SNPs in humans, entitled dbSNP, in collaboration with the National Human Genome Research Institute (NHGRI) (SHERRY *et al.* 2001). In the current build, dbSNP build 151, there are now over 600 M registered variants in humans, with an average density of 1 SNP every 5bp.

SNPs arise in a population via mutation, and since a mutation first occurs in a given individual from a given population, the new variant (SNP) is associated with all the pre-existing alleles/variants in that chromosomal region haplotype in that individual. Over time, these associations will be broken up at a rate dependent upon the distance between variants, the larger the distance the faster the decay, and the recombination rate, the lower the recombination rate the longer the associations will be maintained (LEWONTIN 1988). In population genetics, this non-random association of alleles at two or more loci, not necessarily on the same chromosome, is called linkage (or gametic) disequilibrium (LD). As a result of LD, SNP alleles that are close together in a genome tend to be inherited together, and the term haplotype or haplotype block is used to refer to the set of alleles or DNA sequences that are inherited together (BUSH AND MOORE 2012). SNPs that are in high LD with a set of other SNPs in a haplotype block and thus can be used to represent them are called tag SNPs. Tagged SNPs are used on genotyping arrays and they allow one to impute (or infer) the alleles at neighbouring ungenotyped SNPs based on LD (HALPERIN *et al.* 2005). Because many of the earliest GWAS studies were performed on Caucasian populations, which have a relatively recent shared history and therefore relatively high levels of LD, the first genotyping arrays harboured ~ 1M tagged SNPs, but subsequently imputation can be used to impute another 4M or more SNPs, such that ~ 5M + SNPs can be included in a GWAS.

Measures of LD and the development of a set of genome-wide tagged SNPs were first performed by the HapMap project. The primary goal of the first phase of the project was to map the entire human genome to haplotype blocks that would describe the common patterns of genetic variation in human populations and then to assign a tag SNP to represent each block. The project initially focused on SNPs with a minor allele frequency (MAF) > 5% (INTERNATIONAL HAPMAP 2003), although by the third phase of the project information on common and rare (<5%) variants were integrated (INTERNATIONAL HAPMAP *et al.* 2010). Because the degree of LD, the size of haplotype block and the MAF of alleles of SNPs vary among populations, a major initiative of the HapMap project was to genotype individuals from multiple ethnic groups. In phase I of the project, individuals from 4 ethnic groups were included and by the third phase, the genetic variants and LD structure of 11 ethnic groups was incorporated thereby permitting analyses of population/continental differences in allele frequencies and the size of haplotype blocks (INTERNATIONAL HAPMAP *et al.* 2010).

While the primary goal of the International HapMap project was to identify genetic variants and assign them to haplotype blocks, in 2008 an international research collaboration called the 1000 Genomes Project (1KGP) was launched with the goal of providing a detailed catalogue of human genetic variation for understanding the genetic basis of disease. To this end, they proposed to sequence 1000 genomes collected from diverse global populations. In 2010, the pilot phase of the project was complete (GENOMES

Project *et al.* 2010), and in 2012 the sequencing of 1092 genomes was announced (Genomes Project *et al.* 2012). In 2015, the 1KGP consortium reported the completion of the project with the sequencing of 2504 genomes of individuals from 26 populations. In the current catalogue, there are over 88 million genetic variants (84.7 million SNPs, 2.6 million short indels, and 60,000 structural variants), estimated to have captured diversity at 99% of SNPs with a frequency of 0.1% in diverse ancestries (Genomes Project *et al.* 2015). The 1KGP catalogue has thus become an important tool for understanding the distribution and phenotypic associations of genetic variants in human populations, and allows imputation to be performed at a high level of coverage and accuracy with respect to the LD structure of haplotype blocks.

1.5 Genome-wide and Transcriptome-Wide Association Studies (GWAS & TWAS).

1.5.1 GWAS

As outlined above, GWAS offer a powerful approach to test for the statistical association between common multifactorial diseases and common genetic variants (SNPs). The association can be tested using logistic regression (case-control), or linear regression (continuous phenotype), or another suitable model, in which each SNP is tested independently for its effect on the phenotype while adjusting for covariates. The effect of the disease variant is then estimated by the effect size (beta coefficient) of the SNP effect (Bush and Moore 2012), and a multiple test correction is used to ensure stringency, typically $p < 5.0 \times 10^{-8}$ for 1M indendent tests (Risch and Merikangas 1996). The theoretical framework for GWAS is that common diseases may be caused by multiple common variants with small effect sizes or by multiple rare variants that are in strong LD with the common variants. Complex traits are influenced by both genetic and environmental factors. The portion of the total phenotypic variance due solely to the sum of causal additive genetic variants is termed heritability (h^2) (WRAY *et al.* 2013). In regression models used to test associations between a SNP and a disease, the coefficient of determination (R^2) measures the amount of variation in the trait that is explained by predictors in the model and in models with only genetic variants (i.e. no covariates) an upper bound of R^2 is h^2. It is important to note that when a genetic variant is found to be associated with a disease, it may be unclear whether the disease is associated with the identified SNP, or to one in strong LD with it. Once a potentially causal SNP has been identified, GWAS results should be validated in larger/different cohorts and/or by performing high-resolution sequencing of the putative chromosomal region and fine-mapping the association.

A first step in performing GWAS is to collect high quality genome-wide data of genetic variants (SNPs) present in the study population through the use of genotyping arrays. Once the genotyping data is obtained, the next step is to perform stringent quality control (QC) measures to remove individuals and markers with incongruent or inconsistent data to reduce the false positive rate (BUSH AND MOORE 2012). Typical workflows to remove aberrant individual data consist of removing individuals with a) discordant sex information, b) outlying missing genotype or heterozygosity rates, c) related individuals, and d) divergent ancestry. On the other hand, individual SNPs are removed if a) they have high levels of missing data, b) genotype frequencies differ significant from that expected under Hardy-Weinberg equilibrium (HWE), or c) missing genotype rates are very different between cases and controls (or cohorts depending on the study design) (ANDERSON *et al.* 2010).

As part of or following QC, imputation is performed to fill in missing genotypes where possible and expand the number of genotypes examined in the GWAS to all those that can be well imputed using a reference panel, for example the HapMap data or 1000 Genomes. Genetic variants with MAF below 0.05 (or 0.01) are normally excluded from analyses because estimating the effect size of a variant is difficult when an allele is not present in sufficient frequency in both the cases and controls (BUSH AND MOORE 2012). Increasingly however, the availability of large numbers of fully sequenced genomes in diverse populations (e.g. 1000 Genomes, UK biobank) has improved the availability of LD information for rare variants and robust statistical methodologies are available to test for associations of rare variants with disease (AUER AND LETTRE 2015).

The genotype of individuals identified on the array is assessed using genotype-calling algorithms based on the probe intensity at each SNP, and individuals are assigned one of three possible genotypes (e.g. AA, Aa and aa, where a is the minor allele) at each marker. Genotype calls that do not meet the threshold signal intensity/clarity become missing calls and they may be removed or imputed based on HapMap or 1000 Genomes data. After final genotype QC and imputation has been performed, dosages for the SNPs are calculated from the hard-called and imputed genotypes in which the former take on values of 0, 1 or 2 for AA, Aa and aa respectively, and the imputed dosages take on continuous values from 0 to 2.

Population or subgroup stratification may confound GWAS results, particularly if there are differences in the frequency of alleles between cases and controls. The confounding influence of population structure can also be assessed in multiple ways, including by testing for Hardy-Weinberg equilibrium (HWE), which will also identify genotyping errors. In the absence of selection and population subgroup stratification, the frequency of the three genotypes (AA, Aa and aa) per SNP should follow those expected under Hardy-Weinberg equilibrium (HWE), namely a binomial distribution (with 1 d.f.). This assumption is tested by comparing the expected and observed genotype frequencies using a Pearson χ2 test per locus. The expected genotype frequencies are obtained from using the mean observed allele frequency of the two alleles in the population/cohort assayed. Markers (SNPs) that show significantly different genotype frequencies than expected under HWE in controls are typically removed.

Secondly, cohorts used in GWAS may come from divergent ancestral backgrounds, which may, again, confound interpretation of the results. To control for the effects of divergent ancestry, principal component (PC) analysis is widely employed (ANDERSON *et al.* 2010). To this end, the genotype matrix is used to extract linearly uncorrelated PCs which are in order of decreasing importance. The top PCs describing the ancestry effects may be used to exclude or control for differences in population ancestry.

Once the QC is finished, statistical analysis is performed. To this end, the phenotype(s) is regressed onto each SNP individually, in a model that may or may not include other covariates, and the estimate of the SNP-effect is tested and ranked according to the p-value. For quantitative traits, linear regression models can be used and the SNP effect is modeled as having additive effect on the phenotype (LETTRE *et al.* 2007). If a genotyping arrays assessed ~ 1M independent markers (e.g. SNPs) simultaneously, employing a Bonferonni correction to achieve a FWER of 0.05, would require a SNP to have a p-value of ~ 5×10^{-8} to achieve a genome-wide significance (RISCH AND MERIKANGAS 1996). Since the significance of the SNP effect is dependent on the standard error of the estimate of the SNP effect, having a large sample (population/cohort) size for the study makes it more likely to detect significant associations (SPENCER *et al.* 2009). Since most SNPs are expected to have no effect on the phenotype, the distribution of p-values resulting from the GWAS should follow a uniform distribution. Quantile-quantile (Q-Q) plots are used to visualize the distribution of observed vs expected under the null p-values, in rank order and the majority of variants should follow the line y = x. Some variants will have small p values due to both chance and a real association with the trait, so only the presence of multiple p values that deviate substantially from the null distribution are of particular concern, and may indicate quality control or population stratification effects, or in large samples that the trait is heritable (YANG *et al.* 2011b). These can be evaluated by estimating the genomic control inflation factor for population substructure (λ_{gc}) and then subsequent modifications to the p-values or data-selection can be performed (DEVLIN AND ROEDER 1999). The results of GWAS are visualized using Manhattan plots in which the $-\log_{10}$ p value of the SNP association statistic p value is plotted against the genomic position of the SNP using a gradient of colours to represent chromosome number. Given the predominantly uniform distribution of p-values, most points cluster at the base of the plot, and a few significantly associated SNPs may be present at or above the FWER corrected α level.

1.5.2 Gene-based Association Studies

Although the statistical framework and pipeline for GWAS analyses is robust, the biological link between GWAS significant SNPs and disease or phenotypic effects has been difficult to uncover, in large part because ~ 90% of GWAS significant SNPs have been found in non-coding regions of unclear function (MAURANO *et al.* 2012; ZUK *et al.* 2014). For this reason, a variety of statistical approaches have been developed to aggregate SNPs into more meaningful biological units. Genes are the most-obvious unit around which to aggregate SNPs, and in gene-based approaches, relevant SNPs within and flanking a gene are aggregated and their effect on the phenotype, which is often the amount of RNA produced by the gene, modeled. A window size of 1Mb is typically employed to capture the effect of nearby and long-range cis-regulatory interactions, so in some cases (i.e. high density gene regions) the same SNP may appear in multiple genes. While this violates the assumption of independence of the genes, it has been argued that this is an advantage of gene based approaches and may facilitate identification of more risk alleles (PENG *et al.* 2010; BRODIE *et al.* 2016) because it provides a better molecular basis of phenotypes.

When testing for associations in gene-based approaches, all of the variants in the gene and its' corresponding 1Mb window need not be included in the model, or at least should not be given the same weight, because the LD structure of gene variants enables complete information to be obtained from a subset of variants. Correspondingly, statistical approaches to select variables (the gene variants in this case) are often used to select the set of SNPs (and potentially their weights) that best predicts the phenotype, i.e. RNA abundance. Three of the most common approaches for selecting variables are ridge regression, LASSO (least absolute shrinkage and selection operator) and elastic net. In a normal linear regression with n observations, in which each observation has one response variable, y, and p predictors

$Y = (y_1, \dots y_n)^T, nx1$

$X = \left(x_1, \dots x_p\right), nxp$

We want to find a linear combination β of predictors $X = \left(x_1, \dots x_p\right)$, to describe the relationship between y and $x_1, \dots x_p$,

$\hat{y} = X^T\beta$

Ordinary least squares (OLS) is commonly used to estimate this linear relationship in which

$Y_{nx1} = \ X_{nxp}\beta_{px1} + \sigma\ \epsilon_{nx1}$

And summed over all predictors

$y_i = \ \sum_{j=1}^{p} x_{ij}\beta_j + \ \sigma\varepsilon_i, i = 1, \dots, n,$

given that,

$\epsilon_i \sim^{i.i.d}$ N(0,1)

The OLS solution is uniquely defined when n>p and the matrix X^TX can be inverted. In OLS, $\hat{\beta}$ is found by minimizing the residual sum of squares (RSS) of the (X,y) data set given an n × p matrix X. This generates the best linear unbiased estimator of β, which has a variance given by:

$$Var(\widehat{\beta}) = (X^T X)^{-1}\sigma^2$$

This means that in the OLS regression the variance of the coefficients increases with n and p, and if there is multicollinearity among the predictors, as would normally be the case for SNPs in a gene-based model, the variance of the estimator will be high. Additionally, using OLS to estimate $\hat{\beta}$ requires that n>p (which again may not be the case if there are many SNPs and only a few individuals in the study) and the prediction error of $\hat{\beta}$ increases linearly as a function of p. In situations where there are many predictors (including when p>n) and high levels of multicollinearity, methods to remove correlated variables (e.g. SNPs) and/or shrink the size of their coefficients are preferable. Regression regularization refers to an extension of linear regression techniques in which additional information is introduced to the regression to solve a problem (e.g. p>n) or prevent over-fitting. Regularization of the regression imposes a penalty, λ, on how the RSS is minimized such that:

RSS + λ x penalty on the parameters

In ridge regression, the so-called L2-norm penalty is imposed and the predictors are shrunk to continuous non-zero values. Predictors that are highly correlated will be grouped which reduces over-fitting, but no predictors are eliminated which means that it cannot find the most parsimonious model. In LASSO, the coefficients of the predictors are shrunk as in ridge regression, but there is a second penalty, the so-called L1 norm penalty, which imposes a sparsity penalty and drops weak predictors by setting their coefficients to zero to achieve a parsimonious model. Because LASSO shrinks the coefficients but also enforces sparsity, it is good for eliminating trivial SNPs but not good for grouped selection. The third option, Elastic net, is another regularized regression method that linearly combines the L1 and L2 penalties of LASSO and ridge regression. Elastic net overcomes the limitation of LASSO to select one variable from a group and ignore the others. Furthermore, although all three methods work when p>n, in the case of LASSO, when p>n, it will select at most n variables. Elastic net overcomes these weaknesses by adding a quadratic term to the penalty, first it groups the predictors based on λ2 to reduce multicollinearity, then it uses the λ1 penalty to shrink the number of groups. The relative importance of the λ1 versus λ2 penalties in Elastic net is controlled by a tuning variable α, and when α=0.5, λ1 and λ2 have equal weight. In sum, Elastic net allows selection of a group of correlated predictors; it has proven useful for gene-based association tests.

Recently, a Bayesian approach for modeling the contribution of multiple, perhaps correlated, SNPs to a given phenotype was introduced called Bayesian Sparse Linear Mixed Model (BSLMM) (ZHOU *et al.* 2013). Developed specifically for genetic data, BSLMM employs a linear model with both a polygenic and sparse component to incorporate large and small effect SNPs into the model. It then estimates the distribution of effect sizes (β coefficients) of all SNPs jointly. As in ridge regression, there is no variable selection, only shrinkage (although some coefficients can have a β coefficient of zero), but the method assumes that the large and small genetic effects come from a mixture of two normal distributions. Because the method was developed specifically to estimate the heritability of phenotypes, the authors provide BSLMM solutions to estimate the proportion of phenotypic variance explained by the additive genetic variance (i.e. the narrow sense heritability, h^2_N) and the proportion of the variance in phenotype explained by sparse (large effect) effects.

1.5.3 Functional Genomics and Transcriptome-Wide Association Studies (TWAS).

The transcriptome consists of all of the ribonucleic acids (RNA) in a given cell type at a particular time, i.e. the entirety of gene expression, while the mRNA fraction describes that portion of the transcriptome that could potentially code for proteins (because some mRNA transcripts will not be translated). Non-coding variants up to 1Mb distally and in *cis* to a gene may influence the probability and quantity of transcription that occurs in a given cell type/time. *Cis*-variants influencing gene expression may be in promoters, enhancers, repressors, insulators, and/or chromatin or histone remodeling domains, while trans factors may be in transcription factor proteins that are encoded elsewhere in the genome but which bind to the *cis*-regulatory elements at the binding domain. In the first functional genomic studies, eQTL's were identified using standard linear regression in which

gene expression ~ genotype + covariates + error

(LAPPALAINEN *et al.* 2013). In a complete eQTL analysis, all of the SNPs in a genotyping array (i.e. all possible cis and trans variants, potentially 5M or more SNPs) are used to predict the expression of each of the protein coding genes (typically ~ 12,000 genes) in a given cell type/tissue, rendering there to be a staggering ~ 84 billion tests. Given the severe multiple testing burden of a full eQTL analysis, most eQTL analyses are performed only on possible *cis* variants (LAPPALAINEN *et al.* 2013). The first functional genomic studies employed sequential single SNP analyses, and most genes were found to have at least one *cis* -eQTL influencing expression in at least one tissue, but it was difficult to identify the causal functional variant because of the high LD structure of SNPs in the associated regions (LAPPALAINEN *et al.* 2013). Despite this weakness, a systematic review of functional genomic studies showed that eQTL's were more likely to be associated with complex traits and diseases as annotated in the GWAS catalogue (NICOLAE *et al.* 2010). This further supported the hypothesis that variation in gene expression is associated with variation in disease risk or severity.

A major limitation in performing functional genomic studies has been the cost and difficulty of obtaining RNA sequence data from focal cohorts and relevant tissues. In 2016, two independent labs published methodologies using regression regularization (GAMAZON *et al.* 2015) and BSLMM (GUSEV *et al.* 2016) to 1) create gene-based models to predict the genetic variants contributing to cis-regulatory control of gene expression, and then 2) use these models to impute gene expression values in an independent genome-wide genotyped dataset and 3) test for associations between the imputed gene expression and disease phenotypes. Because of their parallels to GWAS, this approach is referred to as a Transcriptome-Wide Association Studies, or TWAS.

The imputed gene expression values used in these studies were trained on large multi-tissue RNA-seq data sets for which GWAS genotypes were also available, such as the GTEx project. There are several advantages of the approach: 1) by aggregating SNPs into units that are known *a priori* to be functional, the method directly implicates *cis* eQTL's with nearby genes, 2) the multiple hypothesis burden is greatly reduced because traits are associated with genes not SNPs, 3) imputed expression values can be trained on different tissues and even cell types, 4) the association indicates if genes are up or down-regulated with disease, 5) no transcriptome data from a new study cohort is required.

1.5.4 TWAS using PrediXcan

The TWAS approach developed by GAMAZON *et al.* (2015) is called PrediXcan, and is the focus of this thesis. In their approach, all *cis*-acting common SNPs (MAF > 0.05) in HWE (p value > 0.05) from HapMap or 1000 Genomes reference genomic data were aggregated into sets for autosomal genes if the SNPs were within 1 Mb of the start (TSS) or end (stop codon) of a gene. SNP sets were then regressed against the expression of the gene in reference transcriptome data sets available from the Depression Genes and Networks (DGN) (BATTLE *et al.* 2014), which generated RNA-seq data from whole blood in 922 individuals, or from the GTEx database for which data from whole blood and 44 tissues (24 organs) was available for varying number of individuals (CONSORTIUM 2015). Regression analysis was performed on normalized gene expression values by letting

$$Y_g = w_{k,g} X_k + \varepsilon$$

where Y_g was the normalized expression trait for gene g, $w_{k,g}$ the effect size of variant k for gene g, X_k the dosage of variant k, and ε represented other random effects on gene expression (including environmental effects). GAMAZON *et al.* (2015) used LASSO and Elastic Net regularization to estimate the SNPs (X_k) and their weights ($w_{k,g}$) that best impute the genetic component of variation in gene expression using an additive model in which

$$\widehat{GReX}_g = \sum_k \hat{w}_{k.g} X_k$$

[Eqn 1] where $\widehat{GReX}_g$ is the genetically regulated component of gene expression. They propose that one could use the list of SNP's and their weights that they derive from these reference datasets to approximate the transcriptome of an individual from an independent genomic data set. At the PredictDB data repository (predictdb.org), databases containing the list of SNPs (by rsID) and their weights selected using Elastic Net and trained on the DGN/HapMap or GTEx/1000 Genomes paired datasets are available. By combining these values with the individual genotype dosages from an independent GWAS dataset one can impute the genetically controlled component of gene expression across individuals and then test for associations with any phenotype of interest.

GAMAZON *et al.* (2015) performed 10-fold cross-validation to estimate the reliability of $\widehat{GReX}_g$ based on the R^2 from the regression equations using SNPs and weights chosen using Elastic Net regularization. This R^2 serves as the upper bound of the heritability of the $GReX_g$. Using the DGN dataset, Gamazon *et al.*, (2015) compared their cross-validation R^2 to an independent estimate of the heritability of gene expression based on the approach of GCTA, (YANG *et al.* 2011a), showing a close overall agreement between the two estimates, albeit , with the GCTA estimates always somewhat higher (as expected) than R^2. The mean heritability of gene expression in the DGN blood data set based on GCTA was 0.153, while the mean R^2 using prediXcan was 0.137, and the range in h^2/R^2 values was from close to zero to over 0.90. Notably, estimates of the heritability of gene expression not only have a wide range within a tissue, but the same gene can have different heritabilities in different tissues because both gene expression and the SNPs predicted to control it differ among tissues.

1.5.5 The Genotype-Tissue Expression (GTEx) and Depression Genes and Networks (DGN) databases

GTEx is an NIH/Common Fund initiative whose mission is to provide a public resource for the systematic study of genetic variation and regulation of gene expression in multiple reference human tissues (CONSORTIUM 2015). The primary motivation for the initiative was to characterize the regulatory architecture of the human genome to facilitate the interpretation of significant findings from GWAS. They are collecting matched RNA sequence from multiple tissues and GWAS data from > 1000 individuals, characterizing gene expression across these individuals and tissues and identifying eQTL's. In the current version of the project, v7, RNA sequence and GWAS data has been collected from 714 donors each with up to 53 tissues (representing 24 organs) for a total of 11, 688 tissue samples. The donors were of any racial ethnic group or gender between the ages of 21-70 years, with some medical exclusions, and with their biospecimens collected within 24 hours after death. Nevertheless, there is limited ethnic diversity in the study with >85% of donors being white and 12.7% African-American. In addition, 66% of the donors are male, and almost 70% are > 50 years old. The number of tissues per donor varies widely, from 1 to 38 tissues per person with most individuals donating between 17-20 tissues; the most samples are available for muscle and skin (n>500), and the fewest for cervix (n=6). The number of genes with at least one significant eQTL is also highly variable among tissues, with thyroid having the most expressed genes (n=14,313) and the *substantia nigra* region of the brain having the fewest (n=1,807) (www.gtexportal.org, accessed June 30, 2018). GTEx genotype data was generated with the Illumina Omni Core 2.5 and 5M chips, and was used for impution to the 1000 genomes reference panel. RNA sequence data was generated with the Illumina HiSeq4000 (TrueSeq library, non-strand specific poly A+ selected (to capture mRNA fraction) and 76 bp paired end reads). Gamazon *et al.,* (2015) obtained the normalized RNA read counts from the GTEx consortium, and adjusted them for genetic ancestry, PEER factors, genotyping array platform (2.5 vs 5.0M) and sex prior to analyses using the PrediXcan methodology (GAMAZON *et al.* 2015).

GAMAZON *et al.* (2015) also used a paired RNA-Seq and genome-wide genotype data set from DGN (BATTLE *et al.* 2014). The DGN data set consisted of normalized RNA-Seq read counts from whole blood depleted of globin along with GWAS genotype data for 922 individuals of European ancestry from the NIH repository. The original genotyping data was generated with the Illumina HumanOmni1-Quad Bead Chip and processed for QC (BATTLE *et al.* 2014). GAMAZON *et al.* (2015) used ~650K SNPs (MAF>0.05, non-ambiguous strand (no A/T or C/G SNPs)) as input for imputation with both HapMap Phase II and 1000 G Phase1 v3 reference panels (see Table 1.1).

Table 1.1 Differences in the collection, processing and analysis of whole blood samples from the DGN and GTEx training sets.

Factor	DGN blood	GTEx blood
Sample Size	922	338
Participant Age	21-60	20-79
Participant Attributes	Living with or without depression	<24 Post-mortem Medical exclusions
RNA extraction	Whole blood with globin RNA reduction	Whole blood
Genotyping Platform	Illumina Human Omni1-Quad	Illumina 2.5 and 5.0M
RNA sequencing	HiSeq 2000, 50bp single-end reads	HiSeq4000 76 bp paired-end reads
Normalization	14 PEER factors	Genetic ancestry, genotyping array platform, PEER factors

1.6 Rationale for the Study

Previously, a GWAS was performed on the DCCT participants to look for genotypes associated with variation in HbA1c (PATERSON *et al.* 2010). This identified one GWAS hit and three additional associations that were close to genome-wide significance but had unknown functions. Here, I use the open-source PredictDB data repository, the DCCT genotype data set and the PrediXcan workflow to examine the relationship between mean log HbA1c and expression relevant SNPs in the DCCT dataset. The results of this analysis will be compared to the previous GWAS analysis to assess whether significant associations between imputed gene expression and HbA1c overlap with previously identified GWAS hits and/or whether new loci are identified. Such investigation could aid in the interpretation of the functions of the genetic variants identified in the previous

GWAS, tease apart the genetic determinants of hyperglycemia in T1D and ultimately help with therapeutic approaches to controlling blood sugar.

1.7 Objectives of the Thesis

The primary objective of the thesis was to investigate the associations between imputed gene expression and mean log HbA1c levels using 1304 DCCT individuals across selected tissues. Any significant associations will be compared to previously reported GWAS associations in the literature.

A secondary aim of the thesis was to explore differences among the different training sets available to perform transcriptome imputation.

1.8 Hypothesis of the Thesis

I hypothesis that variation in levels of imputed gene expression in blood and/or other tissues will be associated with variation in glycemic control as measured by mean HbA1c. Further, I expect that at least some of the genes found to be associated with variation in glycemic control, will be physically located near to previously identified GWAS variants influencing HbA1c. Secondarily, I hypothesis that for some genes, we may observe an interaction between imputed gene expression and treatment, because of the effect of treatment on mean HbA1c.

Chapter 2: Associations of imputed gene expression with HbA1c

2.1 Research Methods

2.1.1 Study Cohorts

The data used in this study were obtained from the Diabetes Complications and Control Trial (DCCT) and the follow-up study, EDIC (see Section 1.3.4). The eligibility criteria at baseline required that individuals were free of severe complications of T1D except retinopathy at the time of enrolment (Table 2.1), with the aim of examining whether the INT therapy could help prevent the development of retinopathy in individuals without retinopathy at baseline (primary prevention) or could slow the progression of retinopathy in individuals with mild retinopathy at baseline (secondary intervention) (Table 2.1). Thus, the important covariates to include in our analyses are: treatment (CONV vs INT), cohort (primary prevention or secondary intervention), age and duration of T1D at baseline, and gender. As noted, the INT therapy resulted in a significant decrease in mean HbA1c during the course of the trial and following its completion

Table 2.1 Eligibility Criteria at baseline

Age	13-39 years
HbA1c	>6.6%
Serum creatinine	≤ 1.2 mg/dl
Basal C-peptide	< 0.2 nmol/l
Hypertension	No
Hypercholesterolemia	No
Severe Medical Conditions	No

	Primary Prevention	Secondary Intervention
Diabetes duration	1-5 years	1-15 years
Retinopathy (fundus)	No	≥ 1 microaneurysm (level P2)
Albuminuria	<40mg/24 hr	< 200 mg/24 hr

all individuals were changed to intensive therapy and enrolled in the follow-up EDIC study. During the DCCT trial years (1983-1993), measurements of HbA1c were taken quarterly for individuals in CONV therapy, but monthly for individuals in INT therapy, and annually during EDIC for both groups (data for 1994-2009 included here). Thus, the maximum number of HbA1c measurements for individuals in INT therapy is greater than for those originally enrolled in CONV therapy (Table 2.2). The other covariates were quite evenly distributed between therapy groups (Table 2.2).

Table 2.2 Summary statistics of the covariates in the DCCT study. Continuous variables given with standard deviation (SD).

		Conventional	Intensive
Gender	Male	363	332
	Female	304	305
Cohort	1° Prevention	344	307
	2° Intervention	323	330
Age (years ±SD)	DCCT baseline	26.5 ± 7.1	27.2 ± 7.1
Diabetes Duration (months ± SD)	DCCT baseline	66 ± 49	69.75 ± 50
HbA1C measurements^			
DCCT	1983-1993	40	108
EDIC	1994-2009	16	16
TOTAL	1983-2016	62	130

^ Maximum number of HbA1C measurements during DCCT and EDIC

2.1.2 HbA1c

During the study, blood samples were collected from the participants and stored at 4°C until measured for HbA1c content using high-performance liquid-chromatography (HPLC) (1987). HbA1c was measured monthly in the INT group and quarterly in CONV group for up to 10 years and annually for 22 years in both groups (Table 2.2). All of the HbA1c measurements after the third month were used for analysis. To facilitate analyses, the within-person mean HbA1c levels (meanHbA1c) were calculated for each patient, and the logarithm of these values were used as the response variable (mean $\log_{10}$ HbA1c) in our study because the original distribution was skewed (see results).

2.1.3 Genotyping and Imputation

DNA was previously collected from the DCCT patients with their written informed consent and used for genome-wide genotyping with the Human Illumina 1M-duo v3.0 BeadChip (Illumina Inc., San Diego, CA, USA) (PATERSON *et al.* 2010). The Illumina Human1M beadchip is based on the HapMap II reference data set and their SNPs tagged 93% of the common SNPs found in the reference European population (CEU - Utah residents with Northern and Western European ancestry from the CEPH collection) at $R^2 \geq 0.8$. Genotypes were called using the BeadStudio/GenomeStudio software (Illumina Inc., San Diego, CA, USA) and thereafter analyzed using PLINK v1.07 for quality control (QC) (PURCELL *et al.* 2007). All samples passed the cutoff for genotype call rate; however, samples with gender discrepancies, as well as cryptic relatedness and outliers of European ancestral population structure as detected by principal component analyses (PCA) were removed following the QC procedures in (PATERSON *et al.* 2010). The genotype concordance of 24 duplicate samples was 99.9995% (at a call rate threshold of 0.988) and the mean heterozygosity across the genome for each individual fell within normal expectations (0.25-0.32). Autosomal SNPs were excluded from analyses if they deviated from HWE ($p < 10^{-8}$) or showed significant association with gender (PATERSON *et al.* 2010). 841,342 SNPs were retained following these QC procedures.

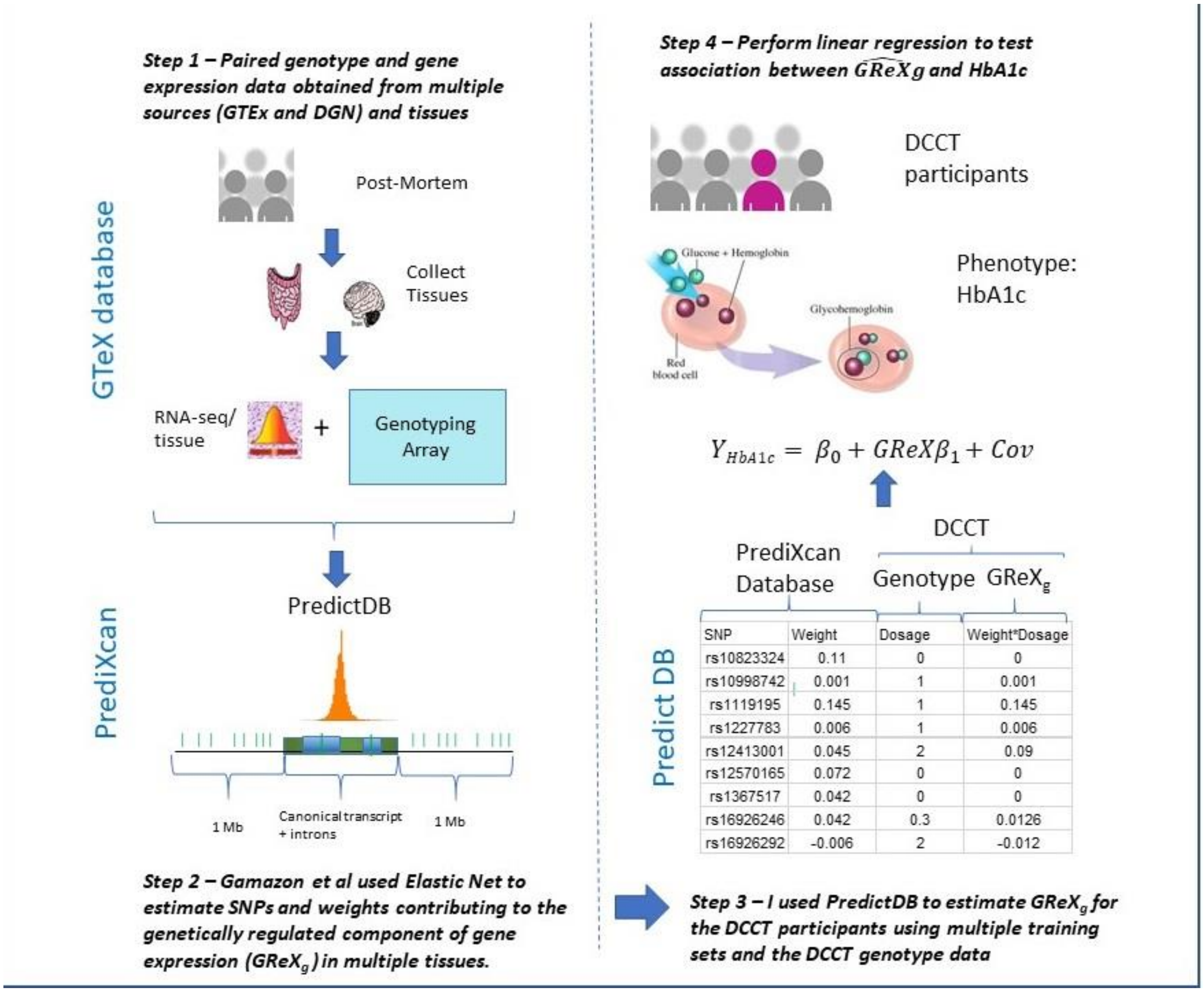
More SNPs were imputed using the Impute v2 software (HOWIE *et al.* (2009) with the reference from the 1000 Genomes (phase 1 version 3) integrated variant release

(March 2012). Only individuals who clustered with CEU and TSI (Toscani in Italy) from phase III of the International HapMap Project based on the PC analysis, and only SNPs that were imputed with high certainty (INFO ≥ 0.8; IMPUTE version 2) were included. In total 1304 of the original 1441 individuals and more than 8 million genotyped and/or imputed SNPs were available for analyses.

2.1.4. Training Sets for estimation of imputed gene expression in DCCT cohort

PredictDB training sets. An overview of the workflow that I implemented to perform the PrediXcan method is given in Figure 2.1. The first two steps of the procedure were done by Gamazon et al (2015): they obtained training sets of matched transcriptome and genotype data from blood (the DGN dataset) or from 44 tissues including blood (the GTEx consortium) from publicly available data repositories (Step 1, described in section 1.5.5) and then used Elastic Net regularization to select the set of SNPs and estimated their weight that best predicted the genetically regulated component of gene expression, $GReX_g$, for each gene (Step 2, Section 1.5.4). The first step of my analyses (Step 3 in Figure 2.1) was then to retrieve the list of SNPs and their weights developed using blood from the DGN dataset or using blood and eight other tissues from the GTEx datasets. I retrieved these ten training sets from the PredictDB data repository (http://hakyimlab.org/predictdb/) in March and August 2017 based on the GTEx V6P data (www.gtexportal.org) which were available on November 15th, 2016. The SNPs and weights were obtained for each of the following ten training sets with their sample sizes: GTEx Ileum (n=77), GTEx liver (n=97), GTEx pancreas (n=149), GTEx Skeletal Muscle (n=361), GTEx Spleen (n=89), GTEx Thyroid (n=278), GTEx Tibial Nerve (n=256), GTEx

Visceral Fat (n=185), GTEx whole blood (n=338) and DGN whole blood (n=922). These datasets were haphazardly chosen because of their potential involvement in T1D (e.g. tissues involved in metabolism or immune functions such as small intestine, spleen, liver, visceral fat) or because of the higher number of eQTLs available in the dataset (e.g. tibial nerve, thyroid). Since some genes are expressed in multiple tissues, different predictors for the same gene will occur in different training sets. Thus an expression- associated SNP may be included as a predictor for a gene in some but not all training sets. Owing to this, associations between imputed gene expression and a disease or trait of interest may be found using training sets not trained on the disease (or trait)-associated tissues *per se*, because by chance, genetic variants may be associated with the disease/trait in non-focal tissues. The ten databases were extracted using the DB Browser for SQLite Version 3.9.2 and read into R studio Version 1.0.16 running R version 3.3.2 (2016-10-31) and 3.5.1 (2018-07-02) for descriptive analyses.


Step 1 – Paired genotype and gene expression data obtained from multiple sources (GTEx and DGN) and tissues
GTeX database
Post-Mortem
Collect Tissues
RNA-seq/ tissue
+
Genotyping Array
PredictDB
PrediXcan
1 Mb
Canonical transcript + introns
1 Mb
Step 2 – Gamazon et al used Elastic Net to estimate SNPs and weights contributing to the genetically regulated component of gene expression ($GReX_g$) in multiple tissues.
Step 4 – Perform linear regression to test association between $\widehat{GReX}g$ and HbA1c
DCCT participants
Glucose + Hemoglobin
Glycohemoglobin
Red blood cell
Phenotype: HbA1c
$Y_{HbA1c} = \beta_0 + GReX\beta_1 + Cov$
DCCT
PrediXcan Database
Genotype
$GReX_g$
Predict DB
Step 3 – I used PredictDB to estimate $GReX_g$ for the DCCT participants using multiple training sets and the DCCT genotype data

SNP	Weight	Dosage	Weight*Dosage
rs10823324	0.11	0	0
rs10998742	0.001	1	0.001
rs1119195	0.145	1	0.145
rs1227783	0.006	1	0.006
rs12413001	0.045	2	0.09
rs12570165	0.072	0	0
rs1367517	0.042	0	0
rs16926246	0.042	0.3	0.0126
rs16926292	-0.006	2	-0.012

Figure 2.1 Overview of the workflow implemented in this thesis to test for associations between imputed gene expression ($GReX_g$) and HbA1c in individuals enrolled in the DCCT trial using the "PrediXcan" database resources developed by Gamazon *et al.* (2015). Steps 1 and 2 (left panel) were used by Gamazon *et al.* to develop their method for gene expression imputation, and steps 3 and 4 were conducted by myself to apply the method to the DCCT genotype and phenotype data.

2.1.4 Estimating GReX for tissue specific gene sets in the DCCT study participants.

After obtaining the SNPs (given by rsID in the PredictDB database, Figure 2.1) for each tissue, I extracted the SNPs overlapping with the markers from the DCCT genotype data and calculated the dosage (number of minor alleles for each SNP) from the 1304 Caucasian participants (Step 3, Figure 2.1). The dosages were used to impute the genetically regulated gene expression, $\widehat{GReX}_g$, for all genes/individuals in each of the ten tissues using the script given in the appendix (B.1). The estimated genetically regulated gene expression was calculated as:

$$\widehat{GReX}_{g,T} = \sum_{k \in S_{g,T}} \widehat{\omega}_{k,g,T} \, \mathrm{X}_k$$

[eqn 2] $S_{g,T}$: set of SNPs in the expression predictors for gene g in the training set or tissue T, $\widehat{\omega}_{k,g,T}$: the estimated weight for variant k from g and T, and X_k, the dosage of SNP, k. Genotypes were extracted in terms of the PredictDB reference and effect alleles to ensure the direction of the effect was correct. The imputed expression was estimated using a script obtained from the PrediXcan website, and run using shell commands listed in Appendix B. The output of this step is a file containing the predicted genetically regulated component of gene expression, $GReX_g$, for every gene found in that training set per individual.

2.1.5. Association of $GReX_g$ with HbA1c.

The next step (step 4, Figure 2.1) was to perform association analyses, using the within person mean $\log_{10}$ HbA1c as the outcome. Then the contribution of $\widehat{GReX}_g$ to HbA1c was assessed using the linear regression model

$$Y_{HbA1c} = \beta_{0,g,T} + \beta_{1,g,T}\widehat{GReX}_{g,T} + \gamma_1 C_1 + \cdots + \gamma_6 C_6 + \epsilon_{g,T}$$

[eqn3] with covariates age (C_1), duration of T1D (C_2), gender (C_3), primary or secondary cohort (C_4), conventional versus intensive treatment (C_5), and the interaction between

cohort and treatment (C_6). The significance of $\widehat{GReX}_g$ was determined using the student's t-distribution with 1296 degrees of freedom (=1304-7-1). The mean $\log_{10}$ HbA1c was also regressed against the covariates in the model

$$Y_{HbA1c} = \gamma_0 + \gamma_1 C_1 + \cdots + \gamma_6 C_6 + \epsilon$$

[eqn4] to assess the proportion of variance in HbA1c that can be explained by the covariates alone. For those genes that showed a significant or suggestive association between $\widehat{GReX}_g$ and HbA1c, an additional model was run to test the interaction between

$\widehat{\mathrm{GReX}}_{\mathrm{g}}$ and treatment (C_7):

$$Y_{HbA1c} = \beta_{0,g,T} + \beta_{1,g,T}\widehat{GReX}_{g,T} + \gamma_1 C_1 + \cdots + \gamma_6 C_6 + \beta_{2,g,T}\widehat{GReX}_{g,T} * C_5 + \epsilon_{g,T}$$

[eqn5]. Kraft et al (2007) demonstrate that an efficient approach for screening for the association between a large number of markers and disease is to perform a two-degree of freedom joint test. In this test, a model including just covariates is compared to a model that includes both the marginal genetic effects and the gene*environment interaction (i.e. eqn 5). These two models are compared using a likelihood ratio test (LRT) of the form

$$2\log\left[L\left(\hat{\beta}_1\right) - L(\left(\hat{\beta}_0\right)\right]$$

[eqn6], where $\hat{\beta}_1$maximizes the likelihood under the alternative hypothesis, and $\hat{\beta}_0$ maximizes it under the constrained null hypothesis (no interaction effect). The statistic is asymptotically chi-squared distributed with two degrees of freedom (the difference in the number of parameters between $\hat{\beta}_0$ and $\hat{\beta}_1$); and tests whether there were significant

associations between the imputed gene expression and HbA1c and whether the associations were mediated by a greater effect in the CONV or INT arms of the DCCT trial.

For all of the linear regression analyses, the distribution of error in the models (ϵ) was assumed to be independently and normally distributed with E(ϵ) = 0 and Var (ϵ) $= \sigma^2$. For the binary covariates, the following categories were included: gender – 1 for male, 2 for female, cohort – 0 for primary prevention and 1 for secondary intervention, and treatment – 0 for conventional and 1 for intensive. The linear models were fit using the lm function in R version 3.5.1 (see script in Appendix). Studentized residuals were used to check the linearity and normality assumptions for the linear regression models with covariates only and for the models including the top hits for $\widehat{GReX}_g$.

A Bonferonni correction significance threshold of α = 0.05/total number of genes across all training sets (N_{total}=58,102, see results Table 2.5) was calculated, and a less stringent threshold of α = 0.05/number of genes in the focal training set (e.g. N_{Ileum}=2,728 to N_{DGN}=11,403, see results Table 2.5) was considered suggestive. Since no genes reached the suggestive significance level, the top genes for which p > 1xE-5 and <0.0001 were explored for possible associations with HbA1c. Manhattan plots were used to visualize suggestive p-values across the genome, and histograms to observe the distribution of p-values.

2.2 Results

2.2.1 Relationship between HbA1c and covariates

The effect of the covariates on mean $\log_{10}$ HbA1c was first examined. Multiple linear regression indicated that a large difference in mean $\log_{10}$ HbA1c among individuals was attributable to the treatment effect (CONV versus INT); the individuals in the INT therapy group had, on average, 6.7% lower (on log scale) mean $\log_{10}$ HbA1c compared to individuals in the CONV therapy group over the duration of the study period between 1983 and 1993 (Table 2.3). Additionally, the age of individuals and their duration of T1D at baseline were significantly associated with mean $\log_{10}$ HbA1c: for every year older that individuals were at baseline, they had a 0.132% lower mean $\log_{10}$ HbA1c, and for every additional year that they had been diagnosed with T1D at baseline, they had a 0.028% decrease in mean $\log_{10}$ HbA1c (Table 2.3). The regression analyses met the assumptions of normally distributed residuals, and no points were deemed to be significant outliers (Figure 2.2). The large effect of treatment on HbA1c was also evidenced by differences in the distribution of HbA1c in the CONV and INT therapy cohorts on both normal and log scales, such that the range of HbA1c values of individuals in the CONV therapy was wider than the range of those on the INT therapy (Table 2.4, Figure 2.2).

Table 2.3 Multiple linear regression results for the model of mean $\log_{10}$ HbA1c ~ Age + Gender + Duration at baseline + cohort + treatment + treatment * Cohort.

Coefficients	Estimate	SE	T	P-value
Age (years)	-1.32E-3	4.75E-4	-2.79	0.00528
Gender	1.15E-2	6.77E-3	1.69	0.0897
Duration (months)	-2.84E-4	9.98E-5	-2.85	0.0045
Cohort	9.76E-3	1.19E-2	0.82	0.412
Treatment	-6.72E-2	9.54E-3	-7.04	3.05E-12
Cohort *Treatment	4.07E-4	1.35E-2	0.03	0.976

R^2_{model}= 0.088, R^2_{adj}=0.084, *F* = 20.91, *d.f.* = 1297

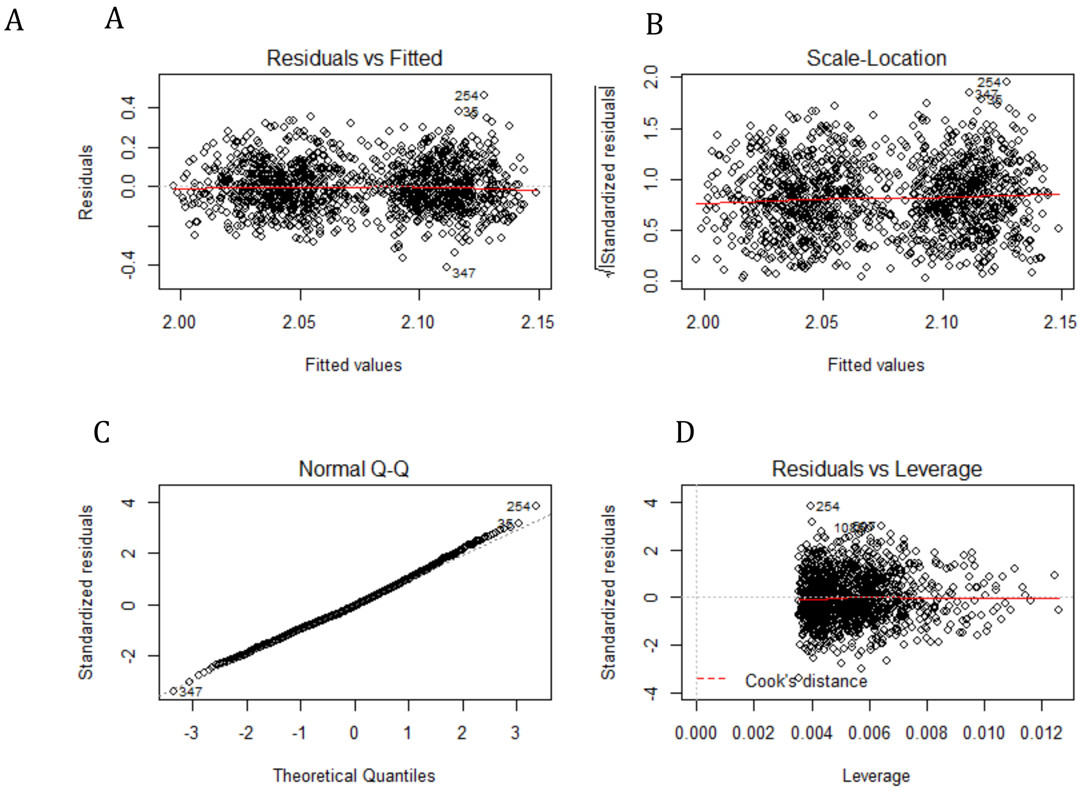


Figure 2.2 Diagnostic plots for the multiple linear regression presented in Table 2.3 of mean log10 HbA1c against the covariates.

(A) Residuals against fitted values, (B) Square root of the standardized residuals against fitted values, (C) normal probability plot standardized residuals, (D) Leverage values. The two clusters in the scatterplots of residuals vs fitted values (A) and scale-location (B) represent differences between treatment arms (INT vs CONV). The one individual (254) that had high leverage (D) did not exceed the threshold value.

Table 2.4 Summary statistics for the mean $\log_{10}$ HbA1c and mean HbA1c of individuals in the CONV and INT therapies during the DCCT trial.

Mean $\log_{10}$ HbA1C

Treatment Group	N	Mean	St. Dev	Min	Max
Conventional	667	2.11	0.123	1.70	2.59
Intensive	637	2.04	0.120	1.76	2.41

Mean HbA1c

Treatment Group	N	Mean	St. Dev	Min	Max
Conventional	667	8.373	1.05	5.33	13.36
Intensive	637	7.802	0.98	5.83	11.25

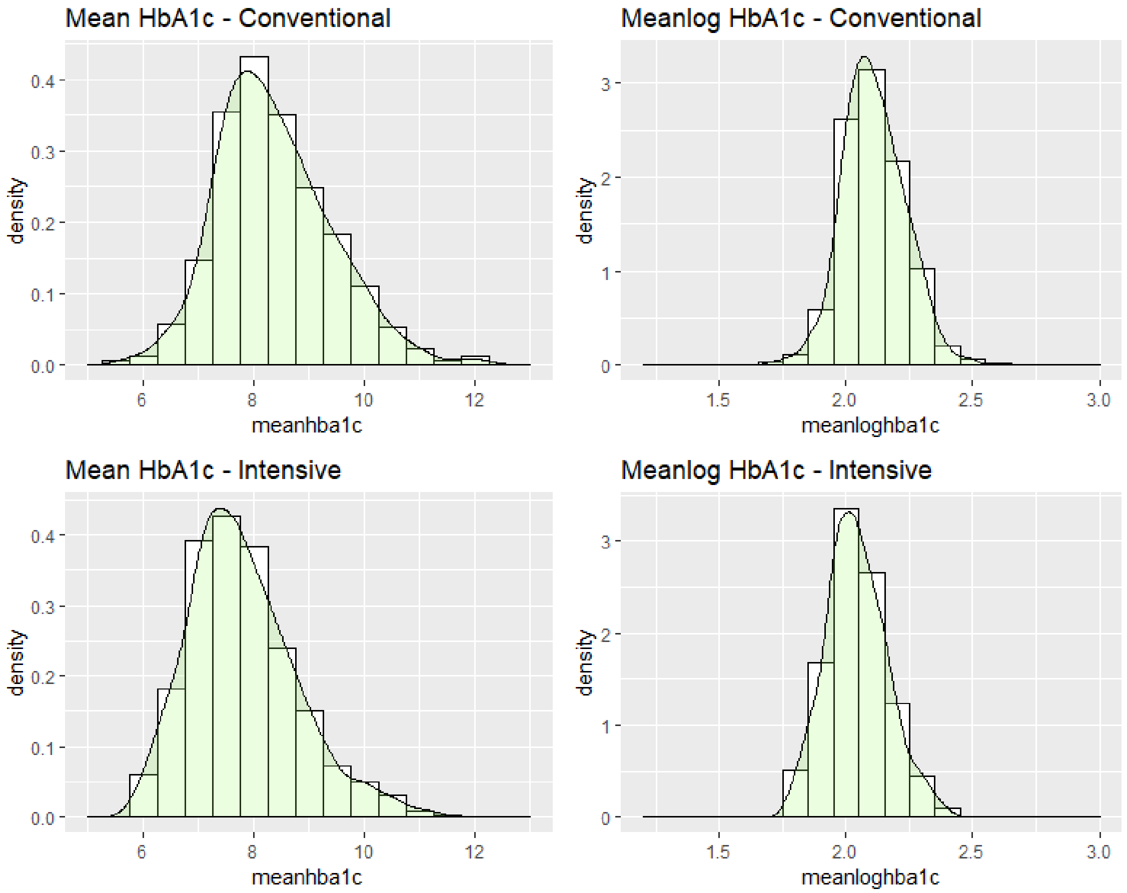


Figure 2.3 Distribution of the mean HbA1c and mean $\log_{10}$ HbA1c of individuals in the CONV and INT therapies during the DCCT trial.

2.2.2 Characteristics of PrediXcan training sets and their application to the DCCT genotype data

The number of genes represented in each of the ten training sets downloaded from the PrediXcan database repository differs because a) tissues/cells express different numbers of genes and b) not all genes were found to have a significant genetically regulated component of expression based on cross-validation estimates of the model R^{2} - and were thus not included in the database (Gamazon *et al.*, 2015). The training set with the most genes was that based on blood from DGN, for which $\widehat{GReX}_g$was estimated for 11,538 genes using 331,425 SNPs (Table 2.5). On the other hand, all of the nine training sets derived from the GTEx data impute $\widehat{GReX}_g$ from fewer genes than the data from DGN, which may be partially related to the fewer number of individuals used in the GTEx training sets (Table 2.5), but may also be related to differences in processing and analysis of the original datasets (Table 1.1). Of the nine training sets derived from GTEx data, estimates of $\widehat{GReX}_g$ were available for between 2,732 (small intestine) to 8,171 (Tibial Nerve) genes, with correspondingly fewer SNPs required to impute $\widehat{GReX}_g$for smaller training sets (Table 2.5). The training set based on GTEx whole blood employed 196,491 SNPs to estimate the $\widehat{GReX}_g$for 6,759 genes, almost 5,000 fewer genes than estimated in blood from the DGN training set (Table 2.5). Following extraction of the genotype data from the DCCT datasets, I estimated $\widehat{GReX}_g$ in the DCCT dataset for almost all of the genes for which prediction were available (last column, Table 2.5).

Table 2.5 Summary of the PrediXcan database characteristics for the ten training sets employed in this thesis.
For each of the ten training sets SNPs and their weights were downloaded from the PrediXcan database repository and were used to impute gene expression ($\widehat{GReX}_g$). Using the genotype data from the DCCT trial, the estimates of $\widehat{GReX}_g$(last column) were then employed in the TWAS for the relationships between $\widehat{GReX}_g$ and HbA1c.

Training Set	**Imputation**	**Tissue**	**N** training	**No. SNPs Training Set**	**No. Genes Training Set**	**No. Genes Imputed in DCCT**
DGN	HapMap	DGN – Blood	922	331,425	11,538	11,403
GTEx	1KG-ph3	Small intestine	77	112,911	2,732	2,728
		Liver	97	108,920	2,913	2,910
		Spleen	89	136,661	3,703	3,696
		Visceral Fat	185	141,838	4,544	4,539
		Pancreas	149	154,298	4,775	4,773
		Skeletal Muscle	361	191,133	6,627	5,926
		Whole Blood	338	196,491	6,759	6,752
		Thyroid	278	241,824	8,149	7,210
		Tibial Nerve	256	253,968	8,171	8,165
						Total: 58,102

A cursory examination of the characteristics of the training sets will be helpful for interpretation of the results. As noted in the introduction, Gamazon *et al.* (2015) employed cross-validation to estimate the reliability (as estimated by R^2) of their gene expression models. These R^2 values serve as an estimate of the upper limit of the heritability of $\widehat{GReX}_g$, and indicate the proportion of variance in gene expression that is attributable only to the cis-acting genetic variants (SNPs) included in the model. The density distribution of the prediction R^2 values for six of the ten training sets (the remaining four were similar) shows that R^2 is close to zero for most estimates of $\widehat{GReX}_g$, and only a few genes have an R^2 value > 0.2 (Figure 2.4). Similarly, histograms of the number of SNPs included in each gene model indicate that the average number of SNPs per gene is ~30 across all tissues (red-dotted line); most gene models have fewer than 30 SNPs, but the distribution is left-skewed, and a few genes have more than 100 SNPs in the gene model (Figure 2.5). Lastly, the density distribution of the weights for all SNPs included in the DGN and GTEx training sets shows that the density is essentially symmetric around zero, with most values falling between - 0.25 and + 0.25 (Figure 2.6). Values >0 indicate a positive association of the contribution of the minor allele of a SNP to gene expression, while values <0 indicate that the alternate allele decreased gene expression.

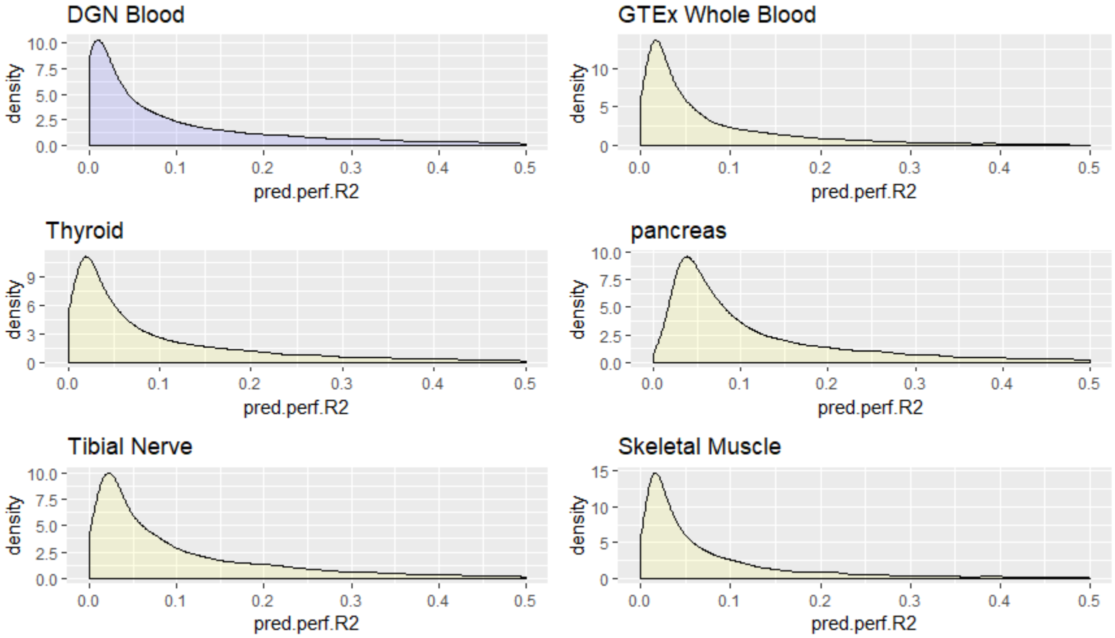


Figure 2.4. Density distribution of prediction R^2 values for six training sets. Prediction R^2 values were obtained from the PrediXcan database (values on the X-axis were limited to be < 0.5, above which the density was very low).

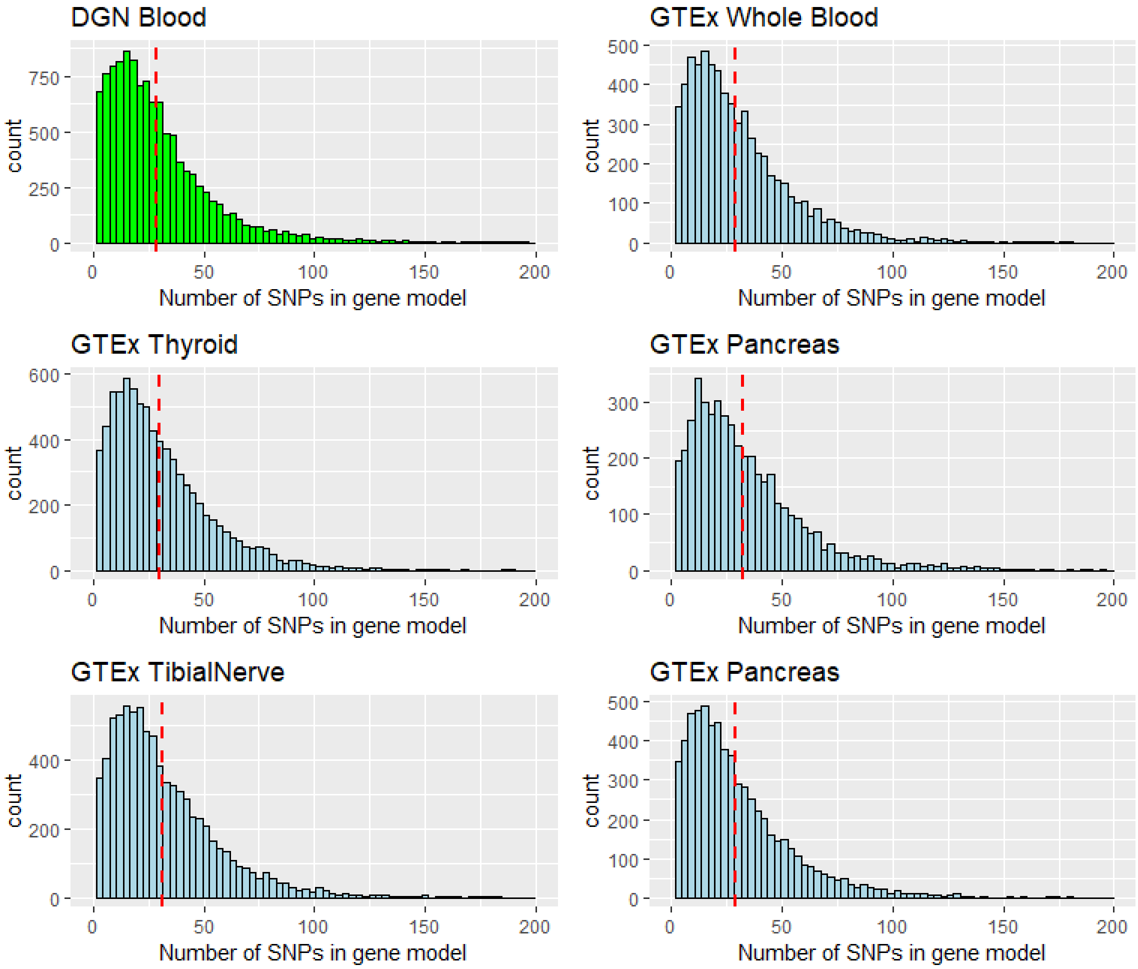


Figure 2.5 Histogram of the number of SNPs used to impute $\widehat{GReX}_g$ for each gene across six of the training sets.

The red-dotted line indicates the mean number of SNPs per gene model, which is close to 30 for all of the tissues presented. The x-axis was constrained to be lie between 0 and 200, a few tissues had genes with > 200 SNPs per gene model. The remaining four tissues not shown here had similar distributions.

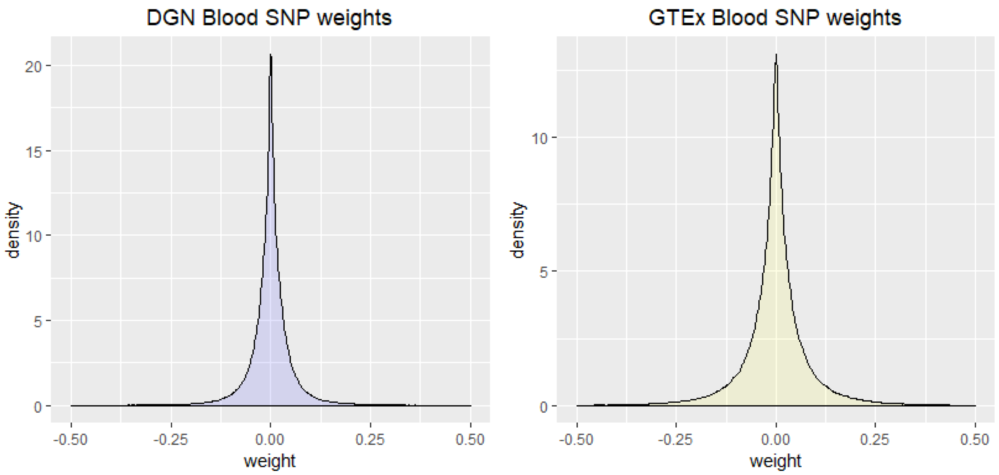


Figure 2.6. Density distribution of the SNP weights for the training sets based on DGN blood (left panel) and GTEx blood (right panel).
Values were obtained from the PrediXcan database (N=331,425 for DGN blood, and N=196,491 for GTEx blood, see Table 2.5).

As noted in the Introduction, there were numerous differences in how the whole blood samples were processed and analysed in DGN versus GTEx (Table 1.1); thus a comparison of the two training sets provides evidence of the robustness of the PrediXcan method to estimate the genetically regulated component of gene expression, $\widehat{GReX}_g$. The Spearman's correlation between the R^2 values in the gene prediction models for co-occurring genes in the DGN and GTEx blood training sets was strong (r=0.65; tested with the Spearman's rank correlation rho S = 1.026 e+10, p-value < 2.2e-16, Figure 2.6), indicating that the same genes were found to have similar heritability's in the two datasets, albeit the estimates based on the DGN dataset were consistently higher.

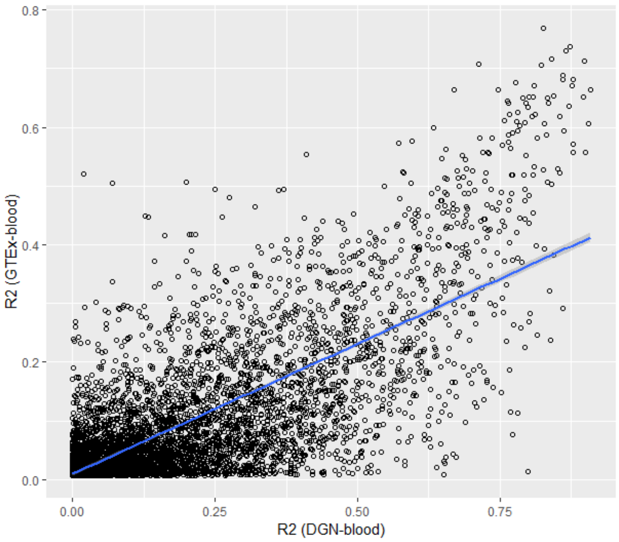


Figure 2.7. Correlation of R^2 prediction values for co-occurring genes in the DGN and GTEx blood training sets.
In total 5584 genes were shared between the two training sets; the Pearson correlation coefficient was r=0.74 (blue line is the best linear model through all points).

The list of SNPs used to impute gene expression in each of the ten training sets was used to extract genotype and dosage information from the DCCT dataset (Step 2, Figure 2.1), in order to impute gene expression in DCCT (Step 3, Figure 2.1). Although the reference and alternate allele frequencies of the overlapping SNPs in the DGN and GTEx training sets were not available from the PredictDB website, I extracted allele frequencies of the SNPs in the DCCT dataset, and plotted the frequency of the alternate allele, the one associated with the SNP effect (Figure 2.8). This revealed that the SNPs used to estimate imputed gene expression based on the DGN-blood training set (left panel) were more common, in general, than those based on the GTEx-blood training set (right panel). This is consistent with the fact that the DGN training set was based on the HapMap SNPs, while those from the GTEx training sets are based on the 1000G genotypes (Table 2.3), since the HapMap project characterized primarily common variants (see the Introduction).

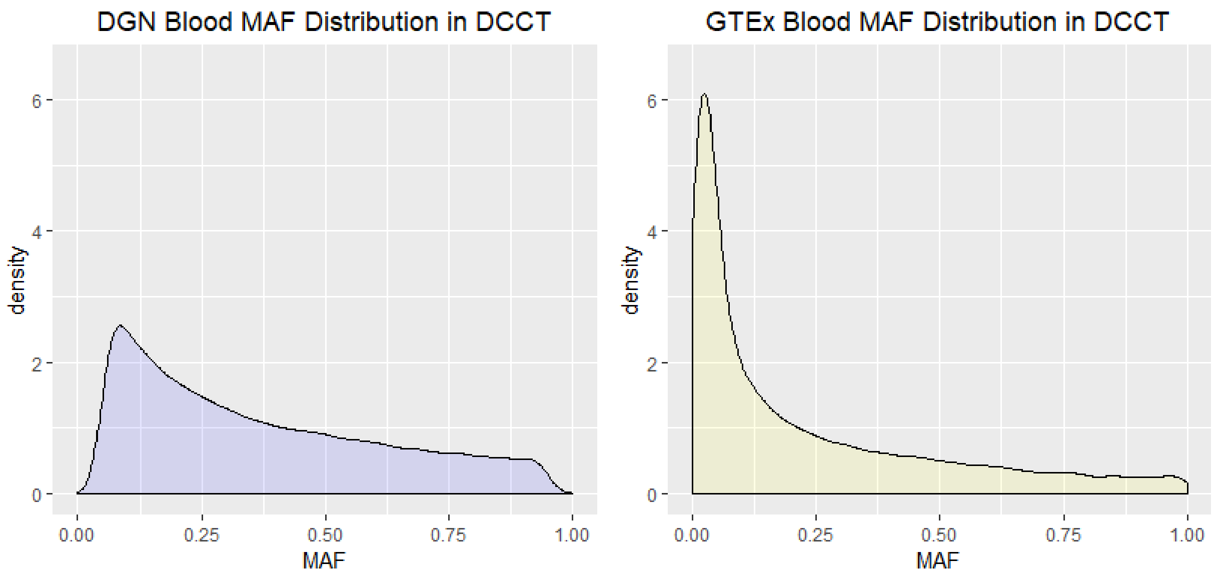
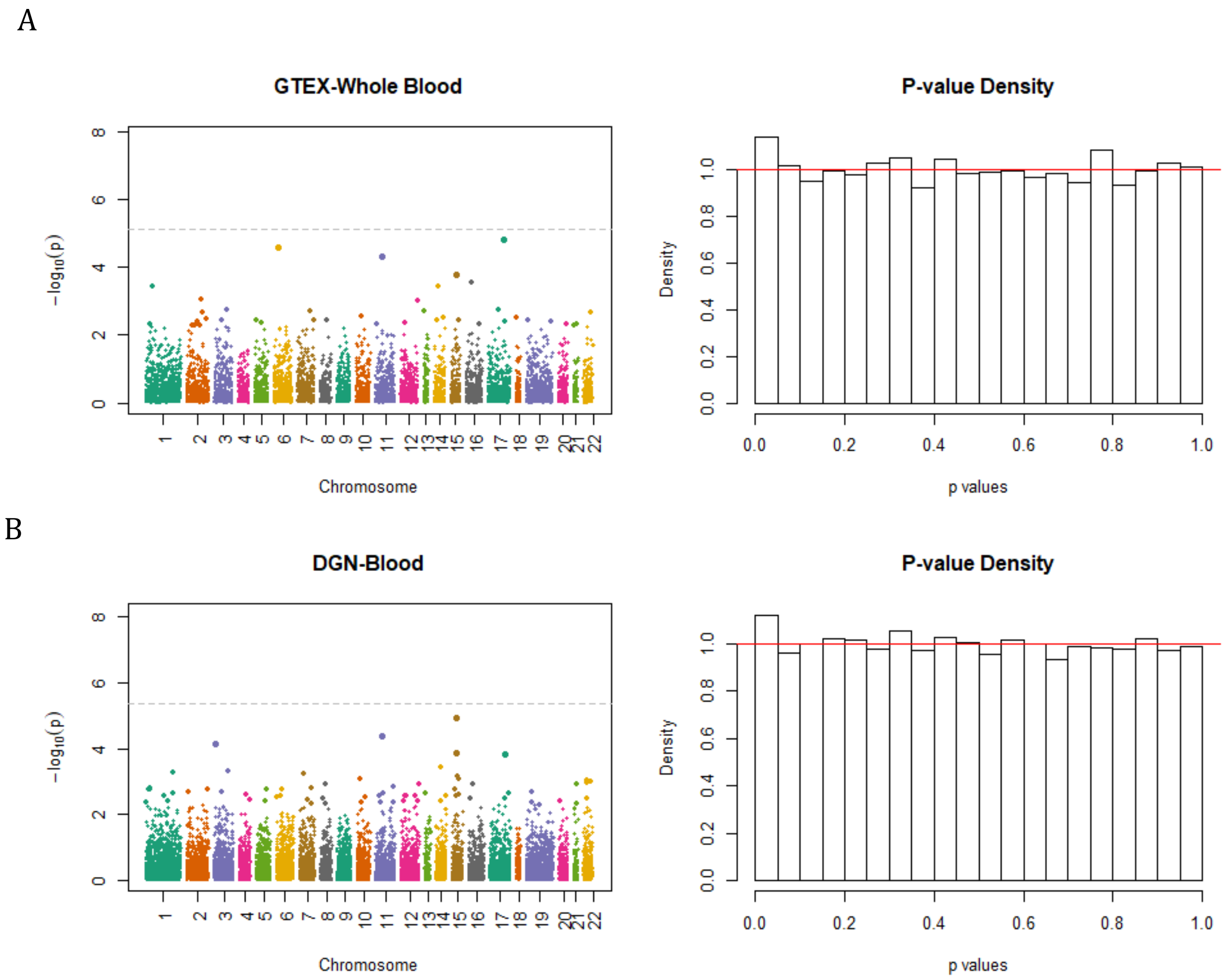


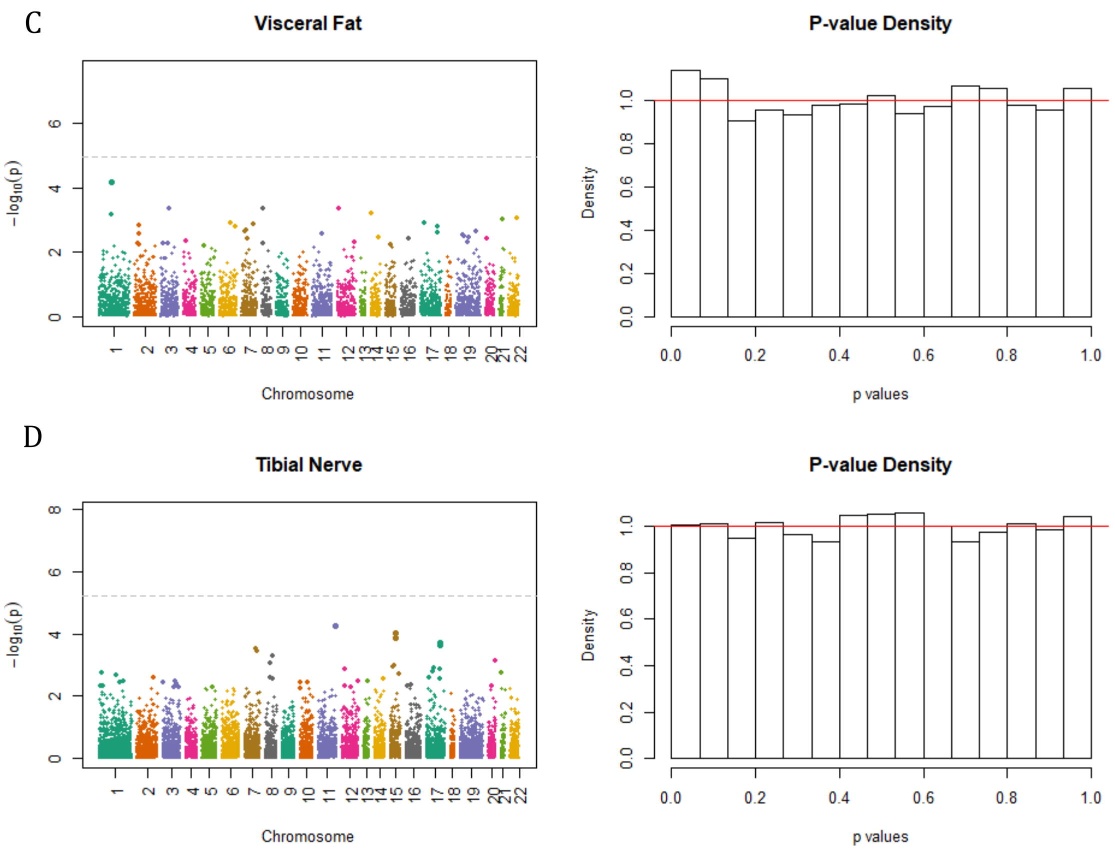
Figure 2.8. Density distribution of the alternate allele frequency of SNPs in 1304 individuals enrolled in DCCT that were used to impute gene expression based on DGN blood (left panel) and GTEx blood (right panel) training sets.
Values were obtained from the DCCT data for all those SNPs for which genotypes were available (289,667 SNPs were extracted based on the SNPs in the DGN blood training set and 182,291 for those in the GTEx blood training set). Note that the number of SNPs extracted is fewer than the maximum possible (Table 2.5) because genotype data was not available at all sites.

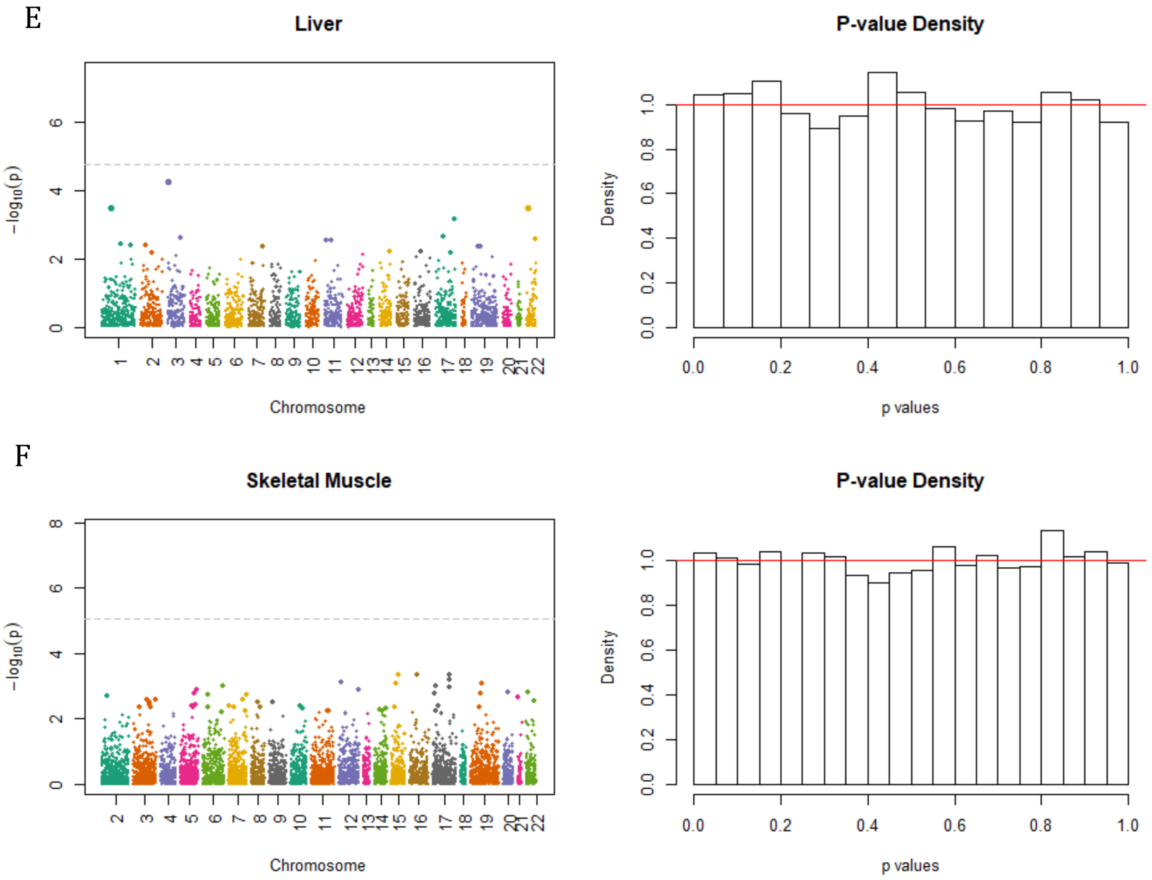
2.2.3 Association between $\widehat{GReX}_g$ and HbA1c in the DCCT dataset

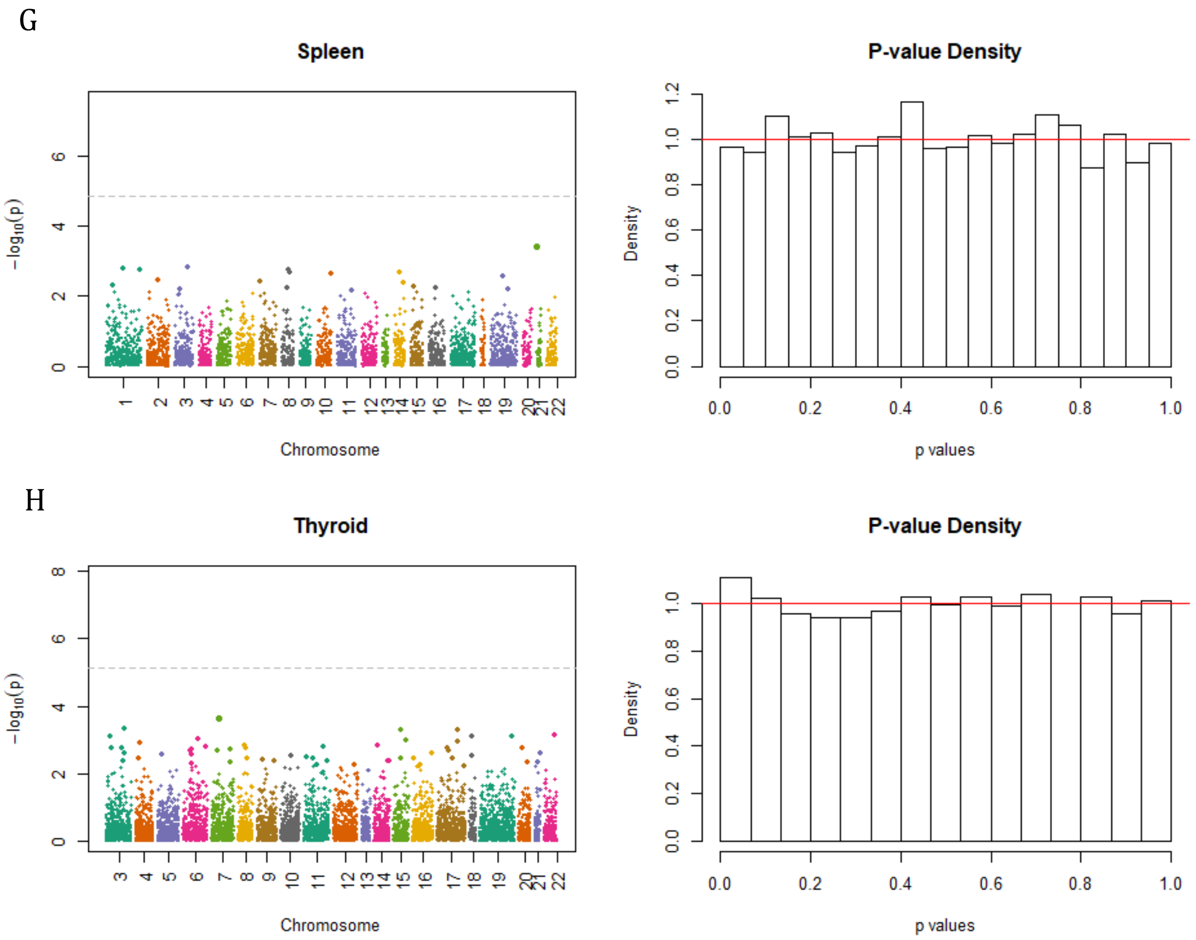
Linear regression analysis was performed to test for associations between imputed gene expression, $\widehat{GReX}_g$, and mean $\log_{10}$ HbA1c (eqn 3) in the 1304 individuals in DCCT in all ten tissues. Given the large number of tests, the results are visualized by Manhattan plots (Figure 2.9a-j) depicting the p-value for the association and the chromosomal position of the gene. In total, 58,102 genes were tested for association (Table 2.5), such that a strict Bonferroni correction p value would be: 0.05/58,102 = 8.6 E-7. Given that many of the same genes occur in different tissues and are not independent, this criterion is very conservative. In any case, no genes reached this global threshold. Thus, for each tissue, a less stringent Bonferroni correction was employed based on the number of genes analysed in that tissue (listed in Table 2.5) and is indicated by a dotted line on each Manhattan plot (Figures 2.9A-J). The p-value density

histograms were approximately uniform in all cases as expected (Figures 2.9A-J).




C
Visceral Fat
−log10(p)
0
2
4
6
Chromosome
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
P-value Density
Density
0.0
0.2
0.4
0.6
0.8
1.0
p values
0.0 0.2 0.4 0.6 0.8 1.0
D
Tibial Nerve
−log10(p)
0
2
4
6
8
Chromosome
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
P-value Density
Density
0.0
0.2
0.4
0.6
0.8
1.0
p values
0.0 0.2 0.4 0.6 0.8 1.0


E
Liver
−log10(p)
0
2
4
6
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome
P-value Density
Density
0.0
0.2
0.4
0.6
0.8
1.0
p values
F
Skeletal Muscle
−log10(p)
0
2
4
6
8
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome
P-value Density
Density
0.0
0.2
0.4
0.6
0.8
1.0
p values


G
Spleen
−log10(p)
0
2
4
6
Chromosome
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
P-value Density
Density
0.0
0.2
0.4
0.6
0.8
1.0
1.2
p values
0.0
0.2
0.4
0.6
0.8
1.0
H
Thyroid
−log10(p)
0
2
4
6
8
Chromosome
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
P-value Density
Density
0.0
0.2
0.4
0.6
0.8
1.0
p values
0.0
0.2
0.4
0.6
0.8
1.0

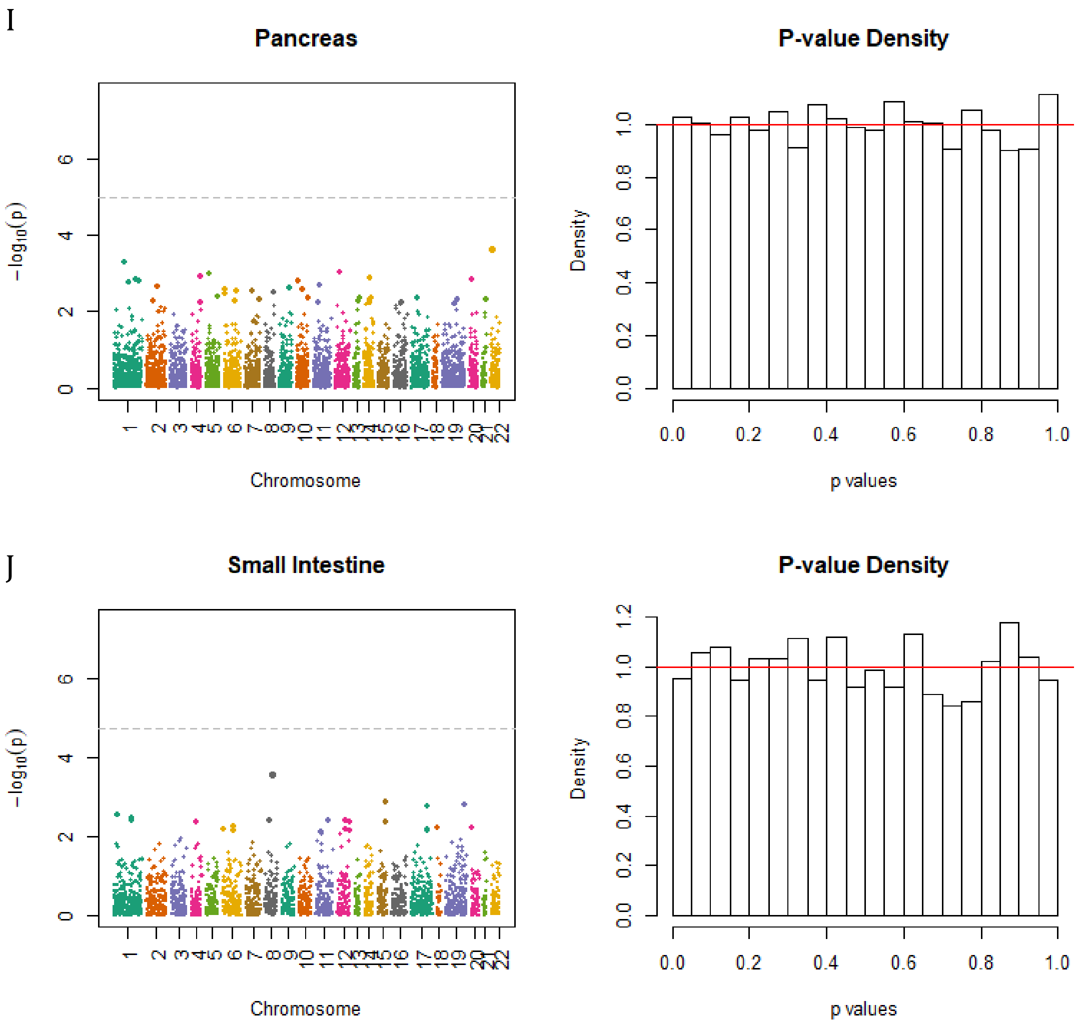


Figure 2.9. Manhattan Plots (left panels) and p-value density (right panels) for the association between mean log10 HbA1c and GReX adjusted for covariates.
The gray dotted line is for a suggestive Bonferroni correction p value in the focal tissue. A: DGN-blood (N=11,403), and from GTeX (B-J) B: blood (N=6752);C: Visceral Fat (N=4539); D: Tibial Nerve (N=8165); E: liver (N=2910); F: Skeletal Muscle (N=5926); G: Spleen (N=3696); H: Thyroid (N=7210); I: Pancreas (N=4776); J: Small Intestine (N=2728).

Although no associations of imputed gene expression with mean $\log_{10}$ HbA1c exceeded the suggestive significance level (Figure 2.9), nine different genes had p-values <0.0001 and this threshold was arbitrarily used to consider genes worthy of further analyses (Table 2.6). Six of the hits were identified based on the training sets from the DGN-blood or GTEx –blood (one gene, MAPK8IP1 was identified in both training sets), two were found based on the training set for tibial nerve, and one liver and visceral fat (Table 2.6). All but one (*SMG8*) of the genes had a positive association with gene expression meaning that the net effect of SNPs controlling imputed gene expression was positive (β1>0.0). For all of the genes except for *MAPK8IP1* in the DGN training set, the R^2 (heritability) of the imputed gene expression was low; several predictors for a gene had R^2 values very close to zero indicating that they explained very little of the variation in gene expression in the training data sets (Table 2.6). Nevertheless, for each of these nine genes, variation in imputed gene expression explained about 10% of the variation in mean log10 HbA1c (last column, Table 2.6).

For each of the top hits, it is instructive to determine their relevance in other training sets. MAPK8IP1 was found in two other training sets (spleen and tibial nerve), with somewhat lower R^2_prediction (0.11 and 0.025) and higher p-values (0.012 and 0.10). GABBR1 was present in two other tissues (visceral fat and pancreas) with similarly low R^2-prediction (0.02 and 0.035) and high p-values (0.06 and 0.22). SPPL2A was present in three other training sets (skeletal muscle, liver and thyroid) and had similar R^2_predictions in each tissue (0.12, 0.14 and 0.12), and had a similarly low p-value in skeletal muscle (p=0.00081), but higher ones in liver (p=0.26) and thyroid (p=0.10).

SKOR1 was present in only one other training set (visceral fat), with a similarly low R^2_prediction (0.05), but also a low p-value (0.0059). The other suggestive genes IL17RC, TMEM136 and IFI44L were only present in one training set.

Interestingly, a few chromosomal regions were found to harbor clusters of genes with relatively low p-values: in particular genes on chromosome 15:50-52Mb, chromosome 17:56-58Mb, and chromosome 11: 45-48Mb consistently identified several of the best possible associations (Supplementary Tables). Collating all of the association results for which the imputed gene expression had an associated p-value <0.01 by chromosome and position, identified multiple genes and gene regions with consistently strong associations across training sets (Supplementary table A.1). For example, on chromosome 17 between 56 and 58 Mb, seven genes (*SMG8, TEX14, SKA2, YBX2, TRIM37, HEATR6* and *C17orf82*) were found to have relatively strong associations with mean $\log_{10}$ HbA1c in one or more tissues (Table 2.6, Supplementary Tables). On chromosome 11: 45-48Mb, *MAPK8IP1* was found to be associated with HbA1c in both DGN and GTEx blood (Table 2.6), and other genes in the same region (*CKAP5, C1QTNF4,* and *SLC35C1*) were found to have strong associations in other tissues (e.g. small intestine, supplementary Tables).

Table 2.6 Top ten hits (nine genes) showing associations between imputed gene expression and mean log10 HbA1c in five tissues.

The gene name, chromosome, approximate position (in Millions of Base pairs), the R^2 of the gene prediction model, the number of SNPs in the predictor, the effect size of the imputed expression (β_1), its associated p-value and the R^2 of the linear regression. Note that one gene, MAPK8IP1 was identified in two tissues.

Training Set	*Gene*	*Chr*	*Position*	*R^2-Pred*	*SNPs*	*β_1*	*p*	*R^2*
DGN-blood	*SPPL2A*	15	51M	0.11	33	0.05	1.16E-05	0.101
DGN-blood	*BRPF1*	3	95M	0.004	10	0.26	7.52E-05	0.099
DGN-blood	*MAPK8IP1*	11	46M	0.56	22	0.02	4.14E-05	0.099
WholeBlood	*MAPK8IP1*	11	46M	0.16	17	0.05	4.87E-05	0.099
WholeBlood	*SMG8*	17	57M	0.009	5	-0.16	1.56E-05	0.101
WholeBlood	*GABBR1*	6	28.5M	0.03	48	0.16	2.75E-05	0.100
liver	*IL17RC*	3	9.3M	0.12	57	0.05	5.44E-05	0.100
TibialNerve	*TMEM136*	11	120M	0.22	12	0.04	5.31E-05	0.099
TibialNerve	*SKOR1*	15	68M	0.09	13	0.06	9.63E-05	0.098
VisceralFat	*IFI44L*	1	78.5M	0.05	24	0.08	6.45E-05	0.099

The relationship between imputed gene expression and mean $\log_{10}$ HbA1c for the 1304 individuals in DCCT was examined graphically by extracting the imputed expression vector for the top ten hits from the output files (script given in the Appendix). In some cases, the imputed gene expression (X-axis) is not continuous but falls into discrete groups. The imputed gene expression will fall into discrete categories when 1) there are few SNPs in the predictor (for example SMG8 has only 5 SNPs in the predictor), 2) some SNP predictors show low genetic variation in the DCCT cohort and/or 3) SNPs have low MAF. In these cases, the imputed gene expression values fall well into haplotype-like groups as is seen for the genes *SMG8, MAPK8IP1* (Figure 2.10) and *MAP2K5* (Figure 2.11), while other genes such as *SPPL2A* and *IL17RC* exhibit continuous variation in imputed gene expression (Figure 2. 10 and 2.11).

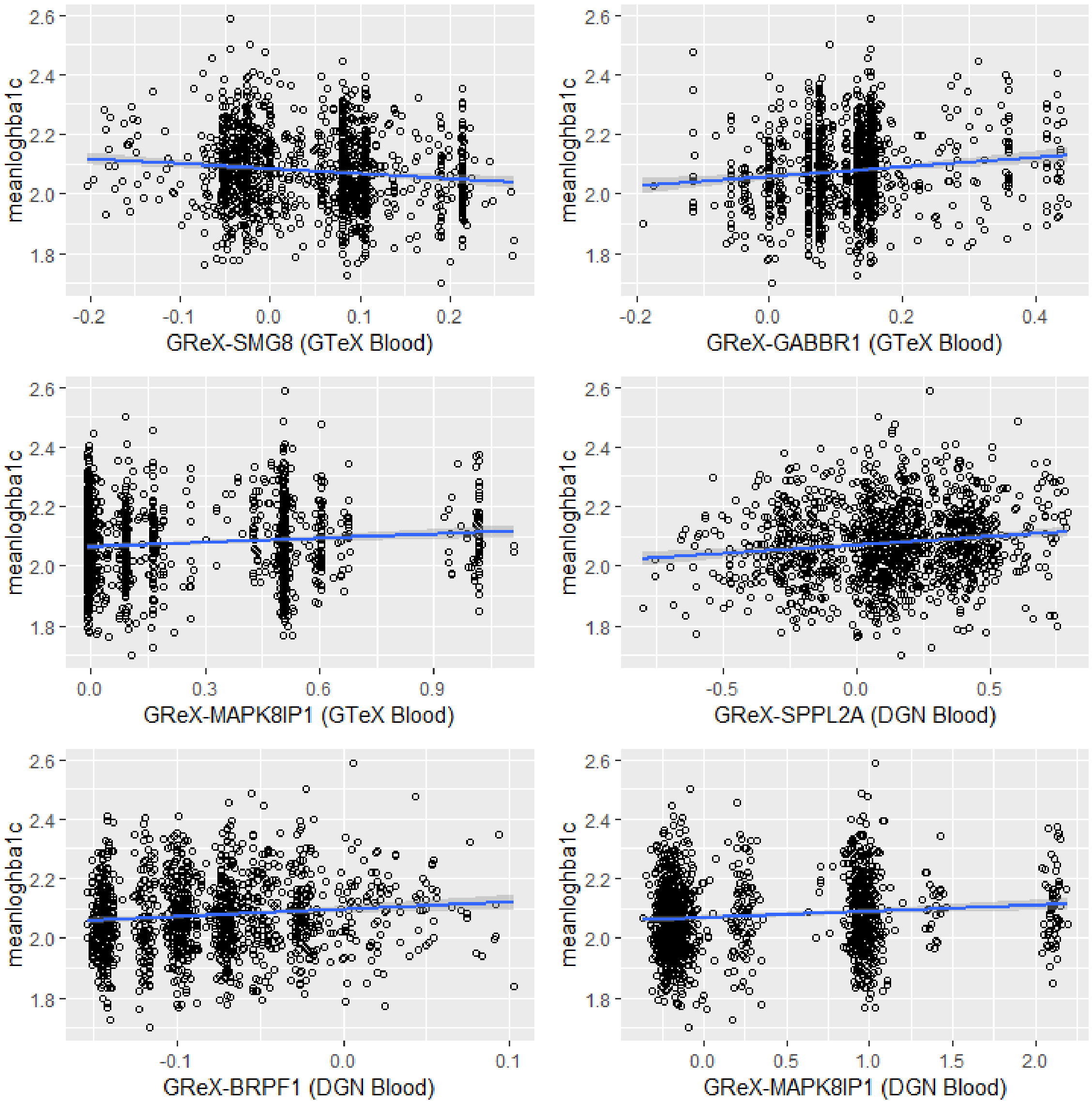


Figure 2.10. Scatterplot of imputed gene expression (X-axis) against mean $\log_{10}$ HbA1c for the 1304 individuals in the DCCT dataset for six of the top ten gene predictors based on the DGN-blood and GTEx-blood training sets.
Slope of the best linear model (blue) with 95 % confidence interval (gray shaded around line) shown.

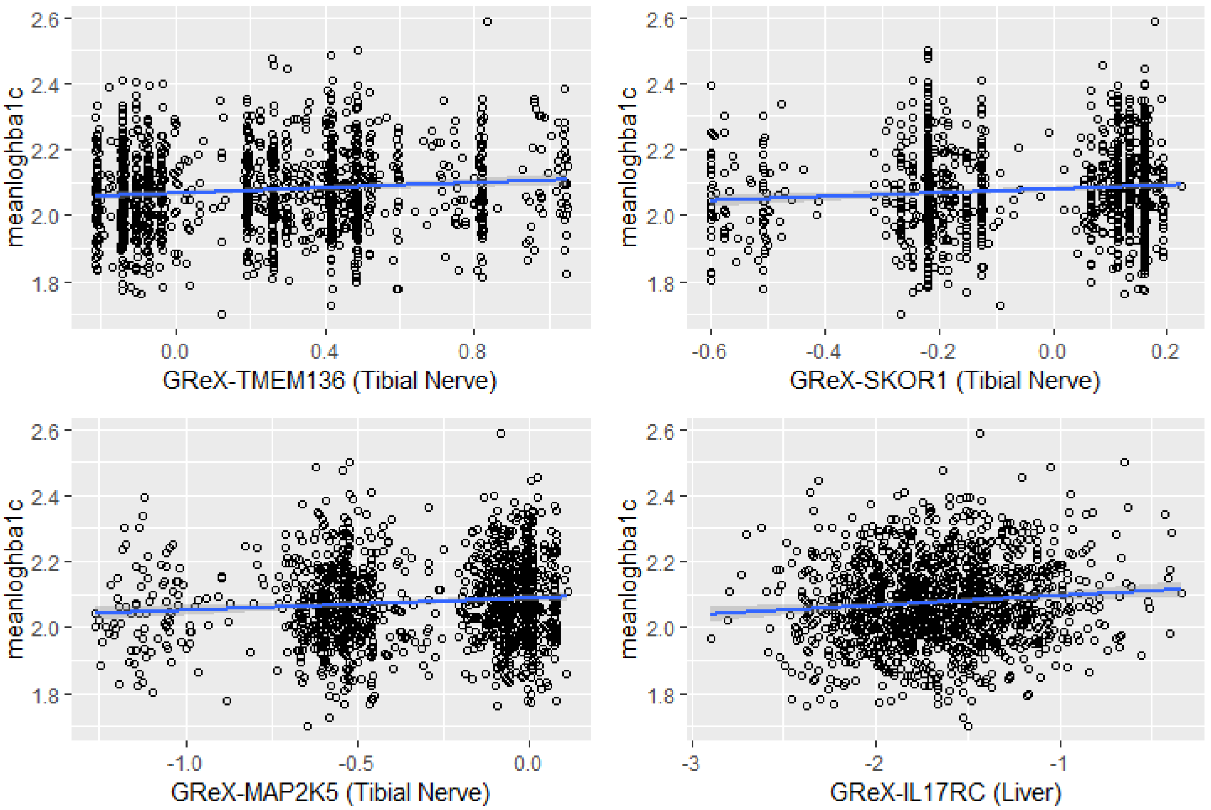


Figure 2.11. Scatterplot of imputed gene expression (X-axis) against mean $\log_{10}$ HbA1c for the 1304 individuals in the DCCT dataset for four of the top ten gene predictors based on the Tibial Nerve, Liver and Visceral Fat training sets.

For the top ten hits, the two-degree of freedom joint test for marginal gene expression and gene-treatment interaction effects was performed (Table 2.7). This test examines whether the association between variation in imputed gene expression and HbA1c was different in the two treatment groups, CONV vs INT. It revealed that the imputed gene expression of IL17RC in liver exhibited a significant gene*treatment interaction, with the individuals in the INT therapy having lower expression than those in the CONV therapy. Additionally, there was a suggestive gene*treatment interaction for the imputed gene expression of GABBR1 based on the GTEx-blood training set (Table 2.7).

Given that the "gene" in this study is genetically controlled component of gene expression, a treatment effect would be expected when differences in the magnitude and/or variance in HbA1c among treatment groups (CONV versus INT) gave greater power to detect a gene-effect.

Table 2.7 Two degree of freedom joint test for marginal gene expression and gene-environment interaction effects for the top ten gene predictors.
On the left set of panels the effect size of the gene*treatment interaction term for the model given in Equation 5 with its associated p-value is given. On the right hand side of the table, the results of the Likelihood Ratio Test comparing the alternative (Equation 5) to null (Equation 4) models are given. In all cases, the number of degrees of freedom (*d.f.*) was 10 for β_1 and 8 for β_0, and the χ^2 value was tested with 2 *d.f.*

Training Set	*Gene*	$\beta_{Gene*}Trt$	*P*	$L(\beta_1)$	$L(\beta_0)$	$\chi 2$	*P*
DGN-blood	*SPPL2A*	3.2E-02	0.17	914.8	904.2	21.23	2.4E-05
DGN-blood	*MAPK8IP1*	1.5E-02	0.12	913.8	904.2	19.3	6.5E-05
DGN-blood	*BRPF1*	-1.7E-01	0.19	912.9	904.2	17.5	1.5E-04
liver	*IL17RC*	-3.5E-02	0.04	912.7	904.2	17.1	1.9E-04
TibialNerve	*TMEM136*	7.2E-04	0.97	912.4	904.2	16.4	2.0E-04
TibialNerve	*SKOR1*	1.2E-02	0.68	911.9	904.2	15.5	4.0E-04
VisceralFat	*IFI44L*	1.8E-02	0.65	904.3	904.2	0.22	0.89
WholeBlood	*SMG8*	-3.3E-02	0.66	913.7	904.2	18.9	7.6E-05
WholeBlood	*GABBR1*	1.3E-01	0.08	914.6	904.2	20.7	3.1E-05
WholeBlood	*MAPK8IP1*	3.3E-02	0.16	913.5	904.2	18.6	9.1E-05

Chapter 3: Discussion

Integrated analysis of genetic and gene expression data provides an elegant way of assessing the consequences of putative regulatory sequence variants on transcription. In this thesis, I present the results of a recently developed approach for performing transcriptome imputation as implemented by Gamazon et al (2015) in the PrediXcan workflow (GAMAZON *et al.* 2015). In particular, I used predictions from ten training sets developed by Gamazon *et al.* (2015) to impute gene expression in 1304 Caucasian individuals with T1D who were participants in the DCCT and then tested for associations between imputed gene expression and variation in HbA1c among individuals.

3.1 Putative functions and disease associations of top hits

Although none of the genes reached statistical significance, the function and disease associations of the nine genes showing the strongest association with HbA1c were explored at the Ensembl genome website (www.ensembl.org), and seven of the nine genes were found to have feasible associations with diabetes mellitus, blood glucose levels or immune function (Table 3.1). For example, two of the genes, SPPL2A and IFI44L are associated with fibrinogen and adiponectin function respectively. Fibrinogen is a glycoprotein found in the blood that is involved with blood clotting while adiponectin is a protein hormone directly involved in regulating blood glucose. The gene GABBR1 maps to chromosome 6p21 within the HLA class I region; in addition to associations with blood protein levels, Crohn's disease and lung cancer (Ensembl website), some studies have investigated its involvement in insulin resistance (BANSAL *et al.* 2011). However, the association between HbA1c and GABBR1 divulged here may be caused by the presence of SNPs in the expression predictor for GABBR1 overlapping with the SNPs in the HLA region which has well-established associations with T1D (discussed below). The gene that was associated with HbA1c in two training sets, MAPK8IP1, has known associations with type 2 diabetes (Table 2.8), and another gene, TMEM136, influences intraocular pressure – a trait impacted by diabetes mellitus. Lastly, two of the genes (IL17RC, IFI44L) were associated with immune responses. Collectively, this indicates that this approach may be promising but more studies are needed to confirm these findings.

Table 3.1 Known phenotypic, disease or trait associations of the top nine genes associated with HbA1c in the PrediXcan analyses.
Data obtained from the Ensembl Release 93 (July 2018) database (Accessed July 20, 2018).

Gene	*Source*	*Phenotype, disease or Trait*
SPPL2A	NHGRI-EBI GWAS catalogue	Fibrinogen levels; neutrophil levels
MAPK8IP1	OMIM	Diabetes mellitus, noninsulin dependent
BRPF1	OMIM	Intellectual developmental disorder
IL17RC	OMIM	Candidiasis, familial 9
	Orphanet	Chronic mucocutaneous candidiasis
TMEM136	NHGRI-EBI GWAS catalogue	Intraocular pressure
SKOR1	NHGRI-EBI GWAS catalogue	Restless legs syndrome
IFI44L	dbGAP	Adiponectin
	NHGRI-EBI GWAS catalogue	ASD; immune response to measles vaccine; schizophrenia
SMG8	none	
GABBR1	NHGRI-EBI GWAS catalogue	Blood protein levels; crohn's disease; lung adenocarcinoma

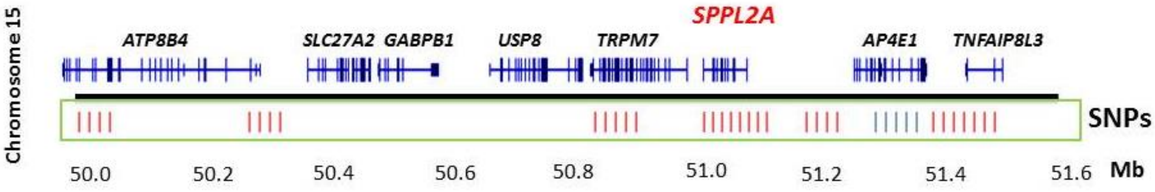
3.2 Loci associated with HbA1c in GWAS studies

Glycated hemoglobin (HbA1c) is a long-term glycemic control indicator for T1D and a diagnostic marker for type 2 diabetes. A number of studies have looked for loci associated with variation in HbA1c (Florez *et al.* 2007; Pare *et al.* 2008; Franklin *et al.*

2010; Soranzo *et al.* 2010; Ryu and Lee 2012). One of the largest is a meta-analysis by Soranzo et al (2010), encompassing 46,368 non-diabetic individuals in which they identified significant loci near 10 genes namely *FN3K* (17:78Mb), *HFE* (6:26Mb), *TMPRSS5* (22:36Mb), *ANK1* (8:42Mb), *SPTA1* (1:156Mb), *ATP11A* (13:112Mb), *HK1* (10:70Mb), *MTNR1B* (11:92Mb), *GCK* (7:44Mb) and *G6PC2* (2:170Mb)(Soranzo *et al.* 2010). It should be noted that in GWAS analysis, the closest gene cannot be assumed to be the causal locus (discussed below) (Consortium *et al.* 2017). As discussed in Soranzo *et al.* (2010), three of these loci are associated with glycemia (*GCK, G6PC2* and *MTNR1B*) and several others are associated with hereditary anemias and iron storage disorders, and thereby probably influence HbA1c levels via their influence on erythrocyte biology. For example, genetic variants near HK1 have been shown to influence HbA1c because they affect the average life-span of RBC's not because they influence blood glucose levels (Cohen *et al.* 2008; Bonnefond *et al.* 2009).

The most relevant GWAS study for this thesis is that performed by Paterson *et al.* (2010), on the same DCCT genotype dataset examined here. In addition to the GWAS association on chromosome 10:108Mb (based on hg19) near the gene *SORCS1*, they additionally identified SNPs near chr 9:17Mb, chr 14:96Mb and chr 15:51Mb, all in non-coding regions with the closest genes being *BNC2, GSC* and *WDR72* respectively. While our study did not identify any genes near chr10:108Mb or chr9:17Mb, we do identify genes at chr14:96Mb and chr15:51Mb (supplementary data). The genes identified near 14:96Mb are *IFI27L1* (an immune system gene), *TCL1A* and *TCL1B*, while the genes identified at 15:51Mb are *LYSMD2* and *SPL22A.* Remarkably, the gene *SPL22A* was one of the top hits identified based on transcriptome imputation suggesting that the SNPs in this predictor warrant further investigation as well as the function of the gene itself. Figure 3.1 presents a cartoon of the SNPs included as the gene expression predictor for SPPL2A based on the DGN blood training set. These SNPs span a region from ~ 50.0 to ~ 51.5Mb (Figure 3.1 and supplementary table A.2), but also clustered in six groups, and span 8 curated genes based on GRCh 37.0. Notably, the five SNPs identified in the Paterson *et al.* (2010) study with near genome-wide significance with HbA1c lie within the expression predictor for SPPL2A.

Figure 3.1 Cartoon diagram of the approximate location of the SNPs in the gene expression predictor for the gene *SPPL2A* based on the DGN blood training sets (red vertical lines) relative to the location of genes that span the region.
Also shown is the approximate location of the SNPs identified to be near genome-wide significant (vertical blue lines) in the GWAS by Paterson *et al.* (2010).



In addition to the overlap between these two genic regions with those identified by Paterson *et al.*, several of the gene regions identified by Soranzo *et al.* are also close to regions harbouring multiple genes associated with HbA1c in this study. For example, Soranzo et al identified a GWAS hit near *MTRN1* on chr11:92Mb (rs1387153); although this gene was not identified *per se*, the gene *MRE11A* on chr 11:93Mb was associated with HbA1c in multiple training sets and three other genes in close vicinity *FAT3* (11:92), *ANKRD49* (11:94) , *ENDOD1* (11:94) and *FOLR4* (11:94) were each identified in a single training set to have associations p<0.01 with HbA1c (Supplementary data). Further work should examine the SNPs in the gene expression predictors for these loci to assess if any

SNPs are overlapping (or in strong LD) with SNPs in tandem gene predictors (thus explaining while multiple genes in the same region are all suggestive), known eQTL's and/or have known GWAS significant associations with HbA1c.

3.3 Association of variation in gene expression with traits and disease.

An implicit assumption of transcriptome imputation, is that variation in gene expression influences phenotype traits and/or disease state. As discussed in the introduction, one of the main impetuses for this hypothesis came from the fact that the majority of GWAS hits occur in non-coding regions and GWAS hits are enriched for eQTL's and vice versa suggesting that variation in regulatory regions influences disease (ALBERT AND KRUGLYAK 2015). The GTEx consortium was initiated in large part to investigate this hypothesis and provide a public repository of data on variation in gene expression among individuals. The GTEx project found that variation in gene expression among individuals accounted for only 5% of variation in gene expression while much larger differences in gene expression were observed among tissues. Inter-individual variation in gene expression was influenced by sex, ethnicity and age; for example about 2,000 genes (or 10% of the protein coding genes) were found to be differentially expressed with age, including genes related to neurodegenerative diseases such as Parkinson's and Alzheimer's diseases (CONSORTIUM *et al.* 2017). Even though gene expression was markedly different among tissues, surprisingly few genes were found to have tissue-specific gene expression, and similarly few eQTL's were found to be associated with expression in a single tissue. Interestingly, eQTL's that influenced gene expression in multiple tissues were depleted for LOF intolerant genes as would be predicted if these loci are subject to purifying selection so as to remove large-effect regulatory variants (CONSORTIUM *et al.* 2017). This provides some background for factors influencing gene expression across individuals and tissues, but what about the association of gene expression with disease?

There are now several well-verified examples of how variation in gene expression may influence disease (ALBERT AND KRUGLYAK 2015). One of the best examples concerns a regulatory variant that influences myocardial infarction. The minor allele of a non-coding SNP in the 3' UTR of the gene *CELSR2* creates a TFBS for a CCAAT/enhancer binding protein (C/EBP) to which the major allele does not bind. However, binding of C/EBP to the minor allele at the newly created TFBS leads to increased expression of the sortilin 1 (*SORT1*) gene in liver cells. Experimental studies in mice have shown that overexpression of *Sort1* in the liver reduces low-density lipoprotein cholesterol (LDL-C) levels which in turn decreases the risk of myocardial infarction (MUSUNURU *et al.* 2010). This provides a complete causal path from a non-coding variant to variation in gene expression to an altered risk for a major human disease (ALBERT AND KRUGLYAK 2015). Other experimentally verified examples are emerging as well. For example, Kapoor *et al.* identified a causal regulatory variant that influences heart function by altering an enhancer in heart tissue which had been missed in previous eQTL studies using different tissues (KAPOOR *et al.* 2014). Additionally, a suite of papers have found associations between imputed gene expression and disease phenotypes, and some of these have experimental support (HOFFMAN *et al.* 2017; HOHMAN *et al.* 2017; WHEELER *et al.* 2017b).

3.4 Strengths and weaknesses of transcriptome imputation using PrediXcan

One of the main advantages of transcriptome imputation is that it provides a statistical framework for integrating the combined effect of multiple SNPs on gene expression. Gamazon et al showed that using only the top SNP to predict gene expression, as in standard eQTL analyses, is not as good as aggregating all putative regulatory SNPs into an expression predictor using regression regularization or BSLMM (GAMAZON *et al.* 2015; WHEELER *et al.* 2016). Even so, although the mean number of SNPs in a predictor was close to 30, many expression predictors are dominated by the influence of a few SNPs (Figure 3.4). For example, in the expression predictors for MAPK8IP1, one of the most robustly associated genes with HbA1c, four of the 22 SNPs in the expression predictor based on the DGN blood training set had weights >0.1 while the remaining 18 SNPs had weights between -0.02 and 0.1 (Figure 3.4). On the other hand, the distribution of weights for the same gene based on the GTEx blood training set was lower, with all weights < 0.1. The presence of higher weights in the DGN data set is in keeping with the overall higher R^2 of expression predictors from DGN than GTEx blood (Figure 2.7) and suggests that differences in the acquisition and processing of the two datasets (outlined in table 1.1) likely influenced the predictive power of the training sets. For example, the DGN blood training set was based on 922 individuals (rather than 336 as for GTEx blood), was depleted of globin prior to RNA sequence, and was subjected to a different process of normalization than the RNA seq data from GTEx. This suggests that variation in sample size, transcriptome sequence and data analyses may all influence the reliability of transcriptome imputation, although curiously it did not appear to influence the number of significant associations found between imputed gene expression and HbA1c in this study.

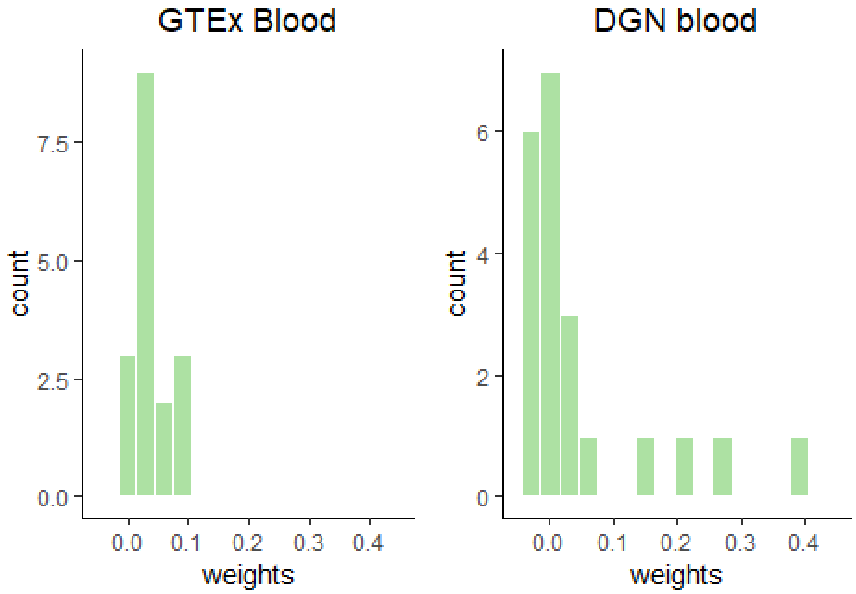


Figure 3.2 Distribution of the weights used to predict GReXg for MAPK8IP1 in both GTEx whole blood (N=17 SNPs) and DGN blood (N=22 SNPs) training sets.

One of the most useful features of both eQTL analyses and transcriptome imputation is that they may help identify the candidate genes influencing traits. As noted by the GTEx consortium, eQTL's frequently lie greater than 0.5Mb from a gene (CONSORTIUM 2015; CONSORTIUM *et al.* 2017). In the most recent release of the GTEx catalogue, they estimated that only 40% of GWAS signals co-localize with the nearest expressed gene (CONSORTIUM *et al.* 2017).

There are a number of caveats with respect to the implementation of transcriptome imputation. One of the primary issues is that regulatory elements and therefore the regulatory impact of sequence variation is highly cell type specific. The GTEx project provides a large-scale cross-tissue eQTL database, however tissues are composed of multiple cell-types and eQTL's identified based on tissue level-expression may be biased towards the most common cell types or, more problematically, identify spurious signals since tissue-level expression is based on the weighted average of gene expression values across all cell types. Thus, it is possible that different cellular conditions change the magnitude and/or direction of individual SNP effects, and incorrect predictions would be obtained by using tissue-level data and experimental conditions that are not representative of those influencing disease conditions. This will improve in the future as eQTL mapping based on single cell transcriptomic data is being developed (KANG *et al.* 2018) which will allow quantification and mapping of cell to cell variability in gene expression. Variation in cellular diversity among tissues may explain why some tissues (i.e. those with fewer cell-types) have higher gene-prediction R^2 than others (e.g. DGN and GTEx blood, pancreas vs blood, Figure 2.4).

Secondly, an important caveat, but also potential benefit of transcriptome imputation is that some variants are associated with multiple nearby genes (personal observation); this observation is consistent with the GTEx consortium eQTL analyses which has shown that some variants are associated with more than 30 different nearby genes (CONSORTIUM *et al.* 2017), but importantly, for over 10% of the eQTL's, the gene for which they drive the strongest associations varies between tissues. The presence of shared high-weight SNPs in the gene predictors for neighbouring genes, or strong LD between such SNPs, is driving the observation that multiple neighbouring genes are found to be strongly associated with HbA1c in this study. This observation can be seen both as a benefit and caveat of transcriptome imputation methods. It is beneficial because it reflects the real complexity of gene regulation in which changes in *cis*-elements affecting chromatin remodulation or response elements would impact multiple nearby genes, acting as local master *cis*-regulators. On the other hand, it is also an important caveat of transcriptome imputation because it can obscure the putatively causal gene or genes.

One of the most important challenges for TWAS as for GWAS is to move from correlation or co-localization toward causation. Clearly this is the most difficult task and always requires experimental validation in addition to rigorous statistical analyses. While transcriptome imputation aids in the identification of functional genetic effects on traits, it still does not provide a direct link to causality. Under the feasible assumption that gene expression mediates the effect of genetics on complex traits, testing for association between predicted gene expression and traits is equivalent to a two-sample Mendelian randomization test for a causal effect of expression on a trait (PICKRELL 2015). This test for causality is valid only when SNPs do not exhibit pleiotropic effects, which is difficult to prove and therefore, TWAS associations do not provide direct evidence of causal relationships between gene expression and complex traits but rather reflect associations between expression levels and traits.

3.4.1 Co-localization Methods

In this context, I should expand upon a class of methods that aims to integrate GWAS and eQTL data under the reasonable assumption that colocalization of both variants suggests that the variant may be causal. One of the first groups to look for agreement between GWAS and eQTL variants was NICA *et al.* (2010) who developed a method called regulatory trait concordance (RTC) that identifies variants that are causal in both GWAS and eQTL of studies while accounting for LD. RTC works on the principal that when a GWAS causal variant colocalizes with an eQTL causal variant, re-computing the marginal statistics for the eQTL variant by conditioning on the GWAS causal variant will remove the significant association signal observed at the locus under true co-localization. Co-localization approaches were limited by the availability of well-powered eQTL datasets, but with the availability of well-powered eQTL datasets for many tissues by the GTEx consortium, new statistical approaches are being develop, such as that by HORMOZDIARI *et al.* (2016) as implemented in the program eCAVAIR and another one by WEN *et al.* (2017) that performs enrichment analysis of *cis*-eQTL's in GWAS hits, as well as fine-scale mapping and co-localization of variants using the program *enloc*.

In the method developed by HORMOZDIARI *et al.* (2016), they assume that the posterior probability that a variant is causal in both GWAS and eQTL's is independent, an assumption which seems valid given that GWAS loci are identified by regressing the phenotype (e.g. disease) against genotyped SNPs, while eQTL's are identified by regressing variance in gene expression against genotyped SNPs. To account for LD, in eCAVIAR HORMOZDIARI *et al.* (2016) considers a collection of variants (up to 50) around the GWAS top hit as a single locus, and then for all variants at the locus, use the marginal statistics of the eQTL's for all eGenes (genes with at least 1 eQTL) in a tissue to derive the posterior probability that a given variant is causal in both the GWAS and eQTL study. Their method purports to accurately quantify the amount of support for co-localization but also identify scenarios where the variants underlying both studies are different.

To test their approach, HORMOZDIARI *et al.* (2016) used the GWAS results from the MAGIC (Meta-analyses of Glucose and Insulin-Related Trait Consortium) dataset for the traits Fasting Glucose (FG) and Fasting ProInsulin (FP). They then obtained eQTL's data for all eGenes in 44 tissues from GTEx (release v.6) as well as eQTL's for pancreatic islet tissue from the study by VAN DE BUNT *et al.* (2015). Using eCAVIAR, they identify GWAS loci that share a causal variant with eQTL's in one or more tissues for both FG and FP, but an equal number of loci where GWAS and eQTL causal variants appear to be different, suggesting that the genetic factors underlying disease mechanisms are more complex than previously thought. They suggest that failure to find co-localization between some GWAS and eQTL variants could be due to a) stronger associations of eQTL's than GWAS variants with gene expression *per se* b) the possibility that GWAS variants affect other aspects of gene regulation (splicing etc.) or c) the possibility that GWAS variants only affect expression under certain conditions (e.g. developmental etc) as iterated by other authors as well.

However, several of the genes identified by HORMOZDIARI *et al.* (2016) as influencing FG were also identified as having strong associations with mean HbA1c in this thesis. As described in Table 1 of HORMOZDIARI *et al.* (2016), co-localization of GWAS variants identified in the MAGIC study with eQTL's from GTEx and in pancreatic islets was observed only for 5 SNPs located on three chromosomes. Two of these SNPs co-localized between a MAGIC GWAS variant and an eQTL associated with gene expressed in pancreatic islets (VAN DE BUNT *et al.* 2015); and for the other three SNPs , co-localisation was observed for one, three or six GTEx tissue and one of the SNPs affected expression in five different genes (HORMOZDIARI *et al.* 2016).

Specifically, HORMOZDIARI *et al.* (2016), identified co-localization for a SNP (rs11717195) on chr 3: 123,082,398 that affected expression of the gene ADCY5 in pancreatic islets but none of the GTEx tissues. Interestingly, however, ADCY5 was identified here (Table A.1) as being negatively correlated with meanlogHbA1c (p<0.01) based on the predicted expression in DGN blood. Similarly, they also found a GWAS variant influencing FG in the MAGIC (rs4607517) study which was an eQTL influencing expression of the gene glucokinase (*GCK*) on chr 7: 44,235,668 in two GTEx tissues (colon, sigmoid and thyroid), which is one of the genes found to be associated with HbA1c in the SORANZO *et al.* (2010) study as well. While we did not find associations of predicted expression of *GCK* with HbA1c, we did find three other genes in the region, chr7:44Mbp, that were associated with mean HbA1C, namely *C7orf44, POLD2* and *STK17A* (Table A.1). Strikingly, the HORMOZDIARI *et al.* (2016), found significant co-localization of a GWAS variant (rs11605924) with gene expression of *MAPK8IP1* in GTEx whole blood. Although this SNP is not in the expression predictor for *MAPK8IP1* that was employed here (not shown), the association of *MAPK8IP1* in both the HORMOZDIARI *et al.* (2016) study and the results presented here, strongly suggests that some of the SNPs in the expression predictor for *MAPK8IP1* are in LD with rs11605924, and that variation in expression of this gene may influence HbA1c in the DCCT cohort as well, something that will be explored in future work. Lastly, HORMOZDIARI *et al.* (2016) identify a GWAS variant (rs7944584) located on chr 11: 47,336,320 bp as being associated with expression of five genes in six tissues. Notably, two of these genes, *MADD* and *NR1H3* were also identified here as influencing HbA1C (at p<0.01), while I also identified a third gene in this region, *C1QTNF4*, which was marginally associated (p<0.01) with gene expression (Table A.1), but was also tabled by HORMOZDIARI *et al.* (2016), as being the focal locus for a GWAS variant (rs 10501320) influencing FP in MAGIC as well as being an eQTL for esophagus based on the GTEx data. The overlap between the co-localization results of HORMOZDIARI *et al.* (2016) with the data presented here is encouraging that the PrediXcan approach could be useful in identifying functional variants in these T1D cohorts.

As more studies employ transcriptome imputation approaches it is becoming evident that it is a technique complementary to GWAS; it may not necessarily identify many new loci on its own but it could be an important piece of identifying causal loci associated with GWAS variants and for understanding the complexity of the effect of gene regulation on phenotypes. For example, a recent study compared three different approaches for transcriptome imputation to GWAS results using the data from the Wellcome Trust Case Control Consortium (WTCCC) for Crohn's disease and T1D (FRYETT *et al.* 2018). The results of this study indicate that the three approaches produced highly correlated results when applied to the same GWAS data (with PrediXcan and MetaXcan providing more consistent results with previous studies than FUSION), except for genes in or near the major histocompatibility complex, which is unsurprising because of the complex haplotype and LD structure at this locus that also renders genotype phasing and imputation challenging. Perhaps more importantly, they find that the genes identified as significant in transcriptome imputation occur near known GWAS risk loci, although fewer associations were identified using transcriptome imputations than with GWAS analyses (FRYETT *et al.* 2018).

3.5 Relevance for Public Health

The role of genetics in public health is a large topic that is difficult to review succinctly, and has been well-reviewed in other works (DAVEY SMITH *et al.* 2005; SEYERLE AND AVERY 2013; FALLIN *et al.* 2016). Here, I give a brief overview of some of the main advancements in genetic epidemiology since its inception and then provide a concrete example of how information gathered from genetic epidemiological approaches is informing clinical interpretation of HbA1c results.

The use of genetic data to help elucidate the aetiology of disease has a history almost as old as that of genetics itself. In 1954, Neel and Schuul studied inheritance patterns in large family pedigrees to investigate the mode of inheritance of disease (DEWAN 2010). In the following decades, the lack of polymorphic molecular markers in humans limited the advancements of genetic epidemiology to the development of its theoretical underpinnings in population genetics and statistical genetics, which, nevertheless, was a profoundly fruitful period, and continues to be a major strength of the field. By 1980, polymorphic molecular markers (e.g. restriction endonucleases length polymorphisms (RFLPs) and simple tandem repeats (STR's)) enabled the development of better genetic linkage maps of the human genome which when combined with pedigree data, provided sufficient resolution to localize and identify the genes underlying some rare inherited Mendelian disorders, such as familial breast cancer, cystic fibrosis and Tay-Sachs disease (FALLIN *et al.* 2016). This initiated the integration of genetic counseling into public health. Toward the end of the 20th century, there was a move away from using only large family pedigrees to identify genes associated with Mendelian diseases, and a move towards the implementation of traditional epidemiological designs especially case-control and cohort designs (FALLIN *et al.* 2016). This was partly due to development of denser genetic maps (precursors to the HapMap project) that allowed researchers to map or localize genes based on LD. Fallin *et al.* (2016) suggest that the period between 1980-2001 could be called the Mendelian era of genetic epidemiology, as it was typified by the development of public health strategies for the screening and prevention of Mendelian diseases. However, there was still little advancement in understanding the genetic basis of complex diseases, such as coronary vascular disease (CVD) or diabetes.

With the completion of the sequencing of the human genome in 2001 (Lander *et al.* 2001; Venter *et al.* 2001) and the development of the HapMap SNP reference panels soon after in 2003-2005 (International HapMap 2003; International HapMap 2005) the field of genetic epidemiology expanded rapidly. These advancements were accompanied by the development of genotyping arrays and tagged SNPs that led to an enormously successful period of genome-wide association studies (GWAS), in which the genetic risk factors for many complex diseases from increasingly large case-control or cohort populations were identified (Visscher *et al.* 2012; Manolio 2013). As described by Manolio (2013) between 2005 and 2013, GWAS identified ~2,000 robust associations with complex diseases, and now over 10,000 novel GWAS loci have been identified (MacArthur *et al.* 2017; Visscher *et al.* 2017). However, the clinical significance of these findings has been limited because the loci identified generally have modest effect sizes and are usually associated with variants in non-coding regions with unclear functions. Thus, although this period was very successful in elucidating the genetic and biological basis of disease, it has translated into relative few applications in public health and clinical care *per se*. Nevertheless, a few loci have delivered strong predictive or prognostic information for clinical care (Visscher *et al.* 2017); for example, there are now prognostic genetic variants for the response of individuals to pharmaceuticals, which is increasingly used to calculate effective medication dosage.

One salient example of how information about a single genetic variant influencing HbA1c can influence public health was recently documented by Wheeler et al (WHEELER *et al.* 2017a) and expounded upon by PATERSON (2017). Wheeler *et al.* (2017) identified an X-linked SNP (rs1050829) that influences levels of HbA1c, which is a diagnostic tool for type 2 diabetes as well as a monitor for blood sugar levels in individuals with T1D. The missense SNP they identified is associated with a particular haplotype of Glucose-6-phosphae dehydrogenase, *G6PD*, which has known associations with lower enzyme activity of *G6PD*, shorter red cell lifespan and, now, lower HbA1c, but not fasting glucose. The variant has a minor allele frequency of 10-15% in African Americans without diabetes but is not polymorphic in the European/Asian populations. However, the variant explains a large amount of variance in HbA1c in African Americans; in males, it explains 13%–20% of the variance in HbA1c, and in females 2%–10%. As described by Paterson (2017), the reason that this finding has considerable public health significance is that diagnosis of type 2 diabetes has shifted from measures of fasting glucose and oral glucose tolerance test towards HbA1c. Currently, HbA1c levels greater than 6.5% [48 mmol/mol] are considered diagnostic of type 2 diabetes. However, in males with this risk variant, average HbA1c levels are 0.8% lower, meaning that an individual diagnosed with an HbA1c level of 6.5% would, on average, have a true HbA1c of 7.3%, putting them at considerably increased risk of long-term complications from pre-existing type 2 diabetes.

This finding led Wheeler *et al.* (2017) to suggest that both sex- and genotype-adjusted recommendations should be in place for Type 2 diabetes diagnosis, a topic that has been the focus of considerable debate (HERMAN 2016; SELVIN 2016). As pointed out by Paterson (2017) another implication of this work is that more effort should be placed on obtaining genotype data from non-European populations now that whole-genome sequencing costs are much lower.

The above finding that a single genetic variant can influence mean HbA1c was discovered using GWAS approaches, however, usually it is difficult to identify the causal (genetic) locus for many GWAS hits, and researchers must turn to functional genomic approaches such as eQTL and TWAS analysis to identify the loci and functional impact of genetic variant(s) on traits or disease risk. Furthermore, as outlined above, there is increasing evidence that variation in gene expression contributes significantly to phenotypic or disease traits. Once identified, animal models can sometimes be employed to test the impact of these variants on animal physiology and/or disease state, which can lead to improved diagnostics or pharmaceutical targets in clinical settings. In this way, TWAS analyses can contribute directly to our understanding of the effect of genetic variants to disease and play a role in informing strategies for personalized medicine in public health.

3.6 Conclusions and Future Directions

In conclusion, by employing transcriptome imputation analyses using genotype data from 1304 Caucasians who participated in the DCCT trial, weakly suggestive associations were found between the imputed gene expression and mean $\log_{10}$ HbA1c levels for nine genes. Several of these genes mapped to chromosomal regions with known GWAS significant or strongly suggestive hits, and several of the genes or genomic regions were also identified by HORMOZDIARI *et al.* (2016) as having an eQTL that co-localized to a GWAS variant influencing FG, most notably the gene *MAPK8IP1*.This suggests that although none of the genes reached tissue-wide significance in this study, the approach should facilitate identification of causal loci influencing HbA1c, which is a marker of long-term glycemia, one of the most important risk factors for developing complications from T1D.

One of the first avenues that will be explored in future work is to repeat this analysis with the training sets developed by the Im lab based on the newest version of the GTEx dataset (v7) (http://predictdb.org/). Once this is completed for the DCCT dataset, I will perform the same analyses on mean $\log_{10}$ HbA1c for other long-term T1D cohort datasets available in the Paterson lab and perform a meta-analysis. Using the results from the meta-analysis, the expression predictors for all genes in regions for which significant associations are identified will be examined to look for evidence of overlap (co-localization) of specific SNPs in the expression predictors with GWAS variants or variants in strong LD with them. This could be done informally by obtaining LD information on the SNPs in the expression predictors from LDlink (https://ldlink.nci.nih.gov/), and overlaying the GWAS variants, expression predictor SNPs as well as other functional data such as epigenomic data (e.g. NIH Roadmap Epigenomics (BERNSTEIN *et al.* 2010)), or more formally by performing a co-localization analysis such as that employed in the approaches of eCAVAIR or *enloc* (HORMOZDIARI *et al.* 2016; WEN *et al.* 2017). Given the encouraging signals thus far that have identified the same regions and/or genes associated with

HbA1C (PATERSON *et al.* 2010) or FG (HORMOZDIARI *et al.* 2016), this work underscores that although transcriptome imputation may not uncover more loci than GWAS analyses, it is a useful and complementary tool to GWAS because it can help to identify causal loci associated with disease phenotypes, and is thus of relevance for work in public health and the growing field of personalized medicine.

Supplementary Data A: Tables

Table A.1 Association results for the relationship between imputed gene expression and mean log HbA1c for 1304 Caucasian individuals that participated in DCCT.
All associations with a p-value <0.01 are listed in order of chromosomal position. R2_pred is the R^2 of the predicted gene expression while R2_Assoc is the R^2 of the linear model to test the association between imputed gene expression and HbA1c.

tissue	*gene*	chr	Mb	Snps	R2_pred	β	SE	T	pval	R2_Assoc
DGNblood	*HES4*	1	1	21	0.184	-0.026	0.01	-2.858	0.00433	0.094
TibNerve	*MEGF6*	1	3	3	-0.062	-0.118	0.04	-2.830	0.00473	0.094
DGNblood	*UTS2*	1	7	111	0.005	0.016	0.01	2.785	0.00543	0.094
TibNerve	*KIF1B*	1	10	13	0.033	0.050	0.02	2.817	0.00493	0.094
TibNerve	*TMEM201*	1	10	44	0.040	0.043	0.02	2.614	0.00906	0.093
pancreas	*KIF1B*	1	11	7	-0.095	0.095	0.04	2.627	0.00872	0.093
DGNblood	*DNAJC16*	1	15	4	-0.022	0.369	0.12	3.136	0.00175	0.095
GTExBlood	*PRDM2*	1	15	23	0.013	-0.075	0.03	-2.838	0.00461	0.094
TibNerve	*C1orf134*	1	16	6	0.055	-0.121	0.04	-3.139	0.00174	0.095
TibNerve	*CASP9*	1	16	45	0.003	0.031	0.01	2.675	0.00756	0.093
ViscFat	*CASP9*	1	16	29	0.000	0.047	0.02	2.621	0.00888	0.094
GTExBlood	*CASP9*	1	16	39	0.044	0.049	0.02	2.635	0.00851	0.093
DGNblood	*CTRC*	1	16	12	0.004	0.022	0.01	2.720	0.00662	0.093
DGNblood	*FBXO42*	1	17	21	0.012	0.054	0.02	3.127	0.00181	0.095
DGNblood	*MST1P9*	1	17	50	-0.017	-0.024	0.01	-3.152	0.00166	0.095
TibNerve	*RSG1*	1	18	28	0.031	-0.031	0.01	-2.829	0.00475	0.094
GTExBlood	*KIF17*	1	20	19	-0.020	-0.075	0.03	-2.762	0.00582	0.094
Ileum	*MUL1*	1	20	33	-0.052	-0.031	0.01	-2.982	0.00292	0.094
TibNerve	*HMGCL*	1	25	24	-0.030	0.063	0.02	2.628	0.00870	0.093
GTExBlood	*RHCE*	1	25	6	0.006	0.947	0.26	3.579	0.00036	0.097
spleen	*TMEM54*	1	33	23	-0.001	-0.042	0.02	-2.818	0.00490	0.094
liver	*RP11-268J*	1	37	36	-0.032	-0.042	0.01	-3.594	0.00034	0.098
ViscFat	*SH3D21*	1	37	5	-0.038	0.137	0.05	2.597	0.00952	0.093
spleen	*MTF1*	1	39	68	-0.031	0.022	0.01	2.689	0.00726	0.093
DGNblood	*YRDC*	1	39	9	-0.004	-0.114	0.04	-2.718	0.00665	0.093
GTExBlood	*BMP8A*	1	40	12	-0.014	-0.065	0.02	-2.689	0.00726	0.093
TibNerve	*AL357673*	1	55	35	-0.012	0.064	0.02	2.635	0.00852	0.093
DGNblood	*OMA1*	1	59	47	0.017	0.029	0.01	2.674	0.00760	0.093
pancreas	*TCTEX1D1*	1	67	4	-0.007	0.210	0.06	3.476	0.00053	0.098
ViscFat	*IFI44L*	*1*	*78*	*24*	*0.006*	*0.078*	*0.02*	*4.009*	*0.00006*	*0.100*
ViscFat	*IFI44*	1	79	41	0.000	0.067	0.02	3.399	0.00070	0.096
DGNblood	*SYDE2*	1	86	11	0.006	-0.079	0.03	-2.590	0.00971	0.093

ViscFat	*LRRC39*	1	100	18	0.037	-0.047	0.02	-2.718	0.00665	0.094
spleen	*SNX7*	1	100	138	-0.056	0.016	0.00	3.172	0.00155	0.095
liver	*DENND2D*	1	111	20	0.008	0.068	0.02	2.915	0.00362	0.096
DGNblood	*KCNA3*	1	111	63	-0.045	-0.027	0.01	-2.996	0.00279	0.094
DGNblood	*OVGP1*	1	112	32	-0.029	0.019	0.01	2.648	0.00819	0.093
TibNerve	*OVGP1*	1	112	34	-0.007	0.036	0.01	3.083	0.00209	0.095
GTExBlood	*OVGP1*	1	113	38	0.034	0.045	0.02	2.739	0.00625	0.093
pancreas	*PPM1J*	1	113	13	-0.035	0.048	0.02	2.648	0.00820	0.095
pancreas	*RHOC*	1	113	60	0.119	0.046	0.01	3.142	0.00172	0.095
DGNblood	*RNF115*	1	147	5	-0.061	0.079	0.03	2.623	0.00882	0.093
DGNblood	*FAM63A*	1	151	5	-0.016	-0.071	0.03	-2.588	0.00976	0.093
Ileum	*SNAPIN*	1	153	51	0.004	-0.029	0.01	-2.889	0.00392	0.094
DGNblood	*AQP10*	1	154	8	0.244	0.045	0.02	2.746	0.00611	0.093
DGNblood	*CREB3L4*	1	154	25	0.009	0.030	0.01	2.876	0.00409	0.094
TibNerve	*CREB3L4*	1	154	24	-0.005	0.027	0.01	2.906	0.00372	0.094
Ileum	*YY1AP1*	1	155	1	0.032	-0.440	0.15	-2.948	0.00326	0.094
TibNerve	*ARHGAP30*	1	161	13	0.001	-0.048	0.02	-2.697	0.00710	0.093
ViscFat	*RCSD1*	1	168	8	0.015	0.067	0.03	2.580	0.00999	0.093
DGNblood	*CACYBP*	1	175	18	-0.038	0.072	0.02	3.468	0.00054	0.097
TibNerve	*CACYBP*	1	175	24	-0.056	0.067	0.02	2.950	0.00323	0.094
pancreas	*RABGAP1L*	1	175	14	0.003	-0.071	0.02	-3.191	0.00145	0.101
DGNblood	*RFWD2*	1	176	10	0.024	-0.129	0.04	-3.069	0.00219	0.095
liver	*TMEM9*	1	202	50	0.023	0.042	0.01	2.870	0.00417	0.095
pancreas	*PFKFB2*	1	207	59	-0.008	0.044	0.01	3.161	0.00161	0.096
DGNblood	*G0S2*	1	210	21	-0.012	0.090	0.03	2.619	0.00893	0.093
DGNblood	*HHAT*	1	211	31	0.008	-0.074	0.03	-2.679	0.00748	0.093
DGNblood	*RCOR3*	1	211	15	-0.035	0.147	0.05	2.682	0.00740	0.093
spleen	*EFCAB2*	1	244	55	0.042	0.020	0.01	3.132	0.00178	0.095
DGNblood	*FAM49A*	2	17	56	0.000	-0.038	0.01	-2.646	0.00824	0.093
spleen	*CGREF1*	2	27	2	0.012	-0.297	0.11	-2.669	0.00771	0.093
DGNblood	*EMILIN1*	2	27	10	-0.002	0.187	0.06	3.093	0.00202	0.095
GTExBlood	*MAPRE3*	2	27	11	-0.002	-0.095	0.04	-2.678	0.00751	0.093
DGNblood	*ABHD1*	2	28	30	-0.038	-0.047	0.02	-2.698	0.00706	0.093
DGNblood	*MAPRE3*	2	28	18	-0.002	-0.065	0.02	-2.742	0.00618	0.093
skmuscle	*ATL2*	2	39	24	0.111	0.043	0.01	3.108	0.00192	0.095
ViscFat	*ATL2*	2	39	38	-0.127	-0.054	0.02	-2.792	0.00532	0.094
GTExBlood	*AC016722*	2	46	48	0.164	-0.045	0.02	-2.605	0.00929	0.093
liver	*GTF2A1L*	2	49	9	-0.020	-0.048	0.02	-2.877	0.00408	0.094
ViscFat	*GTF2A1L*	2	49	78	0.039	-0.022	0.01	-2.998	0.00277	0.095
ViscFat	*STON1-GTF2*	2	50	24	0.055	-0.037	0.01	-3.178	0.00152	0.095

ViscFat	*ACYP2*	2	54	55	0.034	0.049	0.02	2.771	0.00567	0.094
DGNblood	*B3GNT2*	2	61	18	0.002	-0.024	0.01	-2.751	0.00603	0.093
GTExBlood	*B3GNT2*	2	63	20	0.057	-0.021	0.01	-2.784	0.00545	0.094
pancreas	*HK2*	2	74	135	-0.039	0.025	0.01	2.795	0.00526	0.094
GTExBlood	*REG3G*	2	79	7	0.005	-0.113	0.04	-2.789	0.00536	0.094
GTExBlood	*IL18RAP*	2	103	34	0.033	0.024	0.01	2.870	0.00417	0.094
ViscFat	*IL1RL1*	2	103	15	-0.001	0.080	0.03	2.613	0.00907	0.094
liver	*C2orf40*	2	107	23	-0.096	-0.050	0.02	-2.721	0.00660	0.094
spleen	*RANBP2*	2	109	127	-0.024	-0.011	0.00	-2.923	0.00353	0.094
liver	*RGPD5*	2	110	55	0.002	0.040	0.01	2.726	0.00650	0.094
pancreas	*PCDP1*	2	121	29	-0.037	0.038	0.01	3.050	0.00233	0.095
pancreas	*SCTR*	2	121	11	0.048	-0.028	0.01	-2.592	0.00964	0.093
GTExBlood	*SPOPL*	2	138	11	-0.163	0.168	0.06	2.797	0.00524	0.094
GTExBlood	*DYNC1I2*	2	173	35	0.024	0.056	0.02	3.336	0.00088	0.096
DGNblood	*ITGA4*	2	182	45	-0.043	0.019	0.01	2.804	0.00513	0.094
GTExBlood	*ITGA4*	2	183	17	0.152	0.031	0.01	3.084	0.00209	0.095
pancreas	*DUSP19*	2	184	42	-0.019	0.049	0.02	2.675	0.00757	0.093
TibNerve	*TFPI*	2	188	26	0.008	0.109	0.04	2.772	0.00565	0.094
skmuscle	*TFPI*	2	189	25	0.216	-0.026	0.01	-2.668	0.00773	0.093
TibNerve	*MDH1B*	2	207	45	0.023	-0.034	0.01	-3.019	0.00258	0.095
DGNblood	*KLF7*	2	208	15	0.072	0.069	0.03	2.691	0.00722	0.093
GTExBlood	*FARSB*	2	223	8	0.036	0.125	0.04	2.950	0.00324	0.094
DGNblood	*ARMC9*	2	232	30	0.035	-0.071	0.03	-2.644	0.00830	0.093
pancreas	*ITM2C*	2	232	11	0.007	0.077	0.03	2.656	0.00799	0.095
DGNblood	*COPS8*	2	237	37	-0.020	-0.082	0.03	-3.127	0.00181	0.095
TibNerve	*OR6B3*	2	241	29	-0.011	0.029	0.01	2.660	0.00791	0.093
ViscFat	*PPP1R7*	2	241	1	-0.117	0.526	0.19	2.721	0.00660	0.093
skmuscle	*ATG4B*	2	243	11	0.015	0.085	0.03	2.609	0.00918	0.093
liver	*IL17RC*	*3*	9	*57*	*0.070*	*0.046*	*0.01*	*4.049*	*0.00005*	*0.100*
DGNblood	*BRPF1*	*3*	*10*	*10*	*0.005*	*0.255*	*0.06*	*3.972*	*0.00008*	*0.099*
DGNblood	*ATG7*	3	12	61	0.007	-0.034	0.01	-2.738	0.00626	0.093
TibNerve	*ATG7*	3	12	58	-0.058	-0.034	0.01	-2.920	0.00356	0.094
spleen	*ZFYVE20*	3	15	30	0.026	0.047	0.02	2.611	0.00913	0.093
GTExBlood	*CMC1*	3	28	38	-0.237	0.062	0.02	2.701	0.00701	0.093
ViscFat	*PDCD6IP*	3	35	74	0.023	-0.045	0.02	-2.792	0.00532	0.096
thyroid	*OXSR1*	3	39	25	-0.017	0.077	0.02	3.365	0.00079	0.096
spleen	*SNRK*	3	43	17	-0.083	-0.043	0.02	-2.655	0.00803	0.093
skmuscle	*CLEC3B*	3	45	49	0.066	0.039	0.01	2.850	0.00444	0.094
spleen	*ZNF501*	3	45	14	0.017	-0.034	0.01	-2.747	0.00611	0.093
skmuscle	*ZNF502*	3	45	12	0.234	-0.023	0.01	-2.668	0.00773	0.093

DGNblood	*ZNF502*	3	45	46	0.044	-0.012	0.00	-2.620	0.00889	0.093
spleen	*ZNF502*	3	45	22	0.240	-0.022	0.01	-2.620	0.00890	0.093
thyroid	*ZNF502*	3	45	50	0.208	-0.029	0.01	-3.128	0.00180	0.095
DGNblood	*CDCP1*	3	46	41	0.001	-0.025	0.01	-2.723	0.00655	0.093
thyroid	*C3orf18*	3	50	5	0.002	0.141	0.05	2.589	0.00974	0.093
DGNblood	*CACNA2D2*	3	50	7	0.031	0.088	0.03	3.097	0.00200	0.095
GTExBlood	*SEMA3G*	3	52	17	0.062	0.103	0.04	2.907	0.00371	0.094
ViscFat	*SMIM4*	3	52	37	0.015	0.071	0.03	2.778	0.00555	0.094
DGNblood	*ARHGEF3*	3	57	7	-0.012	0.092	0.03	2.749	0.00606	0.093
DGNblood	*CCDC66*	3	57	30	-0.018	-0.022	0.01	-2.767	0.00573	0.094
liver	*FHIT*	3	59	29	-0.033	0.052	0.02	2.653	0.00808	0.094
ViscFat	*ATXN7*	3	65	29	0.013	0.077	0.02	3.533	0.00043	0.097
skmuscle	*GSK3B*	3	119	28	0.032	-0.069	0.02	-3.005	0.00271	0.094
DGNblood	*GSK3B*	3	120	27	0.023	-0.040	0.01	-2.683	0.00739	0.093
thyroid	*PARP14*	3	121	90	0.074	-0.023	0.01	-2.732	0.00638	0.093
GTExBlood	*PARP15*	3	122	31	-0.018	0.044	0.02	2.761	0.00584	0.094
DGNblood	***ADCY5***	3	123	13	0.027	-0.087	0.03	-2.583	0.00989	0.093
thyroid	*OSBPL11*	3	124	45	-0.096	-0.051	0.02	-3.136	0.00175	0.095
TibNerve	*CNBP*	3	130	21	0.003	-0.066	0.02	-2.808	0.00506	0.094
thyroid	*NEK11*	3	130	29	-0.020	0.045	0.02	2.877	0.00408	0.094
GTExBlood	*PLXND1*	3	130	21	-0.124	-0.083	0.03	-3.117	0.00186	0.095
skmuscle	*NEK11*	3	131	15	0.042	0.054	0.02	2.900	0.00379	0.094
TibNerve	*NEK11*	3	131	15	-0.017	0.051	0.02	2.940	0.00334	0.094
spleen	*CDV3*	3	134	39	0.022	0.034	0.01	3.193	0.00144	0.095
skmuscle	*PCCB*	3	136	18	0.077	0.048	0.02	2.976	0.00297	0.094
DGNblood	*PCCB*	3	136	27	0.094	0.027	0.01	3.498	0.00048	0.097
thyroid	*PCCB*	3	136	20	0.040	0.051	0.01	3.518	0.00045	0.097
ViscFat	*PCCB*	3	136	22	-0.030	0.070	0.03	2.690	0.00724	0.093
liver	*IL20RB*	3	137	13	0.003	0.058	0.02	3.034	0.00246	0.095
GTExBlood	*IL20RB*	3	137	11	0.022	0.082	0.03	2.674	0.00759	0.093
skmuscle	*RBP1*	3	139	7	0.094	-0.054	0.02	-2.856	0.00436	0.094
thyroid	*RBP1*	3	140	18	0.046	-0.078	0.03	-3.018	0.00259	0.095
GTExBlood	*AADAC*	3	151	24	-0.037	0.085	0.03	2.747	0.00610	0.093
TibNerve	*P2RY1*	3	152	47	-0.020	-0.042	0.01	-2.887	0.00396	0.094
TibNerve	*DNAJB11*	3	185	41	0.209	-0.049	0.02	-2.788	0.00538	0.094
skmuscle	*DLG1*	3	197	33	0.064	-0.058	0.02	-3.014	0.00263	0.095
thyroid	*PPP2R2C*	4	7	2	-0.141	0.126	0.04	2.913	0.00364	0.094
ViscFat	*SLC2A9*	4	10	35	-0.021	0.046	0.02	2.833	0.00469	0.094
thyroid	*NCAPG*	4	17	76	-0.001	0.044	0.01	3.246	0.00120	0.096
thyroid	*FAM47E*	4	77	21	0.010	0.052	0.02	2.705	0.00691	0.093

Ileum	*CDS1*	4	85	107	-0.004	0.011	0.00	2.870	0.00418	0.094
DGNblood	*KLHL8*	4	88	56	-0.016	0.047	0.02	3.021	0.00256	0.095
skmuscle	*TMEM154*	4	154	5	0.018	-0.125	0.05	-2.643	0.00831	0.093
skmuscle	*CTSO*	4	157	31	0.042	-0.078	0.03	-2.706	0.00690	0.093
DGNblood	*GUCY1B3*	4	157	21	-0.101	-0.102	0.04	-2.905	0.00373	0.094
pancreas	*CTSO*	4	158	35	-0.035	-0.043	0.02	-2.609	0.00918	0.093
pancreas	*NPY1R*	4	164	39	0.009	0.052	0.02	2.762	0.00582	0.094
pancreas	*NPY5R*	4	164	19	0.000	0.072	0.02	3.249	0.00119	0.096
GTExBlood	*CDH6*	5	32	9	0.034	0.102	0.04	2.898	0.00382	0.094
pancreas	*HCN1*	5	46	29	0.105	0.044	0.01	3.284	0.00105	0.096
TibNerve	*EMB*	5	50	58	0.118	-0.040	0.01	-2.764	0.00579	0.094
thyroid	*TRAPPC13*	5	65	2	0.033	-0.178	0.06	-3.003	0.00272	0.094
ViscFat	*ANKRA2*	5	73	70	0.030	-0.041	0.02	-2.732	0.00639	0.094
TibNerve	*ANKDD1B*	5	75	53	-0.003	-0.025	0.01	-2.730	0.00642	0.093
GTExBlood	*SRP19*	5	113	15	-0.010	0.112	0.04	2.862	0.00428	0.094
thyroid	*CDKN2AIPNL*	5	134	5	0.420	-0.058	0.02	-2.644	0.00829	0.093
DGNblood	*PKD2L2*	5	137	25	-0.034	-0.043	0.01	-2.896	0.00385	0.094
skmuscle	*IK*	5	139	33	0.031	-0.061	0.02	-2.611	0.00913	0.093
skmuscle	*CXXC5*	5	140	59	0.036	0.035	0.01	2.873	0.00413	0.094
GTExBlood	*RNF14*	5	141	8	-0.020	-0.102	0.04	-2.702	0.00699	0.093
skmuscle	*NDFIP1*	5	142	47	0.229	0.022	0.01	2.859	0.00432	0.094
DGNblood	*NDFIP1*	5	142	56	0.009	-0.033	0.01	-3.138	0.00174	0.095
TibNerve	*PCDHB15*	5	142	11	-0.047	-0.049	0.02	-2.799	0.00521	0.094
skmuscle	*STK32A*	5	147	3	0.012	-0.138	0.04	-3.147	0.00168	0.095
ViscFat	*MFAP3*	5	153	38	0.037	0.036	0.01	2.685	0.00734	0.094
skmuscle	*FABP6*	5	160	17	0.050	0.064	0.02	2.909	0.00369	0.094
TibNerve	*CCNJL*	5	161	52	0.311	0.041	0.02	2.698	0.00707	0.093
thyroid	*NPM1*	5	171	7	0.001	-0.113	0.04	-2.664	0.00782	0.093
skmuscle	*DOK3*	5	177	47	0.050	0.050	0.02	3.230	0.00127	0.095
pancreas	*RGS14*	5	177	8	-0.156	0.067	0.02	2.884	0.00399	0.097
ViscFat	*TRIM7*	5	181	76	0.018	-0.029	0.01	-2.618	0.00894	0.093
pancreas	*DUSP22*	6	1	38	0.008	-0.062	0.02	-3.018	0.00260	0.096
pancreas	*GMDS*	6	1	2	-0.106	-0.750	0.26	-2.923	0.00353	0.095
DGNblood	*PSMG4*	6	3	39	0.030	-0.020	0.01	-2.965	0.00308	0.094
Ileum	*PPP1R3G*	6	6	63	0.093	0.016	0.01	2.720	0.00661	0.093
skmuscle	*SERAC1*	6	16	21	0.106	0.047	0.01	3.304	0.00098	0.096
GTExBlood	*GABBR1*	*6*	29	*48*	*0.001*	*0.156*	*0.04*	*4.208*	*0.00003*	*0.100*
DGNblood	*FLOT1*	6	31	59	0.043	0.023	0.01	2.595	0.00958	0.093
TibNerve	*ATP6V1G2*	6	32	26	0.141	-0.061	0.02	-2.698	0.00707	0.093
thyroid	*C4A*	6	32	120	0.002	-0.020	0.01	-3.074	0.00215	0.095

GTExBlood	*C4B*	6	32	49	0.009	0.030	0.01	2.725	0.00651	0.093
GTExBlood	*CYP21A2*	6	32	25	0.017	0.091	0.03	2.705	0.00692	0.093
DGNblood	*DHX16*	6	32	2	-0.005	-0.513	0.17	-2.994	0.00280	0.094
skmuscle	*HLA-DQB2*	6	32	65	0.423	0.020	0.01	3.111	0.00191	0.095
skmuscle	*PSMB9*	6	32	40	0.081	0.064	0.02	2.865	0.00424	0.094
DGNblood	*PSMB9*	6	32	38	0.021	0.016	0.01	2.615	0.00902	0.093
DGNblood	*C2*	6	33	45	0.102	0.030	0.01	3.135	0.00176	0.095
thyroid	*HLA-DMB*	6	33	8	-0.055	-0.155	0.05	-3.115	0.00188	0.095
thyroid	*HLA-DQB2*	6	33	37	0.009	0.023	0.01	3.000	0.00275	0.094
thyroid	*HLA-DRA*	6	33	35	0.073	-0.060	0.02	-2.673	0.00762	0.093
thyroid	*PSMB9*	6	33	39	0.021	0.088	0.03	2.710	0.00681	0.093
thyroid	*RNF5*	6	33	21	-0.089	-0.077	0.03	-2.822	0.00484	0.094
thyroid	*ABCC10*	6	44	30	0.018	0.058	0.02	2.754	0.00596	0.093
Ileum	*ENPP4*	6	46	61	0.089	-0.017	0.01	-2.690	0.00723	0.093
Ileum	*BEND6*	6	57	16	-0.029	-0.039	0.01	-2.785	0.00543	0.094
DGNblood	*ME1*	6	84	24	-0.010	-0.080	0.03	-2.602	0.00937	0.093
TibNerve	*NT5E*	6	86	19	-0.036	0.045	0.02	2.622	0.00884	0.093
GTExBlood	*C6orf164*	6	88	32	-0.005	0.052	0.02	2.754	0.00597	0.093
GTExBlood	*RP1-102*	6	88	13	-0.026	0.112	0.04	2.683	0.00738	0.093
thyroid	*SLC35A1*	6	88	43	-0.015	-0.017	0.01	-2.614	0.00906	0.093
TibNerve	*SLC35A1*	6	88	26	-0.010	-0.025	0.01	-2.772	0.00566	0.094
ViscFat	*SLC35A1*	6	88	26	-0.005	-0.052	0.02	-3.230	0.00127	0.097
GTExBlood	*ZNF292*	6	88	44	0.006	-0.047	0.02	-2.597	0.00952	0.093
thyroid	*RP1-102*	6	89	30	-0.020	0.068	0.02	3.317	0.00093	0.096
pancreas	*NDUFAF4*	6	97	12	0.096	0.083	0.03	2.784	0.00545	0.095
TibNerve	*MMS22L*	6	98	29	-0.064	-0.056	0.02	-2.712	0.00678	0.093
pancreas	*FAM26F*	6	117	5	0.192	0.059	0.02	2.985	0.00289	0.096
skmuscle	*RAET1E*	6	151	12	0.059	0.055	0.02	2.754	0.00597	0.093
thyroid	*SYNJ2*	6	158	16	-0.003	-0.046	0.01	-3.166	0.00158	0.095
spleen	*SYTL3*	6	159	1	0.069	0.296	0.11	2.644	0.00829	0.093
ViscFat	*SLC22A3*	6	161	14	0.020	-0.098	0.03	-3.147	0.00169	0.097
TibNerve	*AC187652*	7	0	20	0.141	-0.035	0.01	-2.580	0.00999	0.093
spleen	*FTSJ2*	7	3	20	0.075	0.033	0.01	2.591	0.00967	0.093
DGNblood	*NUDT1*	7	3	52	-0.050	-0.044	0.02	-2.683	0.00738	0.093
skmuscle	*AC073343*	7	6	7	0.019	-0.140	0.05	-2.870	0.00417	0.094
spleen	*CCZ1*	7	6	25	-0.205	0.028	0.01	2.917	0.00360	0.094
TibNerve	*ICA1*	7	7	84	0.000	0.035	0.01	2.759	0.00588	0.094
DGNblood	*ICA1*	7	9	63	-0.016	-0.033	0.01	-2.708	0.00685	0.093
TibNerve	*FAM188B*	7	30	13	0.007	-0.043	0.02	-2.661	0.00789	0.093
ViscFat	*AMPH*	7	39	21	0.039	0.041	0.01	3.059	0.00227	0.097

TibNerve	*MRPL32*	7	42	55	-0.010	0.025	0.01	2.681	0.00743	0.093
GTExBlood	*BLVRA*	7	43	26	0.006	-0.035	0.01	-2.679	0.00748	0.093
DGNblood	*C7orf44*	7	44	41	0.034	0.024	0.01	3.452	0.00057	0.096
thyroid	*POLD2*	7	44	58	-0.063	-0.038	0.01	-3.092	0.00203	0.095
DGNblood	*STK17A*	7	44	7	-0.001	0.255	0.09	2.748	0.00608	0.093
spleen	*SPDYE1*	7	45	21	-0.008	0.036	0.01	2.651	0.00813	0.093
thyroid	*LANCL2*	7	55	20	-0.034	-0.099	0.03	-3.683	0.00024	0.098
skmuscle	*VOPP1*	7	55	8	0.019	0.114	0.04	2.841	0.00457	0.094
ViscFat	*GUSB*	7	65	1	-0.018	-2.072	0.67	-3.074	0.00215	0.095
pancreas	*FGL2*	7	76	36	0.059	-0.062	0.02	-2.981	0.00293	0.095
ViscFat	*TMEM60*	7	78	26	0.044	0.050	0.02	2.893	0.00388	0.094
GTExBlood	*SRI*	7	89	40	-0.002	-0.057	0.02	-2.699	0.00705	0.093
ViscFat	*LRRD1*	7	92	27	0.011	-0.056	0.02	-2.642	0.00834	0.093
DGNblood	*CDK6*	7	93	11	0.016	0.190	0.07	2.904	0.00375	0.094
TibNerve	*KCND2*	7	120	63	0.014	-0.050	0.01	-3.635	0.00029	0.097
DGNblood	*FSCN3*	7	127	26	0.006	0.058	0.02	2.838	0.00461	0.094
skmuscle	*FAM71F2*	7	128	13	0.055	-0.110	0.04	-3.030	0.00249	0.095
DGNblood	*FAM71F2*	7	128	9	0.158	-0.038	0.01	-3.165	0.00159	0.095
GTExBlood	*FAM71F2*	7	128	11	0.000	-0.140	0.05	-3.100	0.00198	0.095
TibNerve	*PRRT4*	7	128	2	0.060	-0.261	0.10	-2.657	0.00798	0.093
TibNerve	*FAM71F2*	7	129	24	-0.016	-0.052	0.01	-3.580	0.00036	0.097
ViscFat	*SMKR1*	7	130	52	-0.064	0.055	0.02	3.221	0.00131	0.096
thyroid	*KLRG2*	7	139	57	-0.006	0.027	0.01	3.117	0.00186	0.095
thyroid	*FAM115C*	7	142	29	-0.021	-0.034	0.01	-2.852	0.00442	0.094
GTExBlood	*FAM115C*	7	143	12	-0.002	-0.049	0.02	-2.709	0.00683	0.093
skmuscle	*ACTR3C*	7	150	39	0.218	0.022	0.01	2.770	0.00568	0.094
liver	*AOC1*	7	150	45	-0.117	-0.025	0.01	-2.857	0.00435	0.095
GTExBlood	*GIMAP1*	7	150	21	-0.022	0.056	0.02	2.900	0.00380	0.094
skmuscle	*GBX1*	7	151	12	0.105	0.044	0.02	2.643	0.00831	0.093
TibNerve	*RARRES2*	7	151	60	0.004	0.025	0.01	2.696	0.00712	0.093
pancreas	*ACTR3B*	7	152	37	-0.085	0.061	0.02	2.811	0.00502	0.094
skmuscle	*CDK5*	7	152	24	0.009	0.077	0.02	3.126	0.00181	0.095
ViscFat	*CSGALNACT1*	8	19	5	0.038	0.135	0.04	3.529	0.00043	0.101
ViscFat	*PHYHIP*	8	22	11	-0.007	-0.081	0.03	-2.779	0.00553	0.095
DGNblood	*C8orf80*	8	28	6	-0.019	0.318	0.11	2.939	0.00335	0.094
skmuscle	*NSMAF*	8	60	1	0.011	-1.745	0.65	-2.691	0.00722	0.093
skmuscle	*LACTB2*	8	71	26	0.107	0.037	0.01	2.613	0.00907	0.093
TibNerve	*LACTB2*	8	71	26	-0.008	0.048	0.02	3.026	0.00253	0.095
DGNblood	*LACTB2*	8	72	19	0.019	0.051	0.02	2.842	0.00455	0.094
skmuscle	*XKR9*	8	72	60	0.479	0.019	0.01	2.966	0.00307	0.094

DGNblood	*XKR9*	8	72	43	0.009	0.019	0.01	3.255	0.00117	0.096
Ileum	*XKR9*	8	72	85	0.003	0.024	0.01	2.894	0.00387	0.094
spleen	*XKR9*	8	72	72	0.003	0.025	0.01	2.762	0.00583	0.094
thyroid	*XKR9*	8	72	54	0.025	0.017	0.01	3.176	0.00153	0.095
TibNerve	*XKR9*	8	72	86	0.002	0.017	0.01	3.323	0.00091	0.096
thyroid	*TMEM70*	8	74	6	-0.016	-0.093	0.04	-2.602	0.00937	0.093
spleen	*LRRCC1*	8	85	40	0.030	-0.027	0.01	-3.139	0.00173	0.095
thyroid	*RP11-122*	8	92	41	-0.064	0.043	0.01	3.125	0.00182	0.095
thyroid	*C8orf37*	8	96	28	0.003	-0.052	0.02	-2.582	0.00994	0.093
GTExBlood	*KIAA1429*	8	96	66	-0.019	-0.043	0.01	-2.921	0.00355	0.094
skmuscle	*SNX31*	8	101	49	0.243	-0.023	0.01	-2.858	0.00433	0.094
Ileum	*SNX31*	8	102	82	0.108	-0.017	0.00	-3.657	0.00027	0.097
thyroid	*SNX31*	8	102	78	-0.073	-0.017	0.01	-2.914	0.00363	0.094
TibNerve	*SNX31*	8	102	37	0.000	-0.017	0.01	-2.984	0.00290	0.094
TibNerve	*CTHRC1*	8	104	27	0.054	-0.057	0.02	-3.484	0.00051	0.097
DGNblood	*SLC25A32*	8	104	29	0.013	-0.037	0.01	-2.710	0.00681	0.093
spleen	*ENY2*	8	109	13	0.004	0.048	0.02	3.090	0.00204	0.095
ViscFat	*ATAD2*	8	124	25	0.118	-0.056	0.02	-2.606	0.00926	0.093
pancreas	*GPIHBP1*	8	144	16	-0.052	-0.058	0.02	-2.707	0.00688	0.094
pancreas	*TIGD5*	8	144	15	0.052	-0.074	0.03	-2.955	0.00318	0.094
skmuscle	*CREB3*	9	36	16	0.023	-0.075	0.03	-2.957	0.00316	0.094
thyroid	*CNTNAP3*	9	38	92	0.011	-0.025	0.01	-2.896	0.00384	0.094
thyroid	*PTPDC1*	9	97	6	0.018	-0.126	0.05	-2.701	0.00701	0.093
thyroid	*ALDOB*	9	105	25	0.198	-0.043	0.02	-2.585	0.00985	0.093
GTExBlood	*GRIN3A*	9	105	21	0.076	-0.077	0.03	-2.751	0.00602	0.093
skmuscle	*LAMC3*	9	133	84	0.110	0.028	0.01	2.655	0.00803	0.093
pancreas	*LAMC3*	9	134	48	-0.018	0.055	0.02	3.036	0.00244	0.095
thyroid	*UBAC1*	9	139	19	0.039	0.064	0.02	2.865	0.00424	0.094
TibNerve	*DIP2C*	10	1	78	-0.102	-0.034	0.01	-2.904	0.00374	0.094
TibNerve	*WDR37*	10	1	27	0.135	-0.022	0.01	-2.781	0.00549	0.094
DGNblood	*IL2RA*	10	5	57	0.024	0.050	0.02	2.601	0.00939	0.093
pancreas	*ITGA8*	10	15	28	0.020	-0.069	0.02	-3.156	0.00164	0.095
thyroid	*RAB18*	10	28	26	-0.009	0.058	0.02	2.716	0.00670	0.093
DGNblood	*FZD8*	10	36	8	0.027	0.112	0.03	3.336	0.00088	0.096
DGNblood	*PARG*	10	50	17	0.011	0.040	0.01	2.688	0.00728	0.093
DGNblood	*TFAM*	10	59	23	-0.032	0.018	0.01	2.702	0.00698	0.093
GTExBlood	*TFAM*	10	60	40	-0.004	0.034	0.01	2.981	0.00293	0.094
DGNblood	*HK1*	10	71	31	-0.043	0.057	0.02	2.853	0.00440	0.094
pancreas	*PALD1*	10	72	5	0.590	-0.038	0.01	-3.003	0.00272	0.096
thyroid	*ZCCHC24*	10	81	52	0.189	-0.039	0.01	-2.697	0.00708	0.093

thyroid	*FAM213A*	10	82	18	-0.035	0.058	0.02	2.986	0.00288	0.094
GTExBlood	*TSPAN14*	10	82	24	-0.013	0.059	0.02	2.704	0.00695	0.093
skmuscle	*OPN4*	10	88	141	0.353	-0.017	0.01	-2.888	0.00395	0.094
TibNerve	*NUTM2D*	10	90	86	-0.010	0.023	0.01	2.920	0.00357	0.094
skmuscle	*R3HCC1L*	10	100	32	0.007	0.081	0.03	2.831	0.00471	0.094
TibNerve	*R3HCC1L*	10	100	13	0.020	-0.087	0.03	-2.778	0.00555	0.094
DGNblood	*COX15*	10	101	93	0.000	-0.027	0.01	-2.967	0.00306	0.094
TibNerve	*LZTS2*	10	102	13	-0.116	0.062	0.02	2.625	0.00876	0.093
thyroid	*LHPP*	10	126	55	-0.058	0.025	0.01	2.719	0.00664	0.093
pancreas	*PWWP2B*	10	134	10	-0.004	-0.087	0.03	-2.833	0.00468	0.095
spleen	*PRAP1*	10	135	25	0.036	-0.041	0.01	-3.050	0.00234	0.095
ViscFat	*LRRC56*	11	0	22	-0.003	-0.049	0.02	-2.689	0.00725	0.093
GTExBlood	*B4GALNT4*	11	1	16	-0.097	-0.093	0.03	-2.812	0.00500	0.094
thyroid	*KCNQ1*	11	3	27	0.049	0.058	0.02	2.940	0.00334	0.094
liver	*PGAP2*	11	4	74	0.000	0.031	0.01	2.978	0.00296	0.094
spleen	*PPFIBP2*	11	8	73	-0.048	-0.023	0.01	-2.593	0.00961	0.093
skmuscle	*RPL27A*	11	8	11	0.008	0.093	0.04	2.594	0.00959	0.093
DGNblood	*SOX6*	11	15	1	-0.007	2.239	0.75	3.002	0.00274	0.094
pancreas	*ARL14EP*	11	30	23	0.002	0.041	0.01	2.760	0.00587	0.094
GTExBlood	*ABTB2*	11	34	47	0.002	0.028	0.01	2.615	0.00901	0.093
GTExBlood	*CAT*	11	34	35	0.030	0.019	0.01	2.684	0.00738	0.093
DGNblood	*CD44*	11	35	8	-0.018	-0.069	0.02	-2.864	0.00425	0.094
Ileum	*SLC35C1*	11	45	31	-0.009	0.026	0.01	2.625	0.00877	0.093
Ileum	*CKAP5*	11	46	19	-0.019	-0.033	0.01	-2.674	0.00759	0.093
DGNblood	***MADD***	11	46	54	-0.101	-0.024	0.01	-2.658	0.00796	0.093
DGNblood	***MAPK8IP1***	*11*	46	*22*	*0.005*	*0.021*	*0.01*	*4.113*	*0.00004*	*0.100*
GTExBlood	***MAPK8IP1***	*11*	46	*17*	*0.027*	*0.048*	*0.01*	*4.075*	*0.00005*	*0.100*
thyroid	***MADD***	11	47	17	-0.003	-0.032	0.01	-2.747	0.00610	0.093
TibNerve	***MADD***	11	47	37	0.008	-0.067	0.03	-2.682	0.00741	0.093
skmuscle	*MTCH2*	11	47	40	0.036	-0.041	0.01	-2.735	0.00632	0.093
thyroid	***NR1H3***	11	47	5	-0.073	-0.174	0.06	-2.930	0.00345	0.094
pancreas	*PTPRJ*	11	47	22	0.112	0.094	0.03	3.097	0.00200	0.095
Ileum	***C1QTNF4***	11	48	34	0.105	0.021	0.01	2.678	0.00749	0.093
liver	*PRG2*	11	57	54	-0.103	0.031	0.01	2.984	0.00290	0.095
DGNblood	*TNKS1BP1*	11	58	33	-0.005	-0.042	0.02	-2.603	0.00936	0.093
DGNblood	*VWCE*	11	60	14	0.033	-0.153	0.05	-3.052	0.00232	0.095
ViscFat	*SCGB2A2*	11	63	1	-0.031	0.777	0.26	2.992	0.00283	0.095
thyroid	*BATF2*	11	66	34	-0.065	-0.057	0.02	-2.784	0.00545	0.094
GTExBlood	*RBM14*	11	66	13	-0.004	0.111	0.04	2.603	0.00935	0.093
thyroid	*MAP6*	11	75	10	-0.023	-0.070	0.03	-2.594	0.00958	0.093

skmuscle	*MAP6*	11	76	31	0.067	-0.042	0.02	-2.761	0.00585	0.094
DGNblood	*PAK1*	11	76	35	0.022	0.044	0.02	2.678	0.00749	0.093
TibNerve	*TMEM126A*	11	85	15	-0.006	-0.051	0.02	-2.729	0.00645	0.093
thyroid	*FAT3*	11	92	39	-0.321	0.060	0.02	2.626	0.00874	0.093
spleen	*MRE11A*	11	93	75	-0.092	-0.013	0.00	-2.717	0.00667	0.093
DGNblood	*ANKRD49*	11	94	7	-0.105	-0.168	0.06	-2.729	0.00643	0.093
DGNblood	*ENDOD1*	11	94	11	0.003	-0.039	0.01	-2.659	0.00794	0.093
thyroid	*FOLR4*	11	94	57	0.089	-0.055	0.02	-3.166	0.00158	0.095
skmuscle	*MRE11A*	11	94	36	0.177	-0.025	0.01	-2.772	0.00566	0.094
Ileum	*MRE11A*	11	94	43	0.009	-0.017	0.01	-2.886	0.00396	0.094
thyroid	*MRE11A*	11	94	14	-0.119	-0.014	0.01	-2.628	0.00869	0.093
TibNerve	*MRE11A*	11	94	12	0.012	-0.013	0.01	-2.604	0.00931	0.093
GTExBlood	*MRE11A*	11	95	55	-0.006	-0.018	0.01	-2.584	0.00986	0.093
DGNblood	*ARHGEF12*	11	120	35	-0.028	0.096	0.03	3.184	0.00149	0.095
DGNblood	*TMEM136*	11	120	12	0.020	0.119	0.04	2.847	0.00448	0.094
TibNerve	*TMEM136*	*11*	120	*12*	*-0.034*	*0.041*	*0.01*	*4.055*	*0.00005*	*0.100*
thyroid	*TMEM136*	11	121	27	0.059	0.056	0.02	2.862	0.00428	0.094
TibNerve	*SRPR*	11	127	18	-0.010	-0.050	0.02	-2.596	0.00953	0.093
pancreas	*LTBR*	12	7	12	-0.061	-0.048	0.02	-2.704	0.00695	0.094
ViscFat	*LTBR*	12	7	21	-0.102	-0.051	0.01	-3.519	0.00045	0.097
skmuscle	*CLEC1A*	12	9	82	0.026	0.046	0.01	3.378	0.00075	0.096
DGNblood	*GABARAPL1*	12	10	31	-0.024	0.081	0.03	2.880	0.00404	0.094
GTExBlood	*GABARAPL1*	12	10	30	0.018	0.063	0.02	2.595	0.00957	0.093
spleen	*KLRC2*	12	10	31	0.001	-0.028	0.01	-2.656	0.00801	0.093
TibNerve	*OLR1*	12	10	15	-0.007	0.087	0.03	3.213	0.00134	0.095
Ileum	*KLRC2*	12	11	18	-0.002	-0.028	0.01	-2.636	0.00849	0.093
TibNerve	*TAS2R13*	12	11	39	-0.005	-0.039	0.01	-2.826	0.00479	0.094
DGNblood	*ETV6*	12	12	18	-0.010	0.065	0.02	2.682	0.00741	0.093
DGNblood	*CASC1*	12	25	33	0.030	-0.026	0.01	-3.013	0.00263	0.095
GTExBlood	*KRAS*	12	25	12	-0.007	0.098	0.03	2.848	0.00447	0.094
DGNblood	*LRMP*	12	25	4	0.011	0.285	0.11	2.682	0.00740	0.093
thyroid	*TMTC1*	12	30	2	-0.003	0.285	0.10	2.718	0.00666	0.093
pancreas	*BICD1*	12	31	26	-0.018	0.066	0.02	3.320	0.00093	0.096
DGNblood	*H3F3C*	12	33	6	-0.011	0.443	0.15	2.997	0.00278	0.094
skmuscle	*ANO6*	12	46	25	0.020	-0.067	0.02	-2.726	0.00651	0.093
skmuscle	*RAPGEF3*	12	47	32	0.109	0.033	0.01	2.713	0.00676	0.093
skmuscle	*SLC38A4*	12	47	18	0.020	0.108	0.04	2.727	0.00648	0.093
thyroid	*RAPGEF3*	12	48	34	-0.002	0.024	0.01	2.665	0.00780	0.093
ViscFat	*RAPGEF3*	12	48	51	-0.026	0.039	0.01	2.633	0.00857	0.096
DGNblood	*CALCOCO1*	12	55	27	0.013	0.080	0.03	2.645	0.00826	0.093

TibNerve	*ATP5B*	12	57	1	-0.029	1.990	0.71	2.792	0.00531	0.094
Ileum	*LLPH*	12	66	116	-0.009	-0.015	0.01	-2.719	0.00664	0.093
Ileum	*SLC6A15*	12	85	45	0.003	-0.024	0.01	-2.878	0.00406	0.094
TibNerve	*NTN4*	12	95	39	0.084	0.046	0.02	2.694	0.00715	0.093
DGNblood	*GLT8D2*	12	103	13	0.005	-0.077	0.03	-2.739	0.00625	0.093
DGNblood	*HVCN1*	12	111	7	-0.041	-0.075	0.02	-3.006	0.00270	0.094
DGNblood	*PPP1CC*	12	111	22	0.002	0.043	0.02	2.779	0.00552	0.094
DGNblood	*RBM19*	12	115	44	-0.053	-0.039	0.02	-2.580	0.00999	0.093
thyroid	*MLEC*	12	121	47	0.048	0.032	0.01	2.768	0.00573	0.094
thyroid	*P2RX4*	12	121	11	0.003	-0.034	0.01	-2.588	0.00977	0.093
ViscFat	*TRIAP1*	12	122	42	-0.016	0.047	0.02	2.714	0.00673	0.093
DGNblood	*ZCCHC8*	12	123	5	0.008	-0.354	0.12	-2.897	0.00384	0.094
TibNerve	*AC226150*	12	133	13	0.000	-0.106	0.04	-2.935	0.00339	0.094
Ileum	*ANKLE2*	12	133	16	-0.039	0.026	0.01	2.848	0.00447	0.094
GTExBlood	*ANKLE2*	12	133	37	-0.098	0.026	0.01	3.299	0.00100	0.096
Ileum	*ZNF605*	12	133	10	0.441	0.035	0.01	2.695	0.00714	0.093
ViscFat	*ZNF605*	12	133	33	-0.002	0.045	0.02	2.818	0.00490	0.094
DGNblood	*ZNF84*	12	133	27	-0.007	0.043	0.01	3.255	0.00116	0.096
liver	*ZNF84*	12	133	15	0.036	0.040	0.01	2.680	0.00745	0.094
thyroid	*P2RX2*	12	134	14	0.085	0.067	0.03	2.594	0.00960	0.093
skmuscle	*ZNF268*	12	134	9	0.014	-0.105	0.03	-3.220	0.00131	0.095
DGNblood	*ALOX5AP*	13	31	60	-0.126	-0.022	0.01	-3.063	0.00224	0.095
GTExBlood	*ALOX5AP*	13	31	12	-0.064	-0.056	0.02	-3.097	0.00199	0.095
pancreas	*GTF2F2*	13	46	48	0.004	0.056	0.02	2.604	0.00931	0.093
skmuscle	*GPR180*	13	95	66	0.509	0.014	0.01	2.716	0.00670	0.093
thyroid	*GPR180*	13	95	37	0.044	0.024	0.01	2.651	0.00813	0.093
GTExBlood	*FARP1*	13	99	7	0.037	-0.082	0.03	-2.581	0.00996	0.093
pancreas	*GPR18*	13	99	18	0.070	-0.156	0.06	-2.800	0.00519	0.094
TibNerve	*LIG4*	13	109	12	-0.016	-0.058	0.02	-2.930	0.00344	0.094
thyroid	*PROZ*	13	113	11	-0.054	0.030	0.01	2.662	0.00786	0.093
pancreas	*PROZ*	13	114	14	0.023	0.037	0.01	2.860	0.00430	0.094
TibNerve	*PNP*	14	21	129	0.011	-0.016	0.01	-2.735	0.00633	0.093
ViscFat	*DAD1*	14	23	55	-0.004	-0.143	0.04	-3.434	0.00061	0.096
pancreas	*MDP1*	14	26	20	0.030	0.081	0.03	2.721	0.00660	0.093
TibNerve	*BAZ1A*	14	34	17	-0.053	-0.068	0.03	-2.679	0.00747	0.093
thyroid	*MIPOL1*	14	39	44	0.020	0.050	0.02	3.193	0.00144	0.095
GTExBlood	*AL139099*	14	49	7	-0.070	-0.134	0.05	-2.897	0.00383	0.094
DGNblood	*ESR2*	14	64	34	-0.009	-0.024	0.01	-3.559	0.00039	0.097
skmuscle	*ESR2*	14	65	8	0.035	-0.089	0.03	-2.801	0.00516	0.094
GTExBlood	*ESR2*	14	65	9	-0.022	-0.108	0.03	-3.558	0.00039	0.097

DGNblood	*FNTB*	14	65	132	0.175	0.015	0.01	2.879	0.00406	0.094
spleen	*MAX*	14	66	63	-0.095	-0.021	0.01	-3.095	0.00201	0.095
pancreas	*SLC39A9*	14	69	56	-0.028	0.052	0.02	3.211	0.00135	0.095
pancreas	*SYNJ2BP*	14	71	46	0.246	0.060	0.02	2.754	0.00596	0.094
skmuscle	*PTGR2*	14	74	34	0.149	0.030	0.01	2.673	0.00761	0.093
thyroid	*PTGR2*	14	74	35	0.011	0.019	0.01	2.645	0.00827	0.093
skmuscle	*ALDH6A1*	14	75	19	0.035	-0.058	0.02	-2.761	0.00585	0.094
pancreas	*ISCA2*	14	76	1	0.022	-1.171	0.42	-2.806	0.00510	0.094
TibNerve	*GPATCH2L*	14	77	38	0.038	-0.025	0.01	-2.623	0.00881	0.093
pancreas	*FOXN3*	14	90	2	-0.065	-0.195	0.07	-2.849	0.00446	0.094
skmuscle	*IFI27L1*	14	95	51	0.332	-0.018	0.01	-2.661	0.00789	0.093
TibNerve	*IFI27L1*	14	95	19	-0.004	-0.070	0.03	-2.698	0.00706	0.093
ViscFat	*IFI27L1*	14	95	17	0.039	-0.076	0.03	-2.927	0.00348	0.094
liver	*TCL1A*	14	96	1	-0.018	1.052	0.38	2.740	0.00622	0.094
DGNblood	*TCL1B*	14	96	23	0.098	-0.020	0.01	-2.711	0.00679	0.093
DGNblood	*ANKRD9*	14	103	19	0.000	-0.028	0.01	-2.714	0.00673	0.093
DGNblood	*CINP*	14	103	15	-0.170	0.033	0.01	2.604	0.00931	0.093
spleen	*CINP*	14	103	5	-0.006	0.053	0.02	2.889	0.00393	0.094
thyroid	*CINP*	14	103	9	0.000	0.043	0.02	2.853	0.00441	0.094
TibNerve	*CINP*	14	103	8	-0.132	0.038	0.01	2.995	0.00280	0.094
GTExBlood	*CINP*	14	103	9	-0.045	0.063	0.02	2.965	0.00308	0.094
skmuscle	*CINP*	14	104	85	0.132	0.025	0.01	2.820	0.00488	0.094
thyroid	*ASPG*	14	105	95	-0.010	0.033	0.01	2.854	0.00438	0.094
DGNblood	*PLD4*	14	105	32	-0.035	-0.026	0.01	-3.010	0.00266	0.095
GTExBlood	*PLD4*	14	105	3	-0.031	-0.215	0.08	-2.724	0.00654	0.093
skmuscle	*LRRC57*	15	43	4	0.011	0.725	0.25	2.851	0.00443	0.094
spleen	*LRRC57*	15	43	16	0.043	0.045	0.02	2.792	0.00531	0.094
DGNblood	*UBR1*	15	43	11	-0.043	-0.161	0.05	-3.142	0.00172	0.095
TibNerve	*SPPL2A*	15	51	36	-0.077	0.037	0.01	3.250	0.00118	0.096
DGNblood	*LYSMD2*	15	51	2	0.000	-2.368	0.79	-3.014	0.00263	0.095
skmuscle	*SPPL2A*	15	51	32	0.121	0.040	0.01	3.356	0.00081	0.096
DGNblood	*SPPL2A*	*15*	51	*33*	*0.024*	*0.052*	*0.01*	*4.402*	*0.00001*	*0.102*
DGNblood	*TRPM7*	15	51	15	0.001	0.187	0.05	3.833	0.00013	0.098
DGNblood	*RORA*	15	62	16	0.025	-0.143	0.06	-2.582	0.00993	0.093
TibNerve	*C2CD4A*	15	63	83	-0.003	0.035	0.01	3.279	0.00107	0.096
GTExBlood	*ANKDD1A*	15	65	11	-0.001	-0.040	0.01	-2.701	0.00700	0.093
thyroid	*MAP2K5*	15	67	18	0.102	0.032	0.01	3.487	0.00051	0.097
thyroid	*SKOR1*	15	67	19	-0.065	0.038	0.01	2.927	0.00348	0.094
DGNblood	*MAP2K5*	15	68	66	-0.005	0.020	0.01	3.359	0.00080	0.096
TibNerve	*MAP2K5*	*15*	68	*18*	*-0.010*	*0.038*	*0.01*	*3.825*	*0.00014*	*0.098*

GTExBlood	*MAP2K5*	15	68	20	-0.043	0.043	0.01	3.782	0.00016	0.098
skmuscle	*SKOR1*	*15*	*68*	*16*	*0.054*	*0.073*	*0.02*	*3.527*	*0.00044*	*0.097*
DGNblood	*SKOR1*	15	68	25	-0.120	0.052	0.02	3.390	0.00072	0.096
TibNerve	*SKOR1*	*15*	68	*13*	*0.016*	*0.060*	*0.02*	*3.912*	*0.00010*	*0.099*
ViscFat	*SKOR1*	15	68	10	-0.061	0.091	0.03	2.757	0.00591	0.096
spleen	*MAP2K5*	15	69	64	-0.016	0.022	0.01	2.679	0.00748	0.093
DGNblood	*KIF23*	15	70	54	-0.018	0.054	0.02	3.027	0.00252	0.095
ViscFat	*UBL7*	15	74	14	0.096	0.084	0.03	2.687	0.00730	0.094
TibNerve	*UBL7*	15	75	47	-0.112	0.028	0.01	2.653	0.00807	0.093
Ileum	*KIAA1199*	15	81	25	-0.010	-0.034	0.01	-2.847	0.00449	0.094
thyroid	*EFTUD1*	15	83	5	0.016	-0.128	0.05	-2.617	0.00898	0.093
TibNerve	*ADAMTSL3*	15	85	16	0.007	-0.080	0.03	-3.108	0.00192	0.095
GTExBlood	*NMB*	15	86	48	-0.008	0.043	0.01	2.906	0.00372	0.094
Ileum	*FANCI*	15	90	113	0.078	0.016	0.01	3.204	0.00139	0.095
thyroid	*ANPEP*	15	91	111	0.100	-0.036	0.01	-3.301	0.00099	0.096
DGNblood	*SYNM*	15	100	41	-0.012	0.013	0.00	2.658	0.00797	0.093
thyroid	*CCDC78*	16	1	10	0.027	0.096	0.03	2.928	0.00347	0.094
liver	*RNPS1*	16	1	11	0.019	-0.052	0.02	-2.621	0.00886	0.094
pancreas	*FAM195A*	16	2	44	-0.077	-0.043	0.02	-2.660	0.00790	0.093
skmuscle	*RNPS1*	16	2	53	0.033	-0.036	0.01	-2.741	0.00620	0.093
DGNblood	*RNPS1*	16	2	13	-0.036	-0.038	0.01	-2.949	0.00324	0.094
thyroid	*RNPS1*	16	2	15	0.049	-0.046	0.02	-2.747	0.00610	0.093
GTExBlood	*PAQR4*	16	3	10	0.132	0.066	0.03	2.599	0.00945	0.093
pancreas	*RNPS1*	16	3	12	-0.106	-0.073	0.03	-2.702	0.00699	0.094
TibNerve	*RNPS1*	16	3	17	-0.013	-0.033	0.01	-2.815	0.00495	0.094
TibNerve	*EMP2*	16	10	54	0.008	-0.021	0.01	-2.769	0.00571	0.094
thyroid	*EMP2*	16	11	27	-0.104	-0.022	0.01	-2.766	0.00575	0.094
GTExBlood	*BFAR*	16	15	5	-0.003	-0.419	0.12	-3.628	0.00030	0.097
DGNblood	*ABCC6*	16	16	20	-0.018	0.049	0.02	3.247	0.00120	0.096
thyroid	*ABCC6*	16	16	20	-0.007	0.031	0.01	2.651	0.00813	0.093
thyroid	*NOMO3*	16	16	28	-0.011	0.049	0.02	2.769	0.00570	0.094
TibNerve	*ABCC6*	16	17	48	-0.025	0.021	0.01	2.851	0.00442	0.094
GTExBlood	*NOMO3*	16	17	45	0.011	0.028	0.01	2.631	0.00862	0.093
spleen	*TNRC6A*	16	24	3	0.002	1.123	0.41	2.767	0.00574	0.094
skmuscle	*AQP8*	16	25	110	0.039	0.041	0.01	3.520	0.00045	0.097
liver	*APOBR*	16	29	2	0.027	-0.411	0.15	-2.743	0.00617	0.094
pancreas	*ALDOA*	16	31	1	0.030	-2.788	1.02	-2.725	0.00651	0.093
ViscFat	*ITGAD*	16	32	7	0.039	0.088	0.03	2.890	0.00392	0.094
pancreas	*CMTM3*	16	66	41	-0.037	0.046	0.02	2.757	0.00592	0.094
DGNblood	*PLA2G15*	16	68	16	-0.046	0.080	0.03	2.744	0.00614	0.093

GTExBlood	*HSBP1*	16	83	49	-0.130	0.056	0.02	2.837	0.00462	0.094
liver	*SPATA33*	16	89	18	-0.367	0.042	0.02	2.602	0.00939	0.093
thyroid	*DEF8*	16	90	47	0.056	-0.033	0.01	-3.036	0.00244	0.095
spleen	*TRPV3*	17	3	18	0.001	0.026	0.01	2.591	0.00966	0.093
skmuscle	*ALOX15*	17	4	12	0.013	0.145	0.05	3.142	0.00172	0.095
TibNerve	*ACAP1*	17	6	20	0.025	0.040	0.01	2.768	0.00573	0.094
skmuscle	*YBX2*	17	6	94	0.084	0.039	0.01	3.295	0.00101	0.096
DGNblood	*CLEC10A*	17	7	35	0.051	-0.027	0.01	-2.701	0.00699	0.093
TibNerve	*CLEC10A*	17	7	34	0.015	-0.045	0.01	-3.022	0.00256	0.095
GTExBlood	*ACADVL*	17	8	12	-0.018	-0.056	0.02	-2.646	0.00824	0.093
ViscFat	*CNTROB*	17	8	1	0.072	1.276	0.40	3.227	0.00128	0.096
skmuscle	*DLG4*	17	8	21	0.099	0.043	0.02	2.765	0.00578	0.094
TibNerve	*HS3ST3B1*	17	14	11	0.043	0.073	0.03	2.713	0.00675	0.093
skmuscle	*PLD6*	17	16	3	0.007	0.188	0.07	2.869	0.00419	0.094
TibNerve	*KSR1*	17	27	22	-0.023	0.071	0.02	3.170	0.00156	0.095
skmuscle	*POLDIP2*	17	27	22	0.012	0.074	0.03	2.725	0.00652	0.093
skmuscle	*TMEM97*	17	27	6	0.013	0.117	0.04	2.749	0.00605	0.093
liver	*ABHD15*	17	28	50	-0.052	0.032	0.01	3.057	0.00228	0.096
TibNerve	*RHBDL3*	17	30	60	0.033	-0.054	0.02	-3.234	0.00125	0.095
thyroid	*RHOT1*	17	31	15	-0.042	-0.106	0.03	-3.149	0.00168	0.095
pancreas	*TMEM98*	17	32	24	0.089	-0.053	0.02	-2.852	0.00442	0.094
DGNblood	*CCT6B*	17	33	49	0.003	-0.032	0.01	-2.801	0.00516	0.094
thyroid	*CCT6B*	17	33	24	-0.011	-0.068	0.02	-3.084	0.00208	0.095
TibNerve	*CCT6B*	17	33	19	0.027	-0.023	0.01	-2.633	0.00857	0.093
pancreas	*CCT6B*	17	34	31	-0.037	-0.040	0.02	-2.587	0.00979	0.096
GTExBlood	*CCL3L3*	17	35	23	0.002	-0.031	0.01	-2.751	0.00603	0.093
DGNblood	*PSMB3*	17	37	22	-0.332	-0.026	0.01	-2.644	0.00828	0.093
GTExBlood	*PSMB3*	17	37	13	-0.216	-0.044	0.01	-3.137	0.00175	0.095
thyroid	*TMEM101*	17	42	30	-0.013	-0.069	0.02	-2.914	0.00363	0.094
DGNblood	*CCDC103*	17	43	36	0.000	-0.020	0.01	-2.599	0.00946	0.093
TibNerve	*BZRAP1*	17	56	17	0.030	-0.033	0.01	-3.007	0.00269	0.094
DGNblood	*TEX14*	17	56	11	0.000	-0.052	0.02	-2.956	0.00317	0.094
liver	*TEX14*	17	56	25	-0.011	-0.047	0.02	-2.718	0.00665	0.094
spleen	*TEX14*	17	56	39	-0.078	-0.025	0.01	-2.685	0.00734	0.093
thyroid	*TEX14*	17	56	42	0.014	-0.025	0.01	-2.595	0.00957	0.093
skmuscle	*SKA2*	17	57	10	0.007	0.100	0.03	3.419	0.00065	0.096
DGNblood	*SKA2*	17	57	11	-0.005	0.129	0.03	3.813	0.00014	0.098
thyroid	*SKA2*	17	57	37	0.269	0.046	0.01	3.267	0.00112	0.096
TibNerve	*SKA2*	17	57	17	-0.011	0.048	0.01	3.736	0.00020	0.098
GTExBlood	*SMG8*	*17*	57	*5*	*-0.129*	*#####*	*0.04*	*-4.337*	*0.00002*	*0.101*

skmuscle	*TEX14*	17	57	11	0.078	-0.061	0.02	-3.513	0.00046	0.097
TibNerve	*TEX14*	17	57	62	0.006	-0.046	0.01	-3.683	0.00024	0.098
skmuscle	*TRIM37*	17	57	51	0.172	-0.029	0.01	-3.284	0.00105	0.096
DGNblood	*TRIM37*	17	57	35	0.004	-0.014	0.00	-2.783	0.00547	0.094
thyroid	*TRIM37*	17	57	11	-0.046	-0.072	0.02	-3.496	0.00049	0.097
TibNerve	*TRIM37*	17	57	56	0.000	-0.028	0.01	-3.219	0.00132	0.095
Ileum	*HEATR6*	17	58	44	-0.021	0.021	0.01	2.698	0.00706	0.093
Ileum	*C17orf82*	17	60	10	-0.036	0.050	0.02	2.711	0.00679	0.093
Ileum	*CEP112*	17	65	77	0.008	0.019	0.01	3.143	0.00171	0.095
GTExBlood	*PRKCA*	17	65	41	-0.029	-0.057	0.02	-2.886	0.00397	0.094
ViscFat	*FAM104A*	17	71	41	0.024	0.076	0.02	3.153	0.00165	0.095
ViscFat	*MYO15B*	17	74	13	0.058	0.081	0.03	3.021	0.00257	0.095
skmuscle	*SPHK1*	17	75	22	0.008	0.109	0.04	2.669	0.00770	0.093
DGNblood	*AFMID*	17	76	3	0.077	0.081	0.03	3.057	0.00228	0.095
liver	*BIRC5*	17	76	26	-0.011	0.104	0.03	3.389	0.00072	0.098
skmuscle	*CYTH1*	17	77	58	0.173	0.024	0.01	2.613	0.00908	0.093
TibNerve	*SYNGR2*	17	77	76	0.004	0.023	0.01	2.707	0.00687	0.093
thyroid	*AC132872*	17	80	20	-0.053	0.069	0.03	2.766	0.00576	0.094
Ileum	*C18orf56*	18	1	30	-0.172	-0.031	0.01	-2.751	0.00602	0.093
TibNerve	*PPP4R1*	18	9	8	0.003	-0.039	0.01	-2.645	0.00827	0.093
GTExBlood	*PPP4R1*	18	10	19	0.001	0.070	0.02	2.957	0.00316	0.094
thyroid	*ABHD3*	18	19	23	-0.005	-0.060	0.02	-2.982	0.00291	0.094
thyroid	*C18orf8*	18	22	24	0.016	-0.121	0.04	-3.364	0.00079	0.096
DGNblood	*GNG7*	19	2	34	-0.075	0.025	0.01	2.683	0.00738	0.093
GTExBlood	*GNG7*	19	3	52	-0.054	0.043	0.01	2.905	0.00374	0.094
DGNblood	*EEF2*	19	4	25	-0.058	-0.085	0.03	-2.650	0.00814	0.093
DGNblood	*QTRT1*	19	10	17	-0.034	0.121	0.04	2.870	0.00418	0.094
DGNblood	*ICAM5*	19	11	58	0.016	-0.023	0.01	-3.087	0.00206	0.095
thyroid	*ICAM5*	19	11	8	0.002	-0.060	0.02	-2.646	0.00823	0.093
liver	*S1PR2*	19	11	21	0.014	-0.044	0.02	-2.860	0.00431	0.094
DGNblood	*DNAJB1*	19	15	17	-0.033	0.044	0.02	2.711	0.00681	0.093
skmuscle	*EMR2*	19	15	49	0.035	-0.058	0.02	-2.842	0.00455	0.094
DGNblood	*IL27RA*	19	15	29	0.005	-0.044	0.02	-2.805	0.00511	0.094
TibNerve	*OR7C1*	19	15	47	0.061	0.019	0.01	2.718	0.00666	0.093
ViscFat	*SYDE1*	19	15	23	0.019	0.096	0.03	2.990	0.00284	0.095
ViscFat	*ABHD8*	19	17	34	0.007	-0.074	0.02	-2.964	0.00309	0.096
thyroid	*MED26*	19	17	31	-0.003	-0.045	0.02	-2.645	0.00828	0.093
DGNblood	*ARRDC2*	19	18	20	-0.012	0.072	0.03	2.725	0.00652	0.093
liver	*IL12RB1*	19	18	34	-0.025	0.043	0.01	2.859	0.00431	0.094
skmuscle	*JUND*	19	19	23	0.036	-0.065	0.02	-3.161	0.00161	0.095

skmuscle	*ZNF101*	19	19	56	0.009	-0.066	0.02	-3.343	0.00085	0.096
TibNerve	*C19orf12*	19	31	1	0.012	-1.143	0.44	-2.588	0.00976	0.093
DGNblood	*PDCD2L*	19	35	30	-0.003	-0.091	0.04	-2.600	0.00941	0.093
DGNblood	*NFKBID*	19	36	10	-0.004	0.132	0.05	2.815	0.00495	0.094
spleen	*THAP8*	19	36	45	0.007	-0.027	0.01	-3.021	0.00257	0.095
ViscFat	*THAP8*	19	36	35	0.023	-0.058	0.02	-2.802	0.00516	0.094
skmuscle	*ZFP82*	19	36	18	0.021	-0.063	0.02	-2.686	0.00732	0.093
ViscFat	*CATSPERG*	19	38	38	0.074	-0.062	0.02	-2.912	0.00365	0.095
GTExBlood	*CATSPERG*	19	39	94	-0.082	-0.018	0.01	-2.706	0.00689	0.093
TibNerve	*FBXO17*	19	39	92	0.229	0.021	0.01	2.619	0.00893	0.093
pancreas	*RPS16*	19	40	75	0.005	-0.026	0.01	-2.740	0.00623	0.095
pancreas	*APOC2*	19	44	43	0.045	0.053	0.02	2.605	0.00930	0.094
thyroid	*ZNF283*	19	44	33	-0.008	0.031	0.01	2.621	0.00887	0.093
spleen	*ZNF225*	19	45	62	0.140	-0.020	0.01	-2.747	0.00609	0.093
pancreas	*ERCC1*	19	47	3	0.002	0.064	0.02	2.805	0.00510	0.094
thyroid	*ZNF114*	19	49	12	-0.014	-0.075	0.03	-2.699	0.00705	0.093
liver	*ALDH16A1*	19	50	1	-0.042	0.505	0.19	2.640	0.00838	0.093
ViscFat	*ZNF701*	19	53	29	0.023	0.083	0.03	3.045	0.00238	0.095
DGNblood	*CACNG8*	19	54	33	0.038	0.033	0.01	2.631	0.00860	0.093
ViscFat	*ZNF665*	19	54	18	-0.048	0.099	0.04	2.651	0.00813	0.094
Ileum	*TMEM150B*	19	55	59	0.047	0.027	0.01	3.155	0.00164	0.095
GTExBlood	*ZNF470*	19	56	7	-0.045	0.154	0.05	2.877	0.00408	0.094
thyroid	*ZNF784*	19	57	4	0.215	-0.121	0.04	-3.357	0.00081	0.096
Ileum	*RBCK1*	20	1	62	-0.090	-0.021	0.01	-2.741	0.00621	0.093
ViscFat	*SIRPB1*	20	3	40	0.006	0.033	0.01	2.909	0.00368	0.095
DGNblood	*PRNP*	20	5	4	0.041	-0.129	0.04	-2.886	0.00397	0.094
TibNerve	*MKKS*	20	10	23	0.155	0.075	0.03	2.744	0.00615	0.093
thyroid	*CST4*	20	23	23	-0.030	0.059	0.02	3.137	0.00175	0.095
pancreas	*PYGB*	20	25	55	-0.077	0.041	0.01	3.198	0.00142	0.095
DGNblood	*CDK5RAP1*	20	31	7	0.021	0.164	0.06	2.708	0.00686	0.093
skmuscle	*DNMT3B*	20	31	1	0.009	-0.526	0.17	-3.166	0.00158	0.095
TibNerve	*BPIFB4*	20	32	54	-0.143	-0.028	0.01	-2.830	0.00473	0.094
thyroid	*SULF2*	20	46	40	-0.038	0.059	0.02	2.851	0.00443	0.094
GTExBlood	*MTG2*	20	61	14	-0.093	-0.062	0.02	-2.813	0.00498	0.094
TibNerve	*ARFGAP1*	20	62	2	0.428	-0.434	0.13	-3.396	0.00071	0.096
DGNblood	*ADAMTS1*	21	28	30	-0.039	-0.047	0.02	-2.735	0.00632	0.093
skmuscle	*N6AMT1*	21	30	38	0.098	0.043	0.01	3.070	0.00218	0.095
DGNblood	*N6AMT1*	21	30	20	-0.037	0.053	0.02	2.726	0.00649	0.093
TibNerve	*N6AMT1*	21	30	9	-0.042	0.103	0.03	3.130	0.00179	0.095
DGNblood	*RWDD2B*	21	30	20	0.008	0.017	0.01	2.585	0.00985	0.093

spleen	*RWDD2B*	21	30	26	-0.006	0.038	0.01	3.575	0.00036	0.097
thyroid	*RWDD2B*	21	30	23	-0.027	0.024	0.01	2.743	0.00616	0.093
TibNerve	*RWDD2B*	21	30	24	-0.042	0.023	0.01	2.778	0.00556	0.094
GTExBlood	*RWDD2B*	21	30	14	0.019	0.047	0.02	2.801	0.00517	0.094
thyroid	*BACE2*	21	42	88	-0.008	0.021	0.01	2.610	0.00916	0.093
thyroid	*PDE9A*	21	44	13	0.060	-0.043	0.02	-2.849	0.00446	0.094
DGNblood	*C21orf67*	21	46	38	-0.001	-0.041	0.01	-2.846	0.00450	0.094
DGNblood	*SUMO3*	21	46	18	0.001	-0.083	0.03	-3.255	0.00116	0.096
ViscFat	*TRPM2*	21	46	7	0.030	-0.096	0.03	-3.307	0.00097	0.097
GTExBlood	*COL6A2*	21	47	77	0.036	0.029	0.01	2.814	0.00497	0.094
TibNerve	*LSS*	21	47	60	0.036	-0.028	0.01	-2.745	0.00614	0.093
thyroid	*DIP2A*	21	48	78	0.150	0.034	0.01	3.024	0.00254	0.095
ViscFat	*PCNT*	21	48	40	-0.002	-0.043	0.02	-2.680	0.00746	0.094
pancreas	*PRMT2*	21	48	6	-0.085	0.075	0.03	2.829	0.00474	0.094
ViscFat	*PRMT2*	21	48	30	0.120	0.053	0.02	2.641	0.00836	0.093
TibNerve	*IL17RA*	22	18	19	0.029	-0.030	0.01	-2.761	0.00584	0.094
thyroid	*TUBA8*	22	19	4	-0.101	0.101	0.04	2.674	0.00758	0.093
DGNblood	*COMT*	22	21	25	-0.012	0.039	0.02	2.612	0.00910	0.093
pancreas	*COMT*	22	21	51	-0.034	0.054	0.01	3.680	0.00024	0.098
skmuscle	*PI4KA*	22	21	22	0.053	-0.056	0.02	-3.186	0.00147	0.095
liver	*GNAZ*	22	23	90	0.179	-0.028	0.01	-3.599	0.00033	0.098
DGNblood	*CHEK2*	22	28	23	-0.030	0.050	0.02	2.948	0.00326	0.094
DGNblood	*HSCB*	22	28	14	-0.010	0.180	0.05	3.274	0.00109	0.096
DGNblood	*TTC28*	22	29	24	0.033	0.037	0.01	2.728	0.00645	0.093
GTExBlood	*TTC28*	22	29	34	0.003	0.076	0.03	2.753	0.00598	0.093
DGNblood	*CCDC117*	22	30	24	0.047	0.070	0.03	2.649	0.00816	0.093
DGNblood	*ZNRF3*	22	30	21	-0.053	0.070	0.02	3.317	0.00093	0.096
DGNblood	*SOX10*	22	38	7	0.003	-0.100	0.04	-2.668	0.00772	0.093
skmuscle	*TTLL12*	22	44	12	0.209	0.028	0.01	2.998	0.00277	0.094
DGNblood	*TTLL12*	22	44	26	-0.034	0.022	0.01	3.291	0.00102	0.096
thyroid	*TTLL12*	22	44	11	-0.149	0.032	0.01	3.391	0.00072	0.096
ViscFat	*TTLL12*	22	45	30	-0.029	0.047	0.01	3.333	0.00088	0.097
GTExBlood	*TTLL12*	22	45	15	-0.002	0.055	0.02	3.071	0.00218	0.095
liver	*SYCE3*	22	51	7	0.128	-0.061	0.02	-3.015	0.00262	0.095

Table A.2 SNPs in the predictor for SPPL2A in the DGN blood training set.

DGN	Weight	Chr	Position	Mb
rs11070715	0.00396265	15	50010853	50.01
rs1484555	0.02200433	15	50014077	50.01
rs559561	-0.0012285	15	50016296	50.02
rs642981	-0.016351	15	50029628	50.03
rs8034382	0.01702327	15	50272085	50.27
rs16953000	0.00182882	15	50275056	50.28
rs6493393	0.02587444	15	50276945	50.28
rs12101586	0.03989513	15	50278691	50.28
rs3105591	0.00062775	15	50869042	50.86
rs2414060	0.00566576	15	50873262	50.87
rs3109891	0.00160804	15	50874964	50.87
rs1395297	0.07401904	15	50878509	50.87
rs3109894	0.00010798	15	50878574	50.87
rs12592778	0.01230376	15	50992311	50.99
rs12442197	0.04868085	15	50995362	50.99
rs12437803	0.04800037	15	50997266	51
rs12437770	0.04771838	15	50997304	51
rs12440864	0.00259943	15	51005957	51
rs12595082	0.04224487	15	51007729	51
rs17646025	0.00494609	15	51019812	51.02
rs12912192	-0.0901485	15	51057181	51.06
rs17521308	-0.2509104	15	51161025	51.16
rs17599691	-0.0992145	15	51169642	51.17
rs17521440	-0.023195	15	51210563	51.21
rs3784301	0.0111016	15	51225825	51.22
rs10851494	0.02365192	15	51393400	51.39
rs4774582	0.00247494	15	51454065	51.45
rs12594156	0.00354991	15	51456518	51.45
rs1438920	0.0023514	15	51458054	51.46
rs4775923	0.00132885	15	51459264	51.46
rs1438922	0.00158764	15	51460148	51.46
rs4775926	3.98E-05	15	51460660	51.46
rs1438924	0.00021399	15	51463624	51.46
Near Genome Wide Significant SNps from Paterson et al., (2010)				
rs493218		15	51277554	51.27
rs572221		15	51291924	51.29
rs690271		15	51291964	51.29
rs566369		15	51295884	51.3
rs482541		15	51296486	51.3

Supplementary Data B: Scripts

B.1 SNP Extraction

```
#$ -S /bin/bash
#! /bin/bash

cd /hpf/largeprojects/andrew/adp/sgood/HbA1C/DCCT/new/WholeBlood

###loop through chromosomes and submit batches

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

qsub  -o ~/queue -e ~/queue -v chr=$a analysis2.sh
let a=$a+1
done

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

cd /hpf/largeprojects/andrew/adp/sgood/HbA1C/DCCT/new/WholeBlood

#### changing to current working directory

chr=$chr

#For all SNPS across genome in DCCT genetic data
#if the file is in gz format first unzip and then extract
#gunzip -c

fgrep -w -f SNPs_GTEX_WB.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 by Joanne Gittens 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 Estimating GReX

GTEx WholeBlood 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/sgood/HbA1C/DCCT/WholeBlood
--dosages_prefix chr --samples samples_1304.txt
--weights TW_Whole_Blood_0.5_1KG --output_dir out
```

B.4 Association Analyses

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

cd /hpf/largeprojects/andrew/adp/sgood/HbA1C/DCCT/new/DGN/output

#### changing to current working directory
#cd $PBS_O_WORKDIR

module load R/3.3.0

/hpf/tools/centos6/R/3.3.0/bin/Rscript Assoc_DCCT_comb.R
```

B.5. Extracting gene expression data for a given gene (using relevant ensemble ID) and training set using shell commands

```
awk '
NR==1 {
  for (i=1; i<=NF; i++) {
    f[$i] = i
  }
}
{ print $(f["FID"]), $(f["ENSG00000138600.5"]), $(f["ENSG00000121653.7"]),
$(f["ENSG00000156983.11"]) }
' predicted_expression.txt > DGNSig.txt
```

B.5 Example R-code

```
#Phenotype Graphs
names(A1C_conv)
as.factor(A1C$treatment)
library(ggplot2)
library(gridExtra)
require('plyr')
library(dplyr)

plot1 <- ggplot(A1C, aes(x=meanhba1c), fill=treatment) +
 geom_histogram(binwidth=0.25, alpha=.5, position="identity", colour = "black", fill="blue") +
geom_vline(aes(xintercept=mean(meanhba1c, na.rm=T)),
      color="red", linetype="dashed", size=1) +
  facet_grid(~treatment)
plot1
ggplot(A1C, aes(x=meanloghba1c), fill=treatment) +
  geom_histogram(binwidth=0.25, alpha=.5, position="identity", colour = "black", fill="white") +
  geom_vline(aes(xintercept=mean(meanloghba1c, na.rm=T)),
             color="red", linetype="dashed", size=1) +
  facet_grid(~treatment)

A1C_conv <- read.table("dcct_conv2.txt", header=TRUE)
names(A1C_conv)
A1C_Int <- read.table("dcct_int2.txt", header=TRUE)
plot1 <- ggplot(A1C_conv, aes(x=meanhba1c)) +
  geom_histogram(aes(y=..density..),      # Histogram with density instead of count on y-axis
                 binwidth=.5,
                 colour="black", fill="white") +
  geom_density(alpha=.2, fill="#99FF66") + xlim(5,13)+
  ggtitle("Mean HbA1c - Conventional")

plot2 <- ggplot(A1C_Int, aes(x=meanhba1c)) +
  geom_histogram(aes(y=..density..),      # Histogram with density instead of count on y-axis
                 binwidth=.5,
                 colour="black", fill="white") +
  geom_density(alpha=.2, fill="#99FF66") + xlim(5,13)+
  ggtitle("Mean HbA1c - Intensive")

plot3 <- ggplot(A1C_conv, aes(x=logmeanhba1c)) +
  geom_histogram(aes(y=..density..),      # Histogram with density instead of count on y-axis
                 binwidth=.1,
                 colour="black", fill="white") +
  geom_density(alpha=.2, fill="#99FF66") + xlim(1.2,3) +
  ggtitle("meanlogHbA1c - Conventional")

plot4 <- ggplot(A1C_Int, aes(x=logmeanhba1c)) +
  geom_histogram(aes(y=..density..),      # Histogram with density instead of count on y-axis
                 binwidth=.1,
                 colour="black", fill="white") +
  geom_density(alpha=.2, fill="#99FF66") + xlim(1.2,3) +
  ggtitle("Meanlog HbA1c - Intensive")

grid.arrange(plot1, plot3, plot2, plot4, nrow =2, ncol=2)


#TWAS results
source("https://bioconductor.org/biocLite.R")
biocLite("GWASTools")
setwd("C:/Users/User/Dropbox/DIABETES - Paterson/MSc thesis/Msc Thesis
Analyses/PredictedExpression&GWAS")
library(GWASTools)
gtexwb <- read.table("WholeBloodSorted.txt", header=TRUE)attach(pancreas)
jpeg("C:/Users/User/Dropbox/DIABETES - Paterson/MSc thesis/Msc Thesis
Analyses/PredictedExpression&GWAS/manhat_HbA1C_gtexwb_comb2.jpeg")
par(mfrow=c(1,1))
library(reshape)
library(gridExtra)
```

```
par(mfrow=c(2,2))
man1 <- manhattanPlot(gtexwb$pval, gtexwb$chr, signif = (0.05/length(gtexwb$Gene)), main="GTEX-
Whole Blood")
pval1 <- hist(gtexwb$pval, prob=TRUE, breaks=(0:20)/20, main="P-value Density", xlab="p values")
abline(h=1, col="red")

#second approach to plot TWAS
library(qqman)
manhattan(pancreas, chr = "chr", bp = "position", p = "pval", snp = "Gene",
        col = c("gray10", "gray60"), chrlabs = NULL,
        suggestiveline = -log10(1e-05), genomewideline = -log10(1e-06))
manhattan(gtexwb, chr = "chr", bp = "position", p = "pval", snp = "Gene",
        col = c("gray10", "gray60"), chrlabs = NULL,
        suggestiveline = -log10(1e-05), genomewideline = -log10(1e-06))

#comparing models for 2df test for a specific gene
PredExp <- read.table("PredExpAssociations.txt", header=TRUE)
names(PredExp)
model1 <- lm(PredExp$meanloghba1c ~ Age + Gender + Duration + retbase + treatment +
retbase*treatment, data = PredExp)
model2 <- lm(PredExp$meanloghba1c ~ Age + Gender + Duration + retbase + treatment +
retbase*treatment + SMG8_WB + SMG8_WB*treatment, data = PredExp)
model3 <- lm(PredExp$meanloghba1c ~ Age + Gender + Duration + retbase + treatment +
retbase*treatment + SMG8_WB, data = PredExp)

lrtest(model1, model2)

#Bland-altman plots
install.packages("BlandAltmanLeh")
library(BlandAltmanLeh)
library(ggplot2)
#par(mfrow=c(2,3))
MAPK <- bland.altman.plot(PredExp$MAPK8IP1_WB, PredExp$MAPK8IP1_DGN, graph.sys = "ggplot2",
conf.int=.95, pch=19) + ggtitle("MapK8IP1") + xlab("Mean MAPK8 expression DGN-GTEX") + ylab("Diff in
MAPK8 expression") + geom_point(size=0.2)

#training set graphs
require(gridExtra)
library(ggplot2)
DGN_weights <- read.csv("weights_DGN_WB_0.5.csv", header=TRUE)
GTEX_weights <- read.csv("weights-GTEX-WB.csv", header=TRUE)
names(DGN_weights)

p7<- ggplot(DGN_weights, aes(weight)) +
  geom_density(alpha=0.1, fill="blue") + xlim(-0.5, 0.5) + ggtitle("DGN Blood SNP weights") +
theme(plot.title = element_text(hjust = 0.5))
p8 <- ggplot(GTEX_weights, aes(weight)) +
  geom_density(alpha=0.1, fill="yellow") + xlim(-0.5, 0.5) + ggtitle("GTEx Blood SNP weights") +
theme(plot.title = element_text(hjust = 0.5))
grid.arrange(p7, p8, nrow=1, ncol=2)

DGN_maf <- read.table("chr_rsid_pos-dgn.txt", header=TRUE)
GTEX_maf <- read.table("chr_rsid_pos-gtex.txt", header=TRUE)
names(DGN_maf)
```

```
source("multiplot.R")
require(gridExtra)
p <- ggplot(data=DGN, aes(x=n.snps.in.model))
plot1 <- p + geom_density(fill="light blue") + xlab("Density SNPS DGN-blood")
q <- ggplot(data=GTEX, aes(x=n.snps.in.model))
plot2 <- q + geom_density(fill="pink")+ xlab("Density SNPs GTEX-blood")
grid.arrange(plot1, plot2, nrow=1, ncol=2)

GTEX_R2 <- read.csv("extra-GTEX-WB-R2.csv", header=TRUE)
DGN_R2 <- read.csv("extra-DGN-WB-R2.csv", header=TRUE)
GTEX_DGN_R2 <- merge(DGN_R2, GTEX_R2, by=c("genename"), all.x=TRUE)
write.table(GTEX_DGN_R2, "Gtex_DGN_R2.txt", quote=F,row.names=F, sep="\t")

R2_complete <- GTEX_DGN_R2[complete.cases(GTEX_DGN_R2), ]
par(mfrow=c(1,1))
names(R2_complete)
p21 <- ggplot(R2_complete, aes(pred.perf.R2.x, y=pred.perf.R2.y)) +
  geom_point(shape=1) +geom_smooth(method=lm)

print(p21 + labs(x="R2 (DGN-blood)", y="R2 (GTEx-blood)" ))
cor(R2_complete$pred.perf.R2.x, R2_complete$pred.perf.R2.y)
cor.test(R2_complete$pred.perf.R2.x, R2_complete$pred.perf.R2.y, method=c("pearson"))
plot(R2_complete$pred.perf.R2.x, R2_complete$pred.perf.R2.y, xlab="R2_DGN ", ylab="R2_GTEX", main = "
Correlation between CV R2 in whole blood DGN vs GTEX")

#LM of covariates only,
HbA1C_Comb <- read.table("dcct_comb.txt", header = TRUE)
HbA1C <- names(HbA1C_Comb[c(5,6)])
output <- data.frame(Estimate=NA, SE=NA, T=NA, pval=NA, R2 = NA )
model <- lm(HbA1C_Comb$logmeanhba1c ~ Age + Gender + Duration + retbase + treatment +
retbase*treatment, data = HbA1C_Comb)
logLik(model)
summary.lm(model)
layout(matrix(c(1,2,3,4),2,2))
plot(model)
anova(model)

#LM with predicted expression of Mean HbA1c and meanlog10HbA1c
genes <- names(df[,-c(1:9)])

for (A1C in HbA1C) {
  output <- data.frame(Gene=genes, Estimate=NA, SE=NA, T=NA, pval=NA, R2 = NA )
  for (i in 1:length(genes)) {
    pred_gene_exp <- df[,i+9]
    model <- lm(df[,A1C] ~ pred_gene_exp + Age + Gender + Duration + retbase + treatment +
retbase*treatment, data = df)
    output[i,2:5] <- coef(summary(model))[2,]
    output[i,6] <- summary(model)$r.squared
    assign(paste("as", "pancreas", A1C, "Comb", sep="_"), output)
    write.table(output, paste("as", "pancreas", A1C, "Comb.txt", sep="_"), quote=F, sep="\t", col.names=T,
row.names=F)
```

References

Adzhubei, I. A., S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova *et al.*, 2010 A method and server for predicting damaging missense mutations. Nat Methods 7**:** 248-249.

Ahlqvist, E., N. R. van Zuydam, L. C. Groop and M. I. McCarthy, 2015 The genetics of diabetic complications. Nat Rev Nephrol 11**:** 277-287.

Albert, F. W., and L. Kruglyak, 2015 The role of regulatory variation in complex traits and disease. Nat Rev Genet 16**:** 197-212.

Anderson, C. A., F. H. Pettersson, G. M. Clarke, L. R. Cardon, A. P. Morris *et al.*, 2010 Data quality control in genetic case-control association studies. Nat Protoc 5**:** 1564-1573.

Arif, S., P. Leete, V. Nguyen, K. Marks, N. M. Nor *et al.*, 2014 Blood and islet phenotypes indicate immunological heterogeneity in type 1 diabetes. Diabetes 63**:** 3835-3845.

Auer, P. L., and G. Lettre, 2015 Rare variant association studies: considerations, challenges and opportunities. Genome Med 7**:** 16.

Bansal, P., S. Wang, S. Liu, Y. Y. Xiang, W. Y. Lu *et al.*, 2011 GABA coordinates with insulin in regulating secretory function in pancreatic INS-1 beta-cells. PLoS One 6**:** e26225.

Battle, A., S. Mostafavi, X. Zhu, J. B. Potash, M. M. Weissman *et al.*, 2014 Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res 24**:** 14-24.

Bernstein, B. E., J. A. Stamatoyannopoulos, J. F. Costello, B. Ren, A. Milosavljevic *et al.*, 2010 The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol 28**:** 1045-1048.

Bizzarri, C., D. Pitocco, N. Napoli, E. Di Stasio, D. Maggi *et al.*, 2010 No protective effect of calcitriol on beta-cell function in recent-onset type 1 diabetes: the IMDIAB XIII trial. Diabetes Care 33**:** 1962-1963.

Blanchette, M., and M. Tompa, 2002 Discovery of regulatory elements by a computational method for phylogenetic footprinting. Genome Res 12**:** 739-748.

Bonnefond, A., M. Vaxillaire, Y. Labrune, C. Lecoeur, J. C. Chevre *et al.*, 2009 Genetic variant in HK1 is associated with a proanemic state and A1C but not other glycemic control-related traits. Diabetes 58**:** 2687-2697.

Brodie, A., J. R. Azaria and Y. Ofran, 2016 How far from the SNP may the causative genes be? Nucleic Acids Res 44**:** 6046-6054.

Bush, W. S., and J. H. Moore, 2012 Chapter 11: Genome-wide association studies. PLoS Comput Biol 8**:** e1002822.

Cain, C. E., R. Blekhman, J. C. Marioni and Y. Gilad, 2011 Gene expression differences among primates are associated with changes in a histone epigenetic modification. Genetics 187**:** 1225-1234.

Chen, Z., S. Li, S. Subramaniam, J. Y. Shyy and S. Chien, 2017 Epigenetic Regulation: A New Frontier for Biomedical Engineers. Annu Rev Biomed Eng 19**:** 195-219.

Cho, J. H., and M. Feldman, 2015 Heterogeneity of autoimmune diseases: pathophysiologic insights from genetics and implications for new therapies. Nat Med 21**:** 730-738.

Cohen, R. M., R. S. Franco, P. K. Khera, E. P. Smith, C. J. Lindsell *et al.*, 2008 Red cell life span heterogeneity in hematologically normal people is sufficient to alter HbA1c. Blood 112**:** 4284-4291.

Consortium, E. P., 2012 An integrated encyclopedia of DNA elements in the human genome. Nature 489**:** 57-74.

Consortium, G. T., 2015 Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 348**:** 648-660.

Consortium, G. T., D. A. Laboratory, G. Coordinating Center -Analysis Working, G. Statistical Methods groups-Analysis Working, G. g. Enhancing *et al.*, 2017 Genetic effects on gene expression across human tissues. Nature 550**:** 204-213.

Dabelea, D., 2009 The accelerating epidemic of childhood diabetes. Lancet 373**:** 1999-2000.

Davey Smith, G., S. Ebrahim, S. Lewis, A. L. Hansell, L. J. Palmer *et al.*, 2005 Genetic epidemiology and public health: hope, hype, and future prospects. Lancet 366**:** 1484-1498.

de Boer, I. H., T. C. Rue, P. A. Cleary, J. M. Lachin, M. E. Molitch *et al.*, 2011 Long-term renal outcomes of patients with type 1 diabetes mellitus and microalbuminuria: an analysis of the Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications cohort. Arch Intern Med 171**:** 412-420.

Devlin, B., and K. Roeder, 1999 Genomic control for association studies. Biometrics 55**:** 997-1004.

DeWan, A. T., 2010 Five classic articles in genetic epidemiology. Yale J Biol Med 83**:** 87-90.

Diabetes, C., G. Complications Trial Research, D. M. Nathan, S. Genuth, J. Lachin *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**:** 977-986.

Diabetes, C., I. Complications Trial/Epidemiology of Diabetes and G. Complications Research, 2016 Risk Factors for Cardiovascular Disease in Type 1 Diabetes. Diabetes 65**:** 1370-1379.

Fallin, M. D., P. Duggal and T. H. Beaty, 2016 Genetic Epidemiology and Public Health: The Evolution From Theory to Technology. Am J Epidemiol 183**:** 387-393.

Fizelova, M., A. Stancakova, C. Lorenzo, S. M. Haffner, H. Cederberg *et al.*, 2015 Glycated hemoglobin levels are mostly dependent on nonglycemic parameters in 9398 Finnish men without diabetes. J Clin Endocrinol Metab 100**:** 1989-1996.

Floor, S. N., and J. A. Doudna, 2016 Tunable protein synthesis by transcript isoforms in human cells. Elife 5.

Florez, J. C., A. K. Manning, J. Dupuis, J. McAteer, K. Irenze *et al.*, 2007 A 100K genome-wide association scan for diabetes and related traits in the Framingham Heart Study: replication and integration with other genome-wide datasets. Diabetes 56**:** 3063-3074.

Franklin, C. S., Y. S. Aulchenko, J. E. Huffman, V. Vitart, C. Hayward *et al.*, 2010 The TCF7L2 diabetes risk variant is associated with HbA(1)(C) levels: a genome-wide association meta-analysis. Ann Hum Genet 74**:** 471-478.

Fryett, J. J., J. Inshaw, A. P. Morris and H. J. Cordell, 2018 Comparison of methods for transcriptome imputation through application to two common complex diseases. Eur J Hum Genet.

Gale, E. A., 2002 The rise of childhood type 1 diabetes in the 20th century. Diabetes 51**:** 3353-3361.

Gamazon, E. R., H. E. Wheeler, K. P. Shah, S. V. Mozaffari, K. Aquino-Michaels *et al.*, 2015 A gene-based association method for mapping traits using reference transcriptome data. Nat Genet 47**:** 1091-1098.

Genomes Project, C., G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks *et al.*, 2010 A map of human genome variation from population-scale sequencing. Nature 467**:** 1061-1073.

Genomes Project, C., G. R. Abecasis, A. Auton, L. D. Brooks, M. A. DePristo *et al.*, 2012 An integrated map of genetic variation from 1,092 human genomes. Nature 491**:** 56-65.

Genomes Project, C., A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison *et al.*, 2015 A global reference for human genetic variation. Nature 526**:** 68-74.

Genuth, S., W. Sun, P. Cleary, X. Gao, D. R. Sell *et al.*, 2015 Skin advanced glycation end products glucosepane and methylglyoxal hydroimidazolone are independently associated with long-term microvascular complication progression of type 1 diabetes. Diabetes 64**:** 266-278.

Group, D. P., 2006 Incidence and trends of childhood Type 1 diabetes worldwide 1990-1999. Diabet Med 23**:** 857-866.

group, T. D. r., 1987 Feasibility of centralized measurements of glycated hemoglobin in the Diabetes Control and Complications Trial: a multicenter study. . Clin Chem 33**:** 2267-2271.

Gusev, A., A. Ko, H. Shi, G. Bhatia, W. Chung *et al.*, 2016 Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet 48**:** 245-252.

Halperin, E., G. Kimmel and R. Shamir, 2005 Tag SNP selection in genotype data for maximizing SNP prediction accuracy. Bioinformatics 21 Suppl 1**:** i195-203.

Harrow, J., A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans *et al.*, 2012 GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res 22**:** 1760-1774.

Herman, W. H., 2016 Are There Clinical Implications of Racial Differences in HbA1c? Yes, to Not Consider Can Do Great Harm! Diabetes Care 39**:** 1458-1461.

Herst, P. M., M. R. Rowe, G. M. Carson and M. V. Berridge, 2017 Functional Mitochondria in Health and Disease. Front Endocrinol (Lausanne) 8**:** 296.

Hindorff, L. A., P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta *et al.*, 2009 Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106**:** 9362-9367.

Hoffman, J. D., R. E. Graff, N. C. Emami, C. G. Tai, M. N. Passarelli *et al.*, 2017 Cis-eQTL-based trans-ethnic meta-analysis reveals novel genes associated with breast cancer risk. PLoS Genet 13**:** e1006690.

Hohman, T. J., L. Dumitrescu, N. J. Cox, A. L. Jefferson and I. Alzheimer's Neuroimaging, 2017 Genetic resilience to amyloid related cognitive decline. Brain Imaging Behav 11**:** 401-409.

Hormozdiari, F., M. van de Bunt, A. V. Segre, X. Li, J. W. J. Joo *et al.*, 2016 Colocalization of GWAS and eQTL Signals Detects Target Genes. Am J Hum Genet 99**:** 1245-1260.

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

Hrdlickova, B., R. C. de Almeida, Z. Borek and S. Withoff, 2014 Genetic variation in the non-coding genome: Involvement of micro-RNAs and long non-coding RNAs in disease. Biochim Biophys Acta 1842**:** 1910-1922.

Hyttinen, V., J. Kaprio, L. Kinnunen, M. Koskenvuo and J. Tuomilehto, 2003 Genetic liability of type 1 diabetes and the onset age among 22,650 young Finnish twin pairs: a nationwide follow-up study. Diabetes 52**:** 1052-1055.

Insel, R. A., J. L. Dunne, M. A. Atkinson, J. L. Chiang, D. Dabelea *et al.*, 2015 Staging presymptomatic type 1 diabetes: a scientific statement of JDRF, the Endocrine Society, and the American Diabetes Association. Diabetes Care 38**:** 1964-1974.

International HapMap, C., 2003 The International HapMap Project. Nature 426**:** 789-796.

International HapMap, C., 2005 A haplotype map of the human genome. Nature 437**:** 1299-1320.

International HapMap, C., D. M. Altshuler, R. A. Gibbs, L. Peltonen, D. M. Altshuler *et al.*, 2010 Integrating common and rare genetic variation in diverse human populations. Nature 467**:** 52-58.

Jacobson, A. M., B. H. Braffett, P. A. Cleary, R. A. Gubitosi-Klug, M. E. Larkin *et al.*, 2013 The long-term effects of type 1 diabetes treatment and complications on health-related quality of life: a 23-year follow-up of the Diabetes Control and Complications/Epidemiology of Diabetes Interventions and Complications cohort. Diabetes Care 36**:** 3131-3138.

Jerram, S. T., and R. D. Leslie, 2017 The Genetic Architecture of Type 1 Diabetes. Genes (Basel) 8.

Kang, H. M., M. Subramaniam, S. Targ, M. Nguyen, L. Maliskova *et al.*, 2018 Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nat Biotechnol 36**:** 89-94.

Kapoor, A., R. B. Sekar, N. F. Hansen, K. Fox-Talbot, M. Morley *et al.*, 2014 An enhancer polymorphism at the cardiomyocyte intercalated disc protein NOS1AP locus is a major regulator of the QT interval. Am J Hum Genet 94**:** 854-869.

Kasowski, M., F. Grubert, C. Heffelfinger, M. Hariharan, A. Asabere *et al.*, 2010 Variation in transcription factor binding among humans. Science 328**:** 232-235.

King, M. C., and A. C. Wilson, 1975 Evolution at two levels in humans and chimpanzees. Science 188**:** 107-116.

Kondrashova, A., A. Reunanen, A. Romanov, A. Karvonen, H. Viskari *et al.*, 2005 A six-fold gradient in the incidence of type 1 diabetes at the eastern border of Finland. Ann Med 37**:** 67-72.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody *et al.*, 2001 Initial sequencing and analysis of the human genome. Nature 409**:** 860-921.

Lappalainen, T., M. Sammeth, M. R. Friedlander, P. A. t Hoen, J. Monlong *et al.*, 2013 Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501**:** 506-511.

Larsen, M. L., M. Horder and E. F. Mogensen, 1990 Effect of long-term monitoring of glycosylated hemoglobin levels in insulin-dependent diabetes mellitus. N Engl J Med 323**:** 1021-1025.

Leete, P., A. Willcox, L. Krogvold, K. Dahl-Jorgensen, A. K. Foulis *et al.*, 2016 Differential Insulitic Profiles Determine the Extent of beta-Cell Destruction and the Age at Onset of Type 1 Diabetes. Diabetes 65: 1362-1369.

Lek, M., K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks *et al.*, 2016 Analysis of protein-coding genetic variation in 60,706 humans. Nature 536: 285-291.

Lenz, T. L., A. J. Deutsch, B. Han, X. Hu, Y. Okada *et al.*, 2015 Widespread non-additive and interaction effects within HLA loci modulate the risk of autoimmune diseases. Nat Genet 47: 1085-1090.

Lettre, G., C. Lange and J. N. Hirschhorn, 2007 Genetic model testing and statistical power in population-based association studies of quantitative traits. Genet Epidemiol 31: 358-362.

Lewontin, R. C., 1988 On measures of gametic disequilibrium. Genetics 120: 849-852.

Maahs, D. M., N. A. West, J. M. Lawrence and E. J. Mayer-Davis, 2010 Epidemiology of type 1 diabetes. Endocrinol Metab Clin North Am 39: 481-497.

MacArthur, J., E. Bowler, M. Cerezo, L. Gil, P. Hall *et al.*, 2017 The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res 45: D896-D901.

Maier, T., M. Guell and L. Serrano, 2009 Correlation of mRNA and protein in complex biological samples. FEBS Lett 583: 3966-3973.

Mannering, S. I., V. Pathiraja and T. W. Kay, 2016 The case for an autoimmune aetiology of type 1 diabetes. Clin Exp Immunol 183: 8-15.

Manolio, T. A., 2013 Bringing genome-wide association findings into clinical use. Nat Rev Genet 14: 549-558.

Maurano, M. T., R. Humbert, E. Rynes, R. E. Thurman, E. Haugen *et al.*, 2012 Systematic localization of common disease-associated variation in regulatory DNA. Science 337: 1190-1195.

Meigs, J. B., C. I. Panhuysen, R. H. Myers, P. W. Wilson and L. A. Cupples, 2002 A genome-wide scan for loci linked to plasma levels of glucose and HbA(1c) in a community-based sample of Caucasian pedigrees: The Framingham Offspring Study. Diabetes 51: 833-840.

Miettinen, M. E., L. Reinert, L. Kinnunen, V. Harjutsalo, P. Koskela *et al.*, 2012 Serum 25-hydroxyvitamin D level during early pregnancy and type 1 diabetes risk in the offspring. Diabetologia 55: 1291-1294.

Mifsud, B., F. Tavares-Cadete, A. N. Young, R. Sugar, S. Schoenfelder *et al.*, 2015 Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat Genet 47: 598-606.

Mignone, F., C. Gissi, S. Liuni and G. Pesole, 2002 Untranslated regions of mRNAs. Genome Biol 3: REVIEWS0004.

Musunuru, K., A. Strong, M. Frank-Kamenetsky, N. E. Lee, T. Ahfeldt *et al.*, 2010 From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature 466: 714-719.

Nathan, D. M., and D. E. R. Group, 2014 The diabetes control and complications trial/epidemiology of diabetes interventions and complications study at 30 years: overview. Diabetes Care 37: 9-16.

Nica, A. C., S. B. Montgomery, A. S. Dimas, B. E. Stranger, C. Beazley *et al.*, 2010 Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. PLoS Genet 6**:** e1000895.

Nicolae, D. L., E. Gamazon, W. Zhang, S. Duan, M. E. Dolan *et al.*, 2010 Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet 6**:** e1000888.

Norris, J. M., X. Yin, M. M. Lamb, K. Barriga, J. Seifert *et al.*, 2007 Omega-3 polyunsaturated fatty acid intake and islet autoimmunity in children at increased risk for type 1 diabetes. JAMA 298**:** 1420-1428.

Oilinki, T., T. Otonkoski, J. Ilonen, M. Knip and P. J. Miettinen, 2012 Prevalence and characteristics of diabetes among Somali children and adolescents living in Helsinki, Finland. Pediatr Diabetes 13**:** 176-180.

Pare, G., D. I. Chasman, A. N. Parker, D. M. Nathan, J. P. Miletich *et al.*, 2008 Novel association of HK1 with glycated hemoglobin in a non-diabetic population: a genome-wide evaluation of 14,618 participants in the Women's Genome Health Study. PLoS Genet 4**:** e1000312.

Paterson, A. D., 2017 HbA1c for type 2 diabetes diagnosis in Africans and African Americans: Personalized medicine NOW! PLoS Med 14**:** e1002384.

Paterson, A. D., D. Waggott, A. P. Boright, S. M. Hosseini, E. Shen *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**:** 539-549.

Peng, G., L. Luo, H. Siu, Y. Zhu, P. Hu *et al.*, 2010 Gene and pathway-based second-wave analysis of genome-wide association studies. Eur J Hum Genet 18**:** 111-117.

Pickrell, J., 2015 Fulfilling the promise of Mendelian randomization. bioRxiv.

Pilia, G., W. M. Chen, A. Scuteri, M. Orru, G. Albai *et al.*, 2006 Heritability of cardiovascular and personality traits in 6,148 Sardinians. PLoS Genet 2**:** e132.

Pociot, F., and A. Lernmark, 2016 Genetic risk factors for type 1 diabetes. Lancet 387**:** 2331-2339.

Prabhakar, S., A. Visel, J. A. Akiyama, M. Shoukry, K. D. Lewis *et al.*, 2008 Human-specific gain of function in a developmental enhancer. Science 321**:** 1346-1350.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira *et al.*, 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81**:** 559-575.

Redondo, M. J., J. Jeffrey, P. R. Fain, G. S. Eisenbarth and T. Orban, 2008 Concordance for islet autoimmunity among monozygotic twins. N Engl J Med 359**:** 2849-2850.

Redondo, M. J., L. Yu, M. Hawa, T. Mackenzie, D. A. Pyke *et al.*, 2001 Heterogeneity of type I diabetes: analysis of monozygotic twins in Great Britain and the United States. Diabetologia 44**:** 354-362.

Rewers, M., and J. Ludvigsson, 2016 Environmental risk factors for type 1 diabetes. Lancet 387**:** 2340-2348.

Risch, N., and K. Merikangas, 1996 The future of genetic studies of complex human diseases. Science 273**:** 1516-1517.

Romero, I. G., I. Ruvinsky and Y. Gilad, 2012 Comparative studies of gene expression and the evolution of gene regulation. Nat Rev Genet 13**:** 505-516.

Ryu, J., and C. Lee, 2012 Association of glycosylated hemoglobin with the gene encoding CDKAL1 in the Korean Association Resource (KARE) study. Hum Mutat 33**:** 655-659.

Sachidanandam, R., D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein *et al.*, 2001 A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409**:** 928-933.

Saleh, J., 2015 Glycated hemoglobin and its spinoffs: Cardiovascular disease markers or risk factors? World J Cardiol 7**:** 449-453.

Samocha, K. E., E. B. Robinson, S. J. Sanders, C. Stevens, A. Sabo *et al.*, 2014 A framework for the interpretation of de novo mutation in human disease. Nat Genet 46**:** 944-950.

Scharfe, C., H. H. Lu, J. K. Neuenburg, E. A. Allen, G. C. Li *et al.*, 2009 Mapping gene associations in human mitochondria using clinical disease phenotypes. PLoS Comput Biol 5**:** e1000374.

Schmidt, D., M. D. Wilson, B. Ballester, P. C. Schwalie, G. D. Brown *et al.*, 2010 Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science 328**:** 1036-1040.

Scotti, M. M., and M. S. Swanson, 2016 RNA mis-splicing in disease. Nat Rev Genet 17**:** 19-32.

Selvin, E., 2016 Are There Clinical Implications of Racial Differences in HbA1c? A Difference, to Be a Difference, Must Make a Difference. Diabetes Care 39**:** 1462-1467.

Seyerle, A. A., and C. L. Avery, 2013 Genetic epidemiology: the potential benefits and challenges of using genetic information to improve human health. N C Med J 74**:** 505-508.

Sherry, S. T., M. H. Ward, M. Kholodov, J. Baker, L. Phan *et al.*, 2001 dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29**:** 308-311.

Sim, N. L., P. Kumar, J. Hu, S. Henikoff, G. Schneider *et al.*, 2012 SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res 40**:** W452-457.

Soderstrom, U., J. Aman and A. Hjern, 2012 Being born in Sweden increases the risk for type 1 diabetes - a study of migration of children to Sweden as a natural experiment. Acta Paediatr 101**:** 73-77.

Soranzo, N., S. Sanna, E. Wheeler, C. Gieger, D. Radke *et al.*, 2010 Common variants at 10 genomic loci influence hemoglobin A(1)(C) levels via glycemic and nonglycemic pathways. Diabetes 59**:** 3229-3239.

Sorensen, I. M., G. Joner, P. A. Jenum, A. Eskild and L. C. Stene, 2012 Serum long chain n-3 fatty acids (EPA and DHA) in the pregnant mother are independent of risk of type 1 diabetes in the offspring. Diabetes Metab Res Rev 28**:** 431-438.

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

van Belle, T. L., K. T. Coppieters and M. G. von Herrath, 2011 Type 1 diabetes: etiology, immunology, and therapeutic strategies. Physiol Rev 91**:** 79-118.

van de Bunt, M., J. E. Manning Fox, X. Dai, A. Barrett, C. Grey *et al.*, 2015 Transcript Expression Data from Human Islets Links Regulatory Signals from Genome-Wide

Association Studies for Type 2 Diabetes and Glycemic Traits to Their Downstream Effectors. PLoS Genet 11: e1005694.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural *et al.*, 2001 The sequence of the human genome. Science 291: 1304-1351.

Visscher, P. M., M. A. Brown, M. I. McCarthy and J. Yang, 2012 Five years of GWAS discovery. Am J Hum Genet 90: 7-24.

Visscher, P. M., N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy *et al.*, 2017 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet 101: 5-22.

Walter, M., T. Kaupper, K. Adler, J. Foersch, E. Bonifacio *et al.*, 2010 No effect of the 1alpha,25-dihydroxyvitamin D3 on beta-cell residual function and insulin requirement in adults with new-onset type 1 diabetes. Diabetes Care 33: 1443-1448.

Wen, X., R. Pique-Regi and F. Luca, 2017 Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. PLoS Genet 13: e1006646.

Wheeler, E., A. Leong, C. T. Liu, M. F. Hivert, R. J. Strawbridge *et al.*, 2017a Impact of common genetic determinants of Hemoglobin A1c on type 2 diabetes risk and diagnosis in ancestrally diverse populations: A transethnic genome-wide meta-analysis. PLoS Med 14: e1002383.

Wheeler, H. E., E. R. Gamazon, R. D. Frisina, C. Perez-Cervantes, O. El Charif *et al.*, 2017b Variants in WFS1 and Other Mendelian Deafness Genes Are Associated with Cisplatin-Associated Ototoxicity. Clin Cancer Res 23: 3325-3333.

Wheeler, H. E., K. P. Shah, J. Brenner, T. Garcia, K. Aquino-Michaels *et al.*, 2016 Survey of the Heritability and Sparse Architecture of Gene Expression Traits across Human Tissues. PLoS Genet 12: e1006423.

White, N. H., 2015 Long-term Outcomes in Youths with Diabetes Mellitus. Pediatr Clin North Am 62: 889-909.

Wray, N. R., J. Yang, B. J. Hayes, A. L. Price, M. E. Goddard *et al.*, 2013 Pitfalls of predicting complex traits from SNPs. Nat Rev Genet 14: 507-515.

Yang, J., S. H. Lee, M. E. Goddard and P. M. Visscher, 2011a GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 88: 76-82.

Yang, J., M. N. Weedon, S. Purcell, G. Lettre, K. Estrada *et al.*, 2011b Genomic inflation factors under polygenic inheritance. Eur J Hum Genet 19: 807-812.

Zhang, Z., and M. Gerstein, 2003 Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements. J Biol 2: 11.

Zhou, X., P. Carbonetto and M. Stephens, 2013 Polygenic modeling with bayesian sparse linear mixed models. PLoS Genet 9: e1003264.

Ziegler, A. G., M. Rewers, O. Simell, T. Simell, J. Lempainen *et al.*, 2013 Seroconversion to multiple islet autoantibodies and risk of progression to diabetes in children. JAMA 309: 2473-2479.

Zipitis, C. S., and A. K. Akobeng, 2008 Vitamin D supplementation in early childhood and risk of type 1 diabetes: a systematic review and meta-analysis. Arch Dis Child 93: 512-517.

Zuk, O., S. F. Schaffner, K. Samocha, R. Do, E. Hechter *et al.*, 2014 Searching for missing heritability: designing rare variant association studies. Proc Natl Acad Sci U S A 111**:** E455-464.