Statistical Methods for Analyzing Population-scale Genomic and Transcriptomic Data

Andrew Liu

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

Doctoral Committee:

Associate Professor Hyun Min Kang (Chair)
Professor Matthias Kretzler
Associate Professor Stephen Parker
Professor Peter Song
Associate Professor Xiaoquan Wen

Andrew Liu

aeyliu@umich.edu

ORCID iD: 0000-0001-5522-1263

Dedication

I dedicate this dissertation to my family – my late paternal grandparents, my parents, and my brother – who have given me unconditional support and love throughout every stage of my life. If not for their hard work, sacrifice, and perseverance, it would be unimaginable to be where we are right now. Words cannot begin to describe the story of our lineage, the trials, and tribulations we have been through. Starting from humble roots in Shanghai, my family has survived through the Second World War, the cultural revolution, and eventually found our way to Canada as poor immigrants. Throughout it all, they have always taught me to value education, honesty, integrity, kindness, compassion, and above all, love. Now as I stand on the precipice of the biggest milestone of my life thus far, I reflect on all they have given me and feel the deepest appreciation and gratitude. I owe everything that I have to them, and I hope I have made them proud. I will continue carrying on our legacy forward to the best my of ability. No matter where life takes me, I will never forget where we came from and they will always be in my heart.

Acknowledgements

Throughout the years, I have received so much support in both my personal and academic life. I would first and foremost like to thank my advisor, Dr. Hyun Min Kang for taking me as his student, spending countless hours teaching me many topics related to statistical genetics and providing academic guidance. If it were not for your patience, understanding, kindness and encouragement I would not be where I am today. Next, I would like to thank Dr. Matthias Kretzler and his lab, Dr. Robert G. Nelson and colleagues at the NIH. Much of the material from this dissertation is drawn from the Pima study, and I am grateful for having opportunity to collaborate on this study. Finally, I would like to extend my sincerest gratitude towards my dissertation committee – Dr. Hyun Min Kang, Dr. Matthias Kretzler, Dr. Stephen Parker, Dr. Peter Song, and Dr. Xiaoquan Wen – for taking the time to review my work and providing input and encouragement throughout this process.

I would also like to thank my family, who have encouraged me through all my life's journeys. I would like to thank my father – who spent so many hours teaching me math as a kid and for guiding me throughout my academic journey – my mother – who has provided me moral support and ensured I never lost sight of what is truly important in life – and countless people outside my family who have become like my family.

Table of Contents

List of Figures

List of Tables

Abstract

The study of genetics is an integral part to understanding the biology behind our complex traits and can be approached in a variety of ways. Technological advancements in the field of genomics have enabled unprecedented large-scale studies which have identified numerous statistical associations between many diseases and our genes. Recently, studies involving gene expression have become an increasingly popular approach to understanding the biological pathways underlying statistical associations. In this dissertation, I address specific challenges related to the study of gene expression, including meta-imputation of expression across multiple datasets with only summary-level imputation models available, correcting for technical biases towards reference alleles in array-based expression assays, and identifying tissue-specific and population-specific regulatory variants and trait-associated loci in the context of systems genetics with whole genome sequencing, transcriptomics profiles, morphometric traits, and clinical endpoints.

In Chapter 2, I develop a method which leverages multiple datasets to accurately impute tissue-specific gene expression levels. Our method, Smartly Weighted Averaging across Multiple Tissues (SWAM) does not train directly from data, but rather performs a meta-imputation by combines extant imputation models by assigning weights based on their predictive performance and similarity to the tissue of interest. I demonstrate that when using the same set of resources, SWAM improves imputation accuracy compared to existing approaches that impute tissue-specific expression by training directly from raw data. The major benefit of using the SWAM meta-imputation framework is the flexibility to combine multiple pre-trained imputation models trained from privacy-protected raw datasets. Indeed, prediction accuracy is substantially improved when integrating multiple datasets, highlighting the importance of using multiple datasets.

In Chapter 3, I examine the benefits of using deep whole genome sequencing to empower and refine existing microarray-based eQTL studies. I revisited a well-known hybridization bias that arises in microarray studies caused by genetic polymorphisms within target probe sequences. In this chapter, I interrogated the impact of genetic variants from whole genome sequencing to accurately identify and characterize this bias at both the probe and probeset level. I evaluated several approaches to account for hybridization bias, including methods to remove variant-overlapping probes, and a novel method to adjust hybridization bias for each probe. I demonstrate that accounting for variant-overlapping probes when quantifying expression levels reduces reference bias and false positives in cis-eQTL analyses. I also demonstrate that adjusting for hybridization bias with deeply sequenced genomes is ideal to avoid reference bias, although leveraging publicly available variant catalogues such as the 1000 Genomes data provides comparable benefits.

In Chapter 4, I performed a systems genetic study of Pima Native Americans enrolled in a diabetic nephropathy study. I integrate whole genome sequences, transcriptomic profiles, and morphometric traits derived from two micro-dissected renal compartments – glomerular and tubulointerstitial – and clinical phenotypes to identify significant associations between these molecular and complex traits. I identified thousands of eQTLs, including kidney-specific and population-specific eQTLs. I also identified many transcriptional associations with morphometric and clinical phenotypes enriched for kidney-specific biological pathways. Moreover, through dimension reduction techniques, I identified genome-wide significant genetic associations with a morphometric trait (podocyte volume), and with a composite trait representing albumin-creatin ration and glomerular surface volume, which was obtained from dimensionality reduction techniques. Studying this unique and richly-phenotyped cohort resulted many population- and tissue-specific regulatory variants, genes, and pathways implicated for renal disease progression.

Chapter 1 Introduction

1.1 Systems genetics: an overview

Genetics is a subject of biology in which we seek to understand genes, which are basic physical units of inheritance and play a major role in the manifestation of traits in living organisms. One major focus in this field is to understand how differences in our genome (DNA variation) affect complex traits. The study of genetics has many important health and medical implications, such as determining genetic pre-disposition to various diseases, and characterizing response to drug treatment. Systems genetics is a study approach which seeks to holistically understand the causal biological pathways that connect our DNA to endpoint traits. By examining many molecular phenotypes such as gene expression, epigenomic marks, protein levels, and metabolite abundance, we gain a deeper understanding of the complicated biology underlying many diseases [1]. Although the study of genetics has long pre-dated our knowledge of the existence of DNA [2], recent rapid developments in technology have facilitated unprecedented research in this topic, providing a high resolution view of many molecular traits. For example, with advances in DNA sequencing technology, we have been able to conduct large-scale genetic studies for many diseases, detecting numerous genetic variants that could potentially influence the disease [3,4]. Advances in technology for gene expression assays have allowed us to study one of the very important intermediate phenotypes in systems genetics [5]. From these developments, thousands of trait-associated genetic loci are being mapped to regulation of gene expression, which in turn directly affect protein building and cell function [6,7].

In this thesis, I delve into some of the topics and challenges that arise in the study of systems genetics, particularly focusing statistical and computational aspects of gene expression studies. In this chapter, I provide a background on the history of related fields, technical developments

and challenges that have arisen in gene expression. We then provide an overview of some of the work we have contributed and how they address some the challenges faced in systems genetics studies.

1.2 Genome-Wide Association Studies (GWAS)

Genome-Wide Association Studies (GWAS) examine the statistical association between many genetic loci and traits. This approach became very popular starting in the early 2000s due to massive improvements in SNP array genotyping and whole-genome sequencing (WGS) technologies. In the past 15+ years, SNP-arrays have been widely used to study the effects of common genetic variants at a large scale. For example, multiple genetic susceptibility variants were identified in a study of type 2 diabetes for over 2,000 Finnish individuals, where SNP-arrays were used to genotype over 300,000 markers [8]. In 2015, a meta-analysis study was performed on over 300,000 individuals, and identified 97 genetic loci associated with obesity [9]. One limitation of SNP-arrays is that it only prior known genomic locations can be genotyped, and rare and population-specific variants can be missed [10].

Another approach to obtain genotypes is whole-genome sequencing, which seeks to characterize the genome of an individual down to a single base-pair resolution. This approach allows for detection of rare and population-specific variants. The history of WGS dates back to the 1990s, where many viruses [11] and bacteria were fully sequenced for the first time, along with a few animals. The Human Genomes project, completed in 2003, was the world's largest collaborative biological project, with the goal of mapping every gene within the human genome. This project however, used Sanger sequencing, which is extremely labor-intensive and low throughput, which would not be viable to study high number of individuals [12]. The advent of short-read sequencing technologies has enabled re-sequencing human genomes in an affordable, massively parallel manner, allowing for population-scale genetic studies. For example, the 1000 Genomes project sought to provide a detailed catalogue of human variation across 2,504 human genomes in 26 populations [13,14]. Recently, the Genome Aggregation Database (gnomAD) has aggregated 125,748 exomes and 15,708 genomes from various human sequencing studies and have identified over 750 million variants, including >400,000 loss-of-

function variants [15]. In addition to providing a high-resolution view in GWAS studies, WGS has also allowed for imputation of genotypes of individuals who were array genotyped. There are now freely available genotype imputation servers which can impute genotypes based on various populations from different reference panels [16].

Currently, the NHGRI-EBI GWAS catalog has publicly available information on >227,000 significant associations across >4,800 (as of Dec 15, 2020) [3,4]. Despite the numerous association signals detected by GWAS, there are still many challenges and the biology behind these associations are still not clearly understood. For example, heritability (which can be calculated without genotypes) for many traits and complex diseases have not been fully accounted for from GWAS alone [17]. A very well-known example is human height, which has an estimated 80% heritability (proportion of variation explainable by genetic variation), yet only 25%-50% of this heritability has been explained by genetic variants [18,19]. One plausible explanation is that many rare variants with high effect size have yet to be discovered from GWAS. Other hypotheses posit that there could be thousands or even millions of variants that all contribute a very small amount of heritability to each trait [20]. Another possible explanation could be attributed to trait heterogeneity, and that trait definitions could be subjective or inconsistent within the same study. In such scenarios, studying intermediate phenotypes such as objective biomarkers or gene expression could potentially provide better insight compared to using endpoint traits. There are other open-ended questions for GWAS that have also been discussed, including the notion that GWAS signals are often not easily interpretable in a biological setting. Some of the reasons for this include linkage-disequilibrium (LD) structures between associated variants, which confounds identification of causal variants. Other reasons include the lack of our understanding on the function of each individual genetic variants beyond the protein-coding regions of the genome, and the complex causal pathways that connect genetic variation to end-point phenotypes [21]. Despite the tremendous successes from GWAS, it is evident that the biological process that links our genotypes to traits is extremely complex. To fully understand the genetic architecture underlying complex traits, it is important to study the intermediate phenotypes that link these two endpoints together, such as gene expression.

1.3 Gene expression: an intermediate phenotype to understand GWAS signals

The functional mechanisms behind trait-associations are very diverse and many avenues of research can be taken to understand their underlying biology. For example, some genetic variants have been shown to directly knockout a gene or disrupt protein function, such as nonsense mutations from the *PCSK9* lowering plasma levels of LDL cholesterol [22]. There has also been mounting evidence that many GWAS risk variants are located outside of coding regions and either co-localize with, or directly regulate gene expression levels [23–26]. To understand the functional aspect of these variants, there are many different intermediate phenotypes that that can be studied, including gene expression (characterized by eQTLs), DNA methylation (meQTLs) [27,28], chromatic accessibility marks (caQTLs) [29,30], and protein levels (pQTLs)[31]. Among these, studies involving gene expression levels have been very popular and have played a central role in understanding biological pathways behind many traits [25,32,33].

The genetic study of gene expression – originally coined as *genetical genomics* [34], and is a part *systems genetics* – has been important in unraveling the complex interaction between our genes, our environment, and many diseases [35]. While our genotypes provide the blueprint for protein coding which in turn affects our traits, these genes must be "expressed" before the phenotype becomes apparent [36,37]. Transcription is the process of copying DNA and converting it into RNA, is a fundamental unit for translation of DNA into proteins and enzymes, which eventually affects phenotypes and clinical endpoints [38]. While our DNA is the same in every cell, gene expression can be different based on cell function. Therefore, studying gene expression can provide insight into the differences between our tissues and cell types, allowing us to compare tissue-specific profiles, which may be more relevant to the trait of interest compared to studying genotypes alone.

1.4 Expression Quantitative Trait Loci Studies (eQTLs): background

Expression quantitative trait loci (eQTL) mapping is an approach which seeks to understand the regulatory function of genetic variants, and to determine regions of the genome that affect transcription. This is typically done by treating gene expression levels as a quantitative trait, and

calculating the statistical association between expression and genotypes [39,40]. A genetic variant can regulate expression levels both proximally (cis-) or distally (trans-). So far, most of the focus in eQTL studies have been for cis-acting variants due to the limited statistical power to detect trans-eQTLs with limited sample sizes [41,42]. Furthermore, cis-eQTLs have been shown to have higher effect sizes, and a substantial proportion of trans-eQTLs is found to be mediated by cis-eQTLs [43,44], For the work conducted in this thesis, we primarily focus on cis-acting eQTLs due to higher power for detection, and the availability of cis-eQTL repositories.

Over the years, many eQTL studies have been conducted and many of these studies have managed to provide interpretable insight into GWAS signals. The first genome-wide mapping of expression levels was performed in 2002 in a genetic linkage study for two strains of yeast [45]. Since then, eQTL studies have been carried out for various cell types in many organisms, including mice and humans [46,47]. In 2007, a study mapping genetic loci with expression levels of genes in EBV-transformed lymphoblastoid cell lines have been able to explain GWAS association signals in childhood asthma [48]. Another study found that variants associated with Crohn's disease were likely to be regulatory variants for the PTGER4 gene [49].

Initially, many eQTL studies performed on humans were based on blood-derived cell types, due to the ease of collection [39]. However, it has been shown that studying the most relevant tissue to the trait in question would provide greater insight into clinical traits. For example, *Emilsson et al.* demonstrated that expression levels of genes for the adipose tissue were correlated with over 50% of obesity related traits, whereas only 10% of blood-derived gene expression levels were correlated with these same traits [23]. With the decreasing cost and increasing availability of obtaining expression data, eQTL databases have been generated for many different tissue types. For example, the GEUVADIS consortium has generated an eQTL repository on lymphoblastoid cell lines for 462 individuals from the 1000 Genomes project [47]. The Depression Genes and Networks (DGN) cohort has 922 participants with RNA sequencing for whole blood, and have discovered over 10,000 eGenes regulated by genetic variation [50]. The GTEx consortium is an on-going project which initially assayed 44 tissues spanning the blood, digestive, respiratory, reproductive, brain and many other tissue types. The list of tissues

has since expanded with more samples being included into the study, with 49 tissues currently having enough sample size to conduct eQTL analysis [51,52].

1.5 Transcriptome-Wide Association Studies (TWAS)

With the increasing availability of external eQTL and measured gene expression reference panels, transcriptome-wide association studies (TWAS) have become popular in recent years. The objective of TWAS is to leverage eQTL (or measured gene expression) information to elucidate the regulatory aspect for many GWAS risk variants [53]. Instead of using genotypes as explanatory variables, TWAS examines the association between trait and gene expression. This is typically done by imputing expression levels from individual-level genotypes and performing association analysis between imputed expression and traits. Because the imputed expression is a function of genotypes, TWAS essentially assigns scores to genetic loci based on their impact on gene regulation. As a result, the association signals found in TWAS are mostly driven by regulatory variants, providing biological insight for many of these GWAS signals [54]. In terms of power, TWAS has a much lower multiple testing burden as genes number in the tens of thousands, as opposed to the millions of SNPs often tested in GWAS. However, power can be lost for signals that are driven by non-regulatory associations.

We do note that while TWAS can be conducted using measured expression, predicted expression is often preferred for several reasons. The first is that genotype data are typically easier and much more feasible to collect compared to tissue-specific expression data. It is overall much more cost effective and to obtain genotype data and use external eQTL databases to impute the expression. Secondly, predicted expression in theory should capture only the genetic regulated component of expression, and should be impervious to potential confounders such as environmental effects [55,56]. Finally, significant associations using predicted expression can be linked to specific genetic markers which can be cross-referenced with GWAS signals. This is particularly useful for determining potential causal candidate SNPs in cases of high linkage-disequilibrium between significant GWAS variants.

Since using imputed expression is the preferred method for conducting TWAS, accurate and powerful eQTL discovery is essential for these studies. A recently developed and widely used tool to impute expression is PrediXcan, which first uses an elastic net to detect cis-eQTLs from tissue-specific expression and genotypes. Next, a prediction database is generated and PrediXcan can automatically create a file with imputed expression levels using individual-level genotypes as an input [55]. Therefore, this tool can leverage many of the previously generated gene expression resources such as GTEx, DGN and GEUVADIS. For example, the authors used PrediXcan to create prediction databases for 44 GTEx tissues (now 49 from GTEx version 8) as well as the whole-blood tissue type from the DGN (depression gene network) cohort, for all tissue-specific "well-predicted" genes (cross-validated R-squared > 0.01). Using these prediction databases to impute expression levels, the authors performed TWAS on seven diseases from the Wellcome Trust Case Control Consortium (WTCCC) study [58], and identified 29 genes associated with type 1 diabetes, with numerous other genes being associated with autoimmune diseases. In 2016, SLINGER, an extension to PrediXcan was developed where the cis-requirement for eQTL discovery was removed. The authors demonstrated that prediction accuracy was improved, increasing the number of estimable genes by more than 2,000 for the DGN whole blood expression data. Furthermore, TWAS conducted on the 7 same WTCCC traits displayed significantly elevated r^2 with many associations being highly reflective of actual variation in expression levels [58].

As the current pool of transcriptomic resources continues to expand and become higher quality, TWAS with predicted expression will become increasingly useful as a means to uncover the regulatory aspect of genetic association.

1.6 The evolution of gene expression technologies

1.6.1 Array-based expression profiling

Traditionally, gene expression levels have been measured using microarrays. This technology was developed in the early 1990s where *Fodor et al* demonstrated that short DNA or RNA

molecules (*oligonucleotides*) could be synthesized onto a glass slide through photolithography [59]. This allowed for miniaturization of the chip, which *Schena et al* in 1994 demonstrated would accommodate high-capacity parallelization of multiple genes [60]. Modern day microarrays typically use short probe sequences known as *features*, which are designed to hybridize with specific known gene regions. The quantification of hybridization of probes compares the relative color intensity of a perfect match probe against a mismatch probe (serving as a baseline), which can then be converted into expression levels using various statistical approaches. This high-throughput method provides a snapshot of the overall gene expression profile of an isolated tissue sample that the researcher is studying.

Since the focus of gene expression experiments are to capture meaningful biological variation between individuals, we ideally want to have a high signal-to-noise ratio. Unfortunately, the microarray technology is highly susceptible to systematic biases which may affect expression estimates. For example, lab conditions and protocols may contribute to systematic differences (known as *batch effects*) between microarray experiments [61]. Furthermore, microarrays may have high sensitivity of the experimental setup to variations in hybridization temperature [62]. In addition, the purity and degradation rate of genetic material [62], and the amplification process [63], may also impact the estimates of gene expression. There have been studies on the lab effects on the quality of gene expression data. For example, *Beekman et al.* demonstrated that to minimize variation, the experiments should be ideally performed in the same lab. However, they also showed that the interlaboratory findings were also generally consistent when the correct statistical methods were applied [64]. *Dobbin et al.* found high between-laboratory concordance for individuals when the same protocols were followed for each lab [65]. In addition to standardization of lab protocols to minimize experimental variation, various statistical methods have been used to normalize microarray data. For example, *Bolstad et al* showed that a quantile-normalization approach for probe intensity values produced the low variance and bias between different arrays, while also being computationally fast [66]. Over the years, other gold standard approaches have been developed such as the Robust Multi-array Averaging (RMA) method which using a median polish approach to convert probe-level data into probeset (or gene) level expression [67]. In 2007, *Johnson et al* implemented both

parametric and non-parametric Bayesian frameworks to combine probeset-level data across multiple microarray platforms [68].

In addition to systematic batch effects, microarrays are also susceptible to cross-hybridization, where unintended sequences hybridize to a probe, artificially inflating the probe intensity levels [69]. This also creates a non-independence between probes, as well as high background levels which limit the ability to detect a high range of difference between genes [70,71]. Indeed, the background noise and cross-hybridization makes it difficult for microarrays to differentiate between low-abundance versus non-expressed transcripts [72]. Another well-known limitation of microarrays is a reduced hybridization for certain probes when individuals have sequence variation within the probe boundaries. This commonly can lead to negatively biased estimates in expression levels, which also could create false positives in association analyses.

1.6.2 RNA Sequencing

A more recent approach to measure gene expression has been to use genome sequencing technology to identify the quantity of RNA in a biological sample. By directly sequencing transcripts, we bypass the requirement of using interrogating probes. This helps overcome many of the limitations of microarrays, avoiding the need for a priori knowledge of RNA target sequences, and reducing susceptibility to hybridization issues. Initially, expression sequencing was done using Sanger sequencing to quantify levels of complementary DNA (cDNA). While Sanger sequencing is still viable on a smaller scale, this approach is low-throughput and has given way to newer methods [73,74]. Tag-based methods such as SAGE (serial analysis of gene expression) and CAGE (cap analysis of gene expression) were developed as a high-throughput methods which also provided precise quantification of expression levels [75,76]. These methods however are unable to discover novel genes and many short tags are unable to be uniquely mapped to the genome. Furthermore, these approaches were unable to distinguish between splice isoforms [77,78].

The development of next generation sequencing (NGS) technologies has greatly enabled the study of transcriptomics. RNA sequencing (RNA-seq) is a high throughput method which refers to the deep sequencing and quantification of (cDNA). These sequence fragments can be

assembled either using a reference genome or done using *de novo* sequencing. RNA-seq has been able to overcome many of the limitations from the older gene-expression technologies (both microarrary and older sequencing approaches) and revolutionizes the study of transcriptomics [77]. For example, RNA sequencing has been able to detect novel transcripts and isoforms, and reveal splice variants [72,79,80], giving it a distinct advantage over the Sanger, SAGE and CAGE sequencing approaches.

Compared to microarrays which require a priori knowledge of the sequences, RNA-seq directly identifies the transcript sequences [78]. RNA sequencing also provides a higher sensitivity to low and high levels of expression, which microarrays often cannot. Since RNA sequencing does not have an upper limit for quantification of sequences, we observe a high dynamic range of expression levels. For example, a 9000-fold range was detected for genes within the yeast genome [81], and a range of five orders of magnitude was detected for 40 million reads within the mouse genome [82]. Because of its high resolution, RNA-seq can also reveal the precise (1 base pair) location of transcript boundaries, give information on how exons are connected and reveal sequence variations [77,80].

RNA-seq traditionally has been performed using bulk tissue, which averages the expression levels over many cell types. Recently there has been evidence that gene expression can be heterogeneous between cells within the same tissue, which lead to substantial functional consequences [83–85]. The first study to profile gene expression using NGS at the cellular level was performed in 2009, using only a single mouse blastomere to detect over 5000 more expressed genes than compared to microarrays [86]. Since then, there have been a plethora of studies that profile expression at the single cell resolution, providing insight that would not be detectable at the bulk-cell level. For example, *Shaffer et al.* characterized the variability in melanoma cells at the single-cell level which predicted resistance to drug treatment [87]. Over the years, single cell RNA-seq (scRNA-seq) has been used for many applications, including tracing cell lineage and classifying cell types, as well as genomic profiling of rare cell types [88]. However, current challenges include cost of sequencing, and high levels of noise compared to bulk RNA-seq, resulting in computational and statistical challenges. As computational methods

improve and sequencing costs continue to decrease, scRNA-seq will provide even greater insight into cell biology and genetics [89].

1.7 Challenges

1.7.1 Accurate imputation of gene expression leveraging multiple datasets

The primary purpose of gene expression imputation is to harness naturally occurring genetic variation to understand the relationships between gene expression and complex traits through TWAS [55,56]. However, expression studies tend to have much smaller sample sizes compared to GWAS datasets, due to many challenges of obtaining RNA samples. For example, extracting RNA from various tissue types requires a biopsy of the tissue sample, which is far more difficult to perform on living individuals compared to obtaining their saliva or blood. In addition, RNA-seq experiments do not have the same level of automation compared to DNA sequencing or genotype-arrays, as lab protocols can differ in terms of extracting and storing many different tissue types. As a result, expression profiles can be heterogeneous across different batches or labs, which presents challenges in combining multiple datasets or performing meta-analysis. Indeed, RNA experiments are currently performed on a much smaller scale compared to those studying DNA. For example, GTEx has examined expression levels for 948 individuals with a tissue-maximum of 803 individuals (skeletal muscle tissue), while the UK Biobank phenome-wide study of depression contained >400,000 subjects [90,91]. Given these sample size differences [92], eQTL detection and hence gene expression prediction accuracy can be limited by the availability of high quality tissue data.

Given the current state of available gene expression data, there can be several ways to improve the power of TWAS based on imputed expression. One idea is to improve prediction accuracy by leveraging information from multiple tissues. This takes advantage of the idea that gene expression profiles can often be shared across different tissue types. This could make the downstream analysis (such as TWAS) much more powerful. Since the original PrediXcan paper [55], there have numerous publications extending the method to include multiple datasets or tissue types. For example, instead of training one tissue at a time, UTMOST jointly trains every tissue simultaneously, producing gene expression estimates for each tissue. While PrediXcan

performs a penalized regression across all genetic variants, UTMOST penalizes across both genetic variants and tissues. By doing so, this method captures the cross-tissue similarity for all genes, and improves imputation accuracy compared to using the single-tissue method [93]. Other methods use multiple single-tissue predictions to perform TWAS directly without further imputing multi-tissue gene expression. For example, MultiXcan (an extension of PrediXcan) improves power for TWAS by performing multivariate regression between predicted expression (derived from PrediXcan) and trait (using principal components to avoid multicollinearity between tissues). The extension of this approach, S-MultiXcan performs TWAS using summary-level GWAS results, in a similar manner to MetaXcan, but across all tissues simultaneously [94].

While these methods have enriched the original PrediXcan by ultimately providing higher power for TWAS, they are not without limitations. For example, UTMOST re-trains prediction models and requires full raw data of genotypes and gene expression measurements for every tissue and individual. While the prediction models derived from UTMOST are freely available for download, researchers are unable create tailored prediction models based selected tissue types unless they have full access to the data. In addition, there must be some overlap between samples for each tissue, and therefore disjoint resources cannot be integrated. MultiXcan integrates multiple imputed expression profiles to enhance power of discovery but does not provide aggregate or multi-tissue gene expression predictions. While the primary objective of imputing expression is to perform TWAS, there are also merits to generating imputations outside of this context. For example, imputed gene expression levels can be used as instrumental variables for Mendelian Randomization purposes.

In Chapter 2, we propose a novel method which integrates information from multiple datasets and tissues using a meta-analysis style approach. This method does not require the full set of raw data, but only measured expression and genotypes for a single tissue of interest. As such, information from multiple disjoint reference panels can be integrated. We demonstrate using GTEx tissues that our method improves prediction accuracy over PrediXcan and UTMOST, using the GEUVADIS consortium measured expression as external validation. We also demonstrate that combining other reference panels (such as GTEx + DGN whole blood) can further improve imputation accuracy, highlighting the importance of using multiple external datasets. Finally,

we demonstrate that our approach increases signal detection in TWAS compared to other approaches that impute gene expression (PrediXcan and UTMOST).

1.7.2 Revisiting array-based eQTL studies in whole genome sequencing era

Moving forward, it is clear that RNA-seq and scRNA-seq are the next generation technologies for gene expression studies. Although microarrays are an aging technology, they are still a viable option and still used for modern day expression studies [95,96]. For example, the NCBI GEO archive as of January 29, 2020 has roughly twice as many microarray datasets (roughly 24,000) compared to RNA-Seq (roughly 12,000) [97–99]. There are also studies in which transcriptomic profiles were already collected on microarrays and the RNA samples are no longer available for assaying with newer technologies.

Despite the relevance of microarrays even in this current era, there are many shortcomings for this technology. One very well-known limitation is the negative bias in probe hybridization when the genetic sequences of the individuals being studied differ from the microarray probe sequences [100,101]. Array-based technologies use target probe sequences based on the reference genome sequences, not necessarily accounting for genetic variations. For individuals carrying non-reference alleles in the probe sequences, the RNA strands are less likely to bind to the oligonucleotides, which in turn results artifactual cis-eQTLs that does not reflect true associations between genetic variants and expression levels. A common characteristics of these artifactual cis-eQTLs is that non-reference alleles are almost always associated with negative effect sizes because the non-reference alleles reduces the hybridization affinity [102].

There have been some solutions proposed to remove or mitigate this bias. For example, *Quigley* uses common variants from the 1000 Genomes reference panel to identify probes that overlap with a genetic variant, and removes said probes from the expression calculation [103]. One potential shortcoming to this approach is that if the 1000 Genomes markers do not match the study population, problematic probes may be missed while other probes may be unnecessarily removed. While this may work well for European samples, populations underrepresented by the 1000 Genomes reference panel may have inaccurate expression estimates. *Dannemann et al.* developed a statistical approach to determine probes with

reduced binding affinity by comparing the hybridization levels for probes of interest to probes from a control group [104]. However, this requires assays of a control group that are not affected by negative binding affinity.

The advent of next generation whole genome sequencing (WGS) technologies enables us to comprehensively catalogue all genetic variants. This affords us a new opportunity to comprehensively account for the negative hybridization effect across all genetic variations. As reported by *Quigley*, the negative hybridization bias in eQTL results appeared to be only partially resolved when using common variants only [103]. With the availability of WGS data for many study cohorts with transcriptomic profiles available through array-based technology, it is now possible to understand the full extent of the hybridization bias and to identify best practice to account for the bias in downstream analysis.

Chapter 3 of this dissertation comprehensively assesses the hybridization bias by leveraging WGS to identify the exact list of variant-overlapping probes. Here, the magnitude and effect of negative hybridization is characterized at both the probe and probeset (gene) level. We then evaluate existing approaches to identify the best practice for probe-level correction and compare these approaches in downstream eQTL analysis. Finally, we explore possible alternative bias-correction methods that leverage whole genome sequence data. We demonstrate that not all variant-overlapping probes have a negative hybridization bias, and that removing them might unnecessarily alter expression estimates by adding noise. This in turn could potentially mask true positives in eQTL studies. We derive and implement a probe-level imputation method, which instead of removing probes, we adjust their expression based on the values from other probes within the probeset. This approach appears to resolve the negative-hybridization bias while also preserving the overall correlation between genes.

1.7.3 Analysis of eQTLs on understudied tissues and populations capitalizing on discovery of novel eQTLs

With advances in high throughput technology, many resources are being generated for transcriptomic profiles of various tissues within the human body. For example, the GTEx consortium has now collected samples of 54 tissues across 948 donors. Some easily accessible

tissues, such as blood or cell lines, have been extensively studied with very large sample sizes. In another study, the eQTLgen consortium has meta-analyzed blood eQTLs of 31,684 samples across 37 studies across 11M common SNPs, identifying ~17,000 cis-eQTL genes [105]. However, while some tissues have been well-studied, not all have been examined equally. For example, kidney transcriptomics is still generally underrepresented and is still an emerging area of research [106]. The kidney is a vital human organ which is responsible for many tasks, including the control of body fluid volume, electrolyte balance, and removal of toxins through filtering processes. In addition, there are many kidney-related diseases that can arise that disrupt function, such as chronic kidney disease (CKD), nephrotic syndrome (NS), end stage renal disease (ESRD), and diabetic nephropathy (DN) [107]. In studying gene expression for the kidney, there are many challenges that include organ heterogeneity, low sample size, lack of healthy tissue samples, disease heterogeneity, as well as the underrepresentation of many populations. These challenges are highlighted below and are key points that Chapter 4 will address.

Although there have been some kidney transcriptomes profiled, these studies have had limitations, such as low sample size, or expression being assayed using bulk tissue only. For example, in the GTEx consortium, the Kidney – Cortex tissue (bulk tissue) has only 73 RNA sequenced samples, while the Kidney – Medulla has only 4 RNA sequenced samples. (This is in part because most of the healthy kidneys of GTEx participants are donated and unavailable for expression studies. The remaining samples available for RNA-seq tend to be lower quality samples.) This is low compared to many other tissues in this cohort, (skeletal muscle, whole blood, subcutaneous adipose, thyroid, lung all have over 500 individuals) and the total number of kidney eGenes as well as ratio of detection (eGenes divided by total number of expressed genes) are near the bottom for all tissues. Another study published in 2017 identified 1,886 candidate eGenes in 96 individuals with chronic kidney disease (CKD). However, the RNA sequencing was performed on bulk tissue [108], which would not differentiate between the heterogeneous compartments of the kidney. Clinical studies of kidney function are often interested in the filtration ability of the glomerulus (GFR) [109]. As such, there has been evidence that microdissection of the kidney into glomerular and tubulointerstitial

compartments would allow for increased specificity of kidney transcriptomic studies compared to studying the cortex [110,111]. Recently, the nephrotic syndrome study network (NEPTUNE) cohort performed such study, describing glomerular- and tubular- specific transcriptomic profiles for 187 individuals with nephrotic syndrome (NS) [106]. While this study provided a detailed compartment-specific eQTL landscape of individuals with NS, it is a rare disease characterized by damaged kidneys, and often progresses serious diseases such as CKD and end-stage renal disease [112]. While this is an invaluable transcriptomic resource, there is also merit in studying relatively healthier kidneys, as it could provide more insight into common disease-related traits.

In addition to tissue heterogeneity, many renal diseases are also heterogeneous. If association studies are performed using heterogeneous traits, latent substructures among individuals could generate spurious results or mask signals in the analyses. To characterize various kidney diseases, clinical measurements such as glomerular filtration rate (GFR) or albumin-creatinine ratio (ACR) can be used. However, there are other fine-resolution phenotypes, such as morphometric measurements that can be obtained from renal biopsies that snapshot the physical state of the kidney at a given moment. Using these phenotypes could potentially be very important pieces of information to connect the dots between genetic variants and complex renal traits.

Finally, there is also importance to studying population-specific expression data. Since most GWAS and transcriptomic studies are heavily biased towards individuals of European descent, this can lead to reduced accuracy in predicted gene expression for non-European individuals [113]. Performing genetic studies of diverse populations allow for trans-ethnic fine mapping, which can pinpoint potential causal variants that may be masked due to LD structures of a single population otherwise [114]. In addition, studying isolated populations have revealed novel population-specific variants that may be too rare for detection in other populations. For example, studying the isolated Sardinian population has detected novel GWAS loci for hemoglobin levels, and lipid and blood inflammatory markers [115,116]. Studies on the Finnish population for 64 quantitative traits have also identified 19 unique or enriched (20-fold more common compared to non-Finnish Europeans) genetic loci [117].

In Chapter 4, we provide a transcriptomic landscape of the kidney on a cohort which addresses many of the challenges in tissue-specific systems genetic studies. In this work, we integrate multiple datasets to provide a genomic and transcriptomic profile for microdissected glomerular and tubulointerstitial tissues from 97 Pima Native American individuals from a diabetic nephropathy. This cohort gives the opportunity to study many high-resolution phenotypes from a population-specific group of individuals, with relatively healthy tissue samples. From this dataset, we use gene expression data assayed from both microarray and RNA-seq platforms, kidney morphometric traits, and whole genome sequencing to conduct various analyses. Although the sample size is relatively low, the abundance of high-quality data provides many insights into the regulatory aspects of the kidney for this population. Here, we discover both tissue-specific and population-specific regulatory variants which highlights the importance of studying diverse populations and accounting for tissue-specificity.

Chapter 2 Meta-imputation of Transcriptome from Genotypes Across Multiple Datasets Using Summary Statistics

2.1 Abstract

Transcriptome wide association studies (TWAS) can be used as a powerful method to identify and interpret the underlying biological mechanisms behind GWAS by mapping gene expression levels with phenotypes [53,54]. In TWAS, gene expression is often imputed from individual-level genotypes of regulatory variants identified from external resources, such as Genotype-Tissue Expression (GTEx) Project [55,56]. In this setting, a straightforward approach to impute expression levels of a specific tissue is to use the model trained from the same tissue type. When multiple tissues are available for the same subjects, it has been demonstrated that training imputation models from multiple tissues types improves the accuracy because of shared eQTLs between the tissues and increase in effective sample size. However, existing methods require access of genotype and expression data across all tissues for joint training [93]. Moreover, they cannot leverage the abundance of various expression datasets across various tissues for non-overlapping individuals.

Here, we explore the optimal way to combine imputed levels across training models from multiple tissues and datasets in a flexible manner using summary-level data. Our proposed method (SWAM) combines arbitrary number of transcriptome imputation models to linearly optimize the prediction accuracy given a target tissue. By integrating models across tissues and/or individuals, SWAM can improve the accuracy of transcriptome imputation or to improve power to TWAS without having to access each individual-level dataset. To evaluate the accuracy of SWAM, we combined nearly 48 tissue-specific gene expression imputation models from the GTEx Project as well as imputation model trained from a large eQTL study of Depression

Susceptibility Genes and Networks (DGN) Project [50] to tested imputation accuracy in GEUVADIS lymphoblast cell lines (LCL) samples [47]. We also extend our meta-prediction method to meta-TWAS to leverage multiple tissues in TWAS analysis with summary-level statistics. Our results capitalize on the importance of integrating multiple tissues to unravel regulatory impacts of genetic variants on complex traits.

2.2 Introduction

Genome wide association studies (GWAS) have been able to identify numerous associations between genetic variants and complex traits. However, interpreting the biological mechanisms underlying the association signals remains a challenge [118]. Recently, studies involving gene expression have become increasingly popular as a means to provide biologically meaningful insight into statistical associations [55,56]. Transcriptome-wide association studies (TWAS) is a widely used method to translate GWAS association signals into more interpretable units by examining the association between phenotypes and gene expression levels imputed from genotypes. Associations identified from TWAS can be interpreted as potentially causal relationships between the traits and the genes through gene regulation [54,119,120]. While TWAS may not detect associations driven by functional mechanisms irrelevant to gene regulation, it increases the specificity and interpretability in identifying GWAS signals driven by gene regulation. Imputed gene expression can be utilized in various contexts of association analysis beyond TWAS, such as Mendelian randomization [121,122] or estimation of trait heritability attributable to cis-eQTLs [123]. Since genotype data from DNA is far easier and cheaper to obtain than expression data from tissues, TWAS based on imputed expression offers excellent augmentation to study the genetic component of gene regulation in addition to RNA-seq-based studies.

The first-generation methods to impute gene expression levels from genotypes train the model from a single-tissue dataset comprising of many individuals with both genotypes and expression profiles [55,56]. For example, a widely-used method PrediXcan [55] uses Elastic net regularization to identify cis-eQTLs (expression quantitative loci) to train the model to impute

gene expressions from genotypes. Other methods, such as TWAS [56], employ different regularization but typically produces a linear model to impute gene expressions as a weighted sum of cis-eQTL genotypes. Imputation models are trained using these methods from various population-scale transcriptomic datasets, such as the Genotype-Tissue Expression (GTEx) project [52,123], Depression Genes and Network (DGN) study [50], and The Cancer Genome Atlas (TCGA) [124], and these models are made available in public repository such as predictDB (http://predictdb.org/) or FUSION (http://gusevlab.org/projects/fusion/) so that expression imputation or TWAS can be performed from any genotyped individuals.

Although these first-generation methods for transcriptome imputation have been quite useful, they have limited accuracy mostly due to limited sample size in the training datasets where both genome-wide genotypes and transcriptome-wide expression levels are available. While millions of individuals have been genotyped or sequenced to date [4,125–127], the sample-size of current population-scale transcriptome data are typically limited only to hundreds or thousands [128] (with the largest study cohort having around 30k participants [129]), primarily due to the difficulty in collecting high quality tissues (other than whole blood) from living donors. Moreover, transcriptomic datasets are prone to potential batch effects between studies [68,130–132], making it difficult to integrate across multiple datasets to build a large and harmonized resource to be trained from. Furthermore, there are hundreds or thousands of different types of tissues or cells, requiring orders of magnitude larger effort to comprehensively profiles transcriptomes in population-scale across tissues, as in GTEx Project.

Recently, methods to address the shortcomings of the first-generation methods have been developed. When transcriptomic profiles are available across many tissues, such as in GTEx Project, transcriptome imputation can improve by leveraging the shared genetic components across tissues. Even though each tissue represents a unique transcriptomic profile, a large fraction of eQTLs are shared across tissues [133], and the availability of multiple expression measurements across tissues can help more precisely identify the shared eQTLs, which in turn can improve the imputation accuracy. For example, UTMOST trains a transcriptome imputation model a simultaneously across all tissues using a combination of L1 and L2 penalization across markers and tissues, respectively [93]. Another multi-tissue approach, MultiXcan, does not

impute transcriptomes, but performs a multi-tissue TWAS across all tissues by including each tissue-specific imputed expression as a predictor variable to improve power to identify association between a trait and a gene, in which the underlying mechanism potentially involves multiple tissues or cell types [94].

Even though UTMOST substantially improves the accuracy of transcriptome imputation, it assumes that expression measurements across multiple tissues are available for overlapping set of genotypes individuals for training imputation models. While this assumption can be met when training from the GTEx dataset (assuming granted access to the individual-level data), it may not be realistic in other circumstances where expression measurements are available for non-overlapping individuals (such as in TCGA), or it is infeasible to obtain individual-level genotypes and expression data due to limited access privilege. As population-scale transcriptomic resources are rapidly increasing, it should be possible in principle to integrate these resources to better impute transcriptomes. While there have been additional methods which have been developed to increase the accuracy of gene expression or TWAS [94,134–136], none of them – to the best of our knowledge – are able to perform "meta-imputation", which systematically integrates multiple imputation models without the need to access to individual-level data.

Here we propose Smartly Weighted Averaging across Multiple tissues (SWAM), a multi-tissue transcriptome imputation method based on a flexible meta-analysis across multiple imputation models. Unlike UTMOST, SWAM does not require access to all genotypes and expression datasets for training its imputation model. Instead, it takes individual transcriptome imputation models trained from individual tissues while optimizing the expected imputation accuracy for a target tissue. Moreover, it can seamlessly integrate imputation models trained from multiple datasets comprising of different individuals and tissues. As a result, SWAM can integrate across hundreds of imputation models across GTEx, DGN, and TCGA projects without requiring all individual-level data to substantially improve the imputation accuracy over existing methods, as we demonstrate with GEUVADIS data. Moreover, we demonstrate that SWAM improves the power of TWAS over single-tissue methods and many alternative multi-tissue methods.

2.3 Results

2.3.1 Smartly Weighted Averaging across Multiple Tissues (SWAM)

We propose *Smartly Weighted Averaging across Multiple tissues* (SWAM), a method that provides a flexible framework to impute tissue-specific expression by integrating single-tissue imputation models across many tissues and datasets (Figure 2.1). The key principle behind SWAM is to improve the accuracy of transcriptomic imputation by determining the optimal linear combination of multiple imputation models in terms of expected imputation accuracy. To do this, SWAM requires a reference tissue (tissue of interest) to be defined as a basis to determine the relative contributions from multiple imputation models. Using the individual-level genotypes and expression of only the reference tissue, SWAM integrates imputation models trained from different tissues and datasets (e.g. GTEx, DGN, and TCGA) without requiring individual-level data except for the reference tissue.

The first step of SWAM is to apply each transcriptomic imputation model to the reference genotypes, which results in individual-level, tissue-specific imputed expression. The second step of SWAM compares each imputed expression with the measured expression of the reference tissue to calculate optimal weights by linearly combining multiple prediction models to maximize expected mean squared error (MSE) (see Methods for the details). The output of second step is an integrated transcriptomic imputation model compatible with the PrediXcan and MetaXcan software tools. Using this SWAM model, we can impute the transcriptome of any samples of interest with genotype information available (via PrediXcan), or to use the model and covariance matrix directly to perform TWAS (via MetaXcan) when GWAS summary statistics are available (Supplementary Figure 2.1).

2.3.2 Simulation study demonstrates the robustness of SWAM across various scenarios

We performed simulation studies to evaluate SWAM's ability to robustly impute expression by leveraging tissue-specific and cross-tissue components across a wide spectrum of parameter settings. To do this, we independently simulated multi-tissue expression levels along with genotype data for both our training and validation sets (see section 2.5, Materials and Methods

for details). We compared SWAM with two heuristic approaches – *naïve average,* which equally weights individual tissue and *best tissue,* which only uses the tissue with the highest expected imputation accuracy – as well as with *single-tissue* imputation.

As expected, we observed *naïve average* to be particularly powerful when the causal variants are shared across all relevant tissues (Figure 2.2A), identifying 94% of genes as significantly imputable at FDR < 0.05. When all causal variants were tissue-specific, the naïve average only identified 25% of genes to be imputable. On the other hand, best-tissue was more powerful (38%) than naïve-average when the all causal variants were tissue-specific, but worse when all causal variants are shared. When only *single-tissue* was used for imputation, the performance stayed similar regardless of the tissue-specificity. Encouragingly, SWAM outperformed all three methods across all ranges of tissue-specific and cross-tissue heritability settings. We believe this is because SWAM learns tissue-specific weights without pre-conceptions of tissue relatedness, and thus determines the weights for relevant tissues while ignoring unrelated ones.

A similar trend is observed when we vary the number of relevant tissues that shares cross-tissue heritability (Figure 2.2B). In the case where there are no relevant tissues other than the target tissue, naïve average is least powerful while SWAM performs as well as the *single tissue* approach. This suggests that in this scenario, SWAM is correctly giving non-zero weights to only the target tissue, making it similar to the *single-tissue* method. In the other scenario where every tissue is relevant, SWAM provides a similar power to the *naïve average* approach, suggesting that SWAM is robustly assigning weights to each relevant tissue. Similarly, when there are more tissues available in overall (assuming 50% are relevant tissue sharing cross-tissue heritability), the power of SWAM and *naïve average* keep increasing while *single-tissue* and *best-tissue* remain similar (Figure 2.2C).

Our simulation study also evaluated the impact of sample size of the reference tissue. We hypothesized that *single-tissue* would perform poorly when the sample size of the reference tissue was small, which was indeed observed in our results (Figure 2.2D). When the reference tissue has sample sizes of 50, 100, 200, we observed that *single tissue* method identified 36%, 66%, and 92% of imputable genes. Because additional tissues are helpful especially when the

reference tissue has smaller sample size, the *best tissue* approach performed better at lower sample size (59% at n=50), but worse at higher sample size (88% at n=200). Similarly, *naïve average* also performed better at lower sample size (63% at n=50), but worse at higher sample size (78% at n=200). However, SWAM consistently outperformed single tissue across all cases (59%, 86%, 97% at n=50, 100, 200). This implies that borrowing information from a relevant tissue (to the reference) is useful in these situations and SWAM robustly estimates the weights from each tissue accounting for the uncertainty from different sample sizes.

2.3.3 SWAM outperforms other transcriptome imputation methods in evaluations with real data by considering the bias-variance tradeoff

We applied SWAM to create multi-tissue imputation models from GTEx v6, using lymphoblastoid cell lines (LCL; the official tissue name in GTEx was "Cells – EBV-transformed lymphocytes") as the reference tissue, to evaluate its imputation accuracy of LCL transcriptomes of 344 European samples from the GEUVADIS consortium [47]. We compared the accuracy of SWAM with various methods, including *single tissue* imputation models (generated by PrediXcan), *naïve average*, *best tissue*, and another multi-tissue method *UTMOST*.

Among the single-tissue imputation models, we observed that the imputation from LCL identified 1,552 genes as significantly imputable at FDR < 0.05. Interestingly, we observed that another tissue, fibroblast cell lines (FCL; the official tissue name in GTEx was "Cells – Cultured fibroblasts"), identified even more genes (1,690 genes) as significantly imputable for GEUVADIS LCL expression levels. One of the outstanding differences between LCL (n=114) and FCL (n=272) models were the sample size used for training. We suspect that this is due to (1) the difference in sample size (i.e., FCL imputation has less variance) and (2) the similarity of transcriptomic profiles between LCL and FCL (i.e., FCL model tends not to introduce large bias). However, tissues with larger sample size did not always result in more accurate imputation. When we examined the results from Skeletal muscle model (n=361), which had the largest sample size in GTEx v6, we identified only 1,197 genes as significantly imputable. This is likely because the large differences of transcriptomic profiles between LCL and Skeletal muscle (i.e., Skeletal

muscle model tends to introduce large bias). These examples demonstrate that both sample size and tissue relevancy are important for maximizing the imputation accuracy. In statistical terms, our primary interest was to reduce the mean-squared error (MSE), which is the sum of $Bias^2$ and Variance. We suspect that FCL model performed better than LCL models due to much smaller variance (because of larger sample size), and better than Skeletal muscle models due to much smaller bias (Supplementary Figure 2.2). We hypothesized that by combining imputations from multiple models, we can minimize MSE by substantially reducing variance without introducing excessive bias, which was our main motivation for developing SWAM.

When evaluating the multi-tissue methods, our two heuristic approaches, *best-tissue* and *naïve average* identified 2,493 and 2,666 significantly imputable genes, respectively, which was >47% and >57% higher than any single tissue models. *UTMOST* (using LCL as the reference) also substantially increased the number of imputable genes (2,238 genes, >32% increase over any single tissue), but surprisingly, it had fewer than the imputable genes compared to the two heuristic approaches. Finally, when we applied *SWAM* specifying GTEx LCL as the reference tissue, the number of imputable genes further increased to 3,040, which is >79% larger than any other single tissue models (Supplementary Table 2.1, Figure 2.3A). Interestingly, *SWAM* improved the imputation accuracy over *UTMOST* even though it requires individual-level data only for one tissue (i.e., LCL) in GTEx while *UTMOST* requires simultaneous access to individual-level data across all tissues. These results demonstrate that SWAM offers an accurate and flexible meta-imputation framework by optimally combining multiple imputation models across tissues.

2.3.4 SWAM enables meta-imputation of expression levels across multiple heterogeneous datasets

One of the important advantages of SWAM compared to other multi-tissue imputation methods is the ability to integrate imputation models across heterogeneous datasets where samples may not necessarily overlap. To evaluate the benefit of SWAM's ability for multi-dataset "meta-imputation", we integrated imputation models trained from GTEx v7 and v8, as well as 922 whole blood transcriptomes from Depression Gene Network (DGN). The rationale to

include GTEx v7 and v8 models is that the datasets are slightly different from v6 (for example, v7 has more samples in all tissues except for LCL, FCL, and whole blood) and integrating multiple training models from slightly different versions of datasets may improve the accuracy. The reason to include DGN whole blood is that the sample size is much larger than any individual tissue GTEx, so it may help further reduce the variance and MSE of the imputation model.

When applying SWAM to GTEx v6, v7, or v8 datasets individually, the number of significantly imputed genes at FDR < .05 were 3,040, 3,060, and 3,203, respectively (Figure 2.3B). However, when all datasets were combined, the number of imputable genes increased to 3,342. These results suggest that imputation across multiple datasets can help even when the datasets are highly overlapping. When we additionally integrated SWAM with the DGN whole blood model, which detected 2,390 imputable genes by itself, the number of imputable genes by the integrated SWAM model further increased to 3,413. Note that we needed individual-level data only for the reference tissue/data (GTEx v6 LCL in our experiment), so an arbitrary combination of imputation models, which consist of only summary-level data, can be seamlessly added to the meta-imputation framework of SWAM.

Overall, using all 49 GTEx v8 tissues in combination with the DGN whole blood model provided the highest number of predictable genes, with a 112.9% improvement over the corresponding GTEx v8 PrediXcan-LCL model (single tissue), and a 13.5% improvement over the GTEx v6 version of SWAM-LCL (multi-tissue) (Figure 2.3B). Regardless of the version of GTEx used, including the DGN whole blood model gives a substantial improvement in number of predictable genes compared to not including it in the model. Another interesting observation is that while PrediXcan-LCL (v6) appears to perform better than PrediXcan-LCL (v7), SWAM-LCL derived from v7 performs better than v6 SWAM-LCL. This may suggest that while GTEx v7 PrediXcan-LCL may not have had a significant improvement in eQTL detection compared to its predecessor, other tissues may have improved in more substantial ways. This is because the sample size for LCL in v7 decreased by 18 samples, whereas other non-blood tissues had substantial sample size gains of up to 89 individuals. Here, SWAM leverages the increase in

quality from other tissues, which allows for better overall prediction regardless of the quality of the target tissue itself.

2.3.5 SWAM robustly captures both tissue-specific and cross-tissue regulatory components

The key component behind the robust performance of SWAM is that it learns how to distribute weights across multiple imputation models for each gene individually. If a gene shares eQTLs across many tissues, the SWAM's weights will be distributed evenly across tissues and the model will behave similarly to the naïve average heuristic. For example, *ERAP2* is a well-known gene with shared eQTLs profiles across most tissues. In the GTEx (v6), *ERAP2* can be reliably imputed with any of the 44 single-tissue imputation models from PrediXcan with $r^2 > 0.77$ or more eQTLs. As a result, the weights from SWAM is almost evenly distributed across the tissues, ranging from 0.018 to 0.027 (Supplementary Figure 2.2A), and the accuracy of SWAM ($r^2 = 0.795$) is very similar to the accuracy of naïve average ($r^2 = 0.796$).

On the other hand, when the imputation model from the reference tissue is not particularly good due to smaller sample size or other technical issues, SWAM can substantially improve accuracy by leveraging eQTL sharing from other tissues. For example, the single-tissue imputation accuracy of *GSTM1* is relatively low in LCL tissue ($r^2 = 0.368$) compared to the accuracy of the 38 other tissues in which a PrediXcan imputation model is available (average $r^2 = 0.61$). Using SWAM, the predictive R-squared increases to $r^2 = 0.741$ by assigning positive weights to 31 tissues (Supplementary Figure 2.2B).

Finally, for genes that are highly tissue-specific, the SWAM's weights will be distributed similarly to the best tissue heuristic. For example, CTSK is expressed in most tissues, but has eQTLs in only 16 tissues, (Supplementary Figure 2.2C). SWAM assigns weights to 7 of these tissues, and substantially improves the predictive accuracy from $r^2 = 0.111$ to $r^2 = 0.447$.

2.3.6 Comparison of imputation models in the context of TWAS

We conducted TWAS analysis using SWAM, UTMOST, and PrediXcan models via MetaXcan [137]. In addition, we also used S-MultiXcan [94] to simultaneously test all of the PrediXcan models using their PCA regression approach. We used a Bonferroni correction to establish p-

value threshold for each analysis separately, based on the number of genes imputed. Overall, we found that among the methods that directly estimate expression levels (SWAM, UTMOST, PrediXcan), SWAM outperformed the other methods in terms of number of associations detected (see Supplementary Tables 2.4, 2.5, 2.6). For example, PrediXcan models on average detected 23.7, 23.2 and 4.0 transcriptome-trait associations for HDL, LDL and T2D respectively. For SWAM, we observed an average of 79.7, 77.8 and 8.4 associations per tissue, whereas UTMOST yielded an average of 69.3, 61.6 and 8.8 associations per tissue, for the three traits respectively.

We plotted transcriptome-wide signals for the LDL trait using the GTEx v6 liver model for PrediXcan, UTMOST and SWAM (Figure 2.4). One interesting signal gained from the SWAM analysis is the APOC1 gene, which is primarily expressed in the liver and has been implicated in playing a role in HDL and LDL/VLDL (very low-density lipid) metabolism [138].

One potential shortcoming for both multi-tissue approaches (SWAM and UTMOST) appear to be that the number of unique signals (across all tissues) is fewer than those generated by PrediXcan's single tissue models. For example, SWAM produces 210 unique associations for the HDL trait, while we see 187 unique associations from UTMOST and 248 unique associations from PrediXcan. Similarly, MultiXcan detects 284 significant associations when scanning across all tissues (based off the PrediXcan models). It appears that while the multi-tissue methods can leverage information from other tissues to predict expression accurately, marginal association signals in TWAS are potentially lost using these approaches. However, we found that a high number of these unique signals from the PrediXcan TWAS appeared only in one or two tissues (92.5% for HDL, 98.2% for LDL and 100% for T2D).

With all these various considerations, SWAM appears to improve TWAS power for a given tissue, although ultimately may yield fewer signals compared to comprehensive tissue scans using PrediXcan or MultiXcan. While SWAM outperforms other methods in terms of prediction accuracy, there may not be a clear-cut winner in terms of performance in TWAS. The best approach to use will likely depend on the needs of the researcher, and each approach may

provide different yet complementary insights into understanding the biological mechanisms from these association studies.

2.4 Discussion

The transcriptome serves as an intermediate phenotype linking genetic variants to complex traits. Association studies between traits and gene expression, when used in conjunction with GWAS, provide additional insight into the biological mechanisms of complex traits. Prediction of gene expression in the context of transcriptome wide association studies is a promising approach to understanding the connection between our genes and many traits. Yet, there are still many challenges that arise when performing association studies with predicted expression. Current tissue-specific prediction models are trained using data obtained from their respective tissues, which can vary greatly in data quality and sample size. As such, there is a great deal of variability among tissues in the prediction accuracy of tissue-specific gene expression levels. For example, PrediXcan was able to significantly predict only 2086 vagina-specific genes, while it discovered 8171 genes specific to the tibial nerve tissue. Furthermore, the prediction accuracy of significant genes within a tissue are also highly variable, with some genes such as ERAP2 having very high (>80% of variation explained by eQTLs) predictability and other genes (~1% of variation explained by eQTLs) with low predictability.

In this paper we developed SWAM, a method that determines the level of eQTL sharing between tissues and uses the shared information from other tissues to improve the prediction accuracy for the target tissue. By simultaneously examining the relatedness of multiple tissues, SWAM in essence increases the effective sample size of prediction models. Using GEUVADIS LCL data, we compared SWAM to single-tissue approaches. We found that our multi-tissue approach, in addition to increasing the number of significantly predictable genes for each tissue, also improved the overall prediction accuracy for genes that were already significantly predictable using PrediXcan. We improved the power of TWAS by running a SWAM-adapted version of MetaXcan for various traits, finding an increased number of significant

transcriptome-trait associations, even when correcting for the larger number of genes predicted.

Although SWAM provides a substantial improvement for the number of significantly predictable genes for many tissues and generally increases power for TWAS, there are some shortcomings and caveats to consider with the approach. It is important to note that unlike PrediXcan, SWAM does not actually perform model training or eQTL discovery. Instead, it evaluates the efficacy of various single-tissue prediction models (in this case, the GTEx tissues) and assigns weights to the models based on their relatedness to the target tissue. Therefore, for SWAM to work, there must already be a database of prediction models that it can use to derive the multi-tissue weighting. Because we are utilizing existing prediction models, we acknowledge that there will be cases where the SWAM prediction accuracy could be similar or worse to the single-tissue prediction, especially if the gene has shared eQTLs across many tissues or if the single-tissue prediction model was already performing well. The improvement observed in our validations and TWAS are an overall trend, and as with any analysis, interpretation of any specific results should be approached with caution. Furthermore, the improvement for any given gene has an upper limit which is dependent on the pool of single tissue models available. There may be tissues that have very few relevant other tissues to draw information from. For any given gene within the target tissue, SWAM automatically assigns weights of non-relevant tissues to zero based on a threshold. However, for the purposes of our study, the threshold was tuned to be more lenient, allowing for more tissues to be included in the prediction of each gene's expression levels. A more lenient threshold will yield more genes, but a lower sensitivity to the target tissue. A stricter threshold will provide predictions that are more specific to the target tissue but will provide predictions for fewer genes and may reduce prediction accuracy in some genes. Optimal tuning of this threshold may depend on the target tissue, and the goals of the analysis. Further work could help determine the ideal way to tune these thresholds, perhaps using a different threshold depending on the gene and tissue in question.

Next, our empirical validation of prediction accuracy was tested on European individuals (344 samples from GEUVADIS) and thus SWAM's performance with other populations has not yet been determined. A future direction of research could be to examine whether a single model

derived from mixed populations would represent each of the populations accurately, or if a different model should be trained on each population separately. Currently, evidence suggests that training from the correct ancestry group is the ideal approach for population-specific prediction [139], which emphasizes the importance of reference panel resources derived from a wide array of ancestries. Alternative approaches could be to leverage trans-ancestry correlation, which has been shown to increase predictive R^2 in the context of polygenic risk scores [140].

Finally, while SWAM improved the number of association signals for any given tissue in TWAS compared to UTMOST and single tissue PrediXcan, aggregation of signals (MultiXcan/combining PrediXcan signals) suggest that other approaches may yield more unique signals. It is unclear which approach is preferable in this scenario, and the answer may depend on unraveling the causality of association signals. Recently, there have been a number of publications which have addressed this issue, such as PTWAS which uses instrumental variables (IVs) to investigate the causal relationship between expression levels and complex traits [134], or phenomeXcan, which integrates GWAS and gene expression and regulation data to identify likely causal pathways [141]. Future directions could include using IVs or functional annotation to interpret TWAS signals.

To conclude, we propose a novel method for gene expression prediction, which extends already established single-tissue prediction models into a multi-tissue setting. By combining information from multiple models, we were able to increase overall tissue-specific prediction accuracy for many genes and increase power for transcriptome-wide association studies.

2.5 Materials and Methods

2.5.1 SWAM Notation and Framework

Our framework for *SWAM* is designed to find the optimal linear combination of imputed expression levels from multiple tissues and datasets. For simplicity, we will denote each (tissue, dataset) combination as a source. We assume there are K imputation models from individual

sources, with each model indexed as $j \in (1,..,K)$. We also denote $r \in \{1,\dots,K\}$ to represent the index of the reference source. The inputs for *SWAM* are: (1) $f_j(\cdot)$ – the single-source imputation models and (2) $\boldsymbol{Y_r}$ and $\boldsymbol{X_r}$ – the individual-level gene expression measurements and genotypes for the reference source. For each gene g, let $\hat{\boldsymbol{s}}_j^g = f_j(\boldsymbol{X}|g)$ be imputed expression from a single source. Then we can represent any linearly combined multi-tissue imputed expression $\widehat{\boldsymbol{m}}_r^g$ as

$$\widehat{\boldsymbol{m}}_r^g = \sum_{j=1}^{K} w_j^g \hat{\boldsymbol{s}}_j^g$$

where w_j^g is the weight contributed by j-th source. *SWAM* learns w_j^g by leveraging individual-level data from the reference source as we describe later.

2.5.2 Multi-tissue methods using naïve average or best-tissue

There are two heuristic approaches to impute expressions from multiple sources - *naïve average* and *best tissue*. *Naïve average* defines weights uniformly as $w_1^g = \dots = w_K^g = {}^1\!/_K$. For *best tissue,* the weights are defined as a dichotomous variable:

$$w_j^g = \begin{cases} 1 & \text{if } j = \underset{i}{\operatorname{argmax}}(cor(\hat{\boldsymbol{s}}_i^g, \boldsymbol{y}_r^g)) \\ 0 & \text{otherwise} \end{cases}$$

where $\boldsymbol{y}_r^g$ represents the individual-level expression measurements of the reference source.

2.5.3 Smartly Weighted Average across Multiple Tissues (SWAM)

Here we describe how SWAM calculates optimal w_j^g, whose derivation is shown in the Supplementary Text. It is important to note that *SWAM* works ideally when the tissue type intended to be imputed matches to the tissue types of the reference source. We define $\boldsymbol{y}_r^g$ as the $n \times 1$ vector of individual-level expression measurements for the reference source, and as before, $\boldsymbol{X_r}$ to be the corresponding $n \times m$ matrix of individual-level genotypes. The first step is to impute expression using each of the K models using the reference genotypes. Thus, we obtain K sets of imputed expressions, $\hat{\boldsymbol{s}}_j^g = f_j(X_r|g)$, with each being a single-source

prediction for the samples in the reference data. The weights for *SWAM* are given by

$$\boldsymbol{w^g} = \left(w_1^g, w_2^g, \ldots, w_K^g\right)^T = \left[\begin{bmatrix} cor(\hat{\boldsymbol{s}}_1^g, \hat{\boldsymbol{s}}_1^g) & \cdots & cor(\hat{\boldsymbol{s}}_K^g, \hat{\boldsymbol{s}}_1^g) \\ \vdots & \ddots & \vdots \\ cor(\hat{\boldsymbol{s}}_K^g, \hat{\boldsymbol{s}}_1^g) & \cdots & cor(\hat{\boldsymbol{s}}_K^g, \hat{\boldsymbol{s}}_K^g) \end{bmatrix} + \lambda I\right]^{-1} \begin{bmatrix} cor(\hat{\boldsymbol{s}}_1^g, \boldsymbol{y}_r^g) \\ \vdots \\ cor(\hat{\boldsymbol{s}}_K^g, \boldsymbol{y}_r^g) \end{bmatrix}$$

Here, the correlation matrix account for the similarity between the imputation models, and the vector containing the entries $cor\left(\hat{\boldsymbol{s}}_j^g, \boldsymbol{y}_r^g\right)$ account for the empirical similarity of imputed expressions from each model to the measured expressions in the reference source. When $j = t$, because $cor\left(\hat{\boldsymbol{s}}_t^g, \boldsymbol{y}_r^g\right)$ will be prone to overfitting, we replace this value to a 5-fold cross-validated correlation instead, which is available from PrediXcan output. Finally, λI acts to regularize the weights, providing numerical stability for the inversion of the covariance matrix. The calibration of λ is further discussed in the Supplementary Text.

2.5.4 Simulations

Our simulation study sought to examine SWAM's ability to detect the correct shared components between related tissues across a wide spectrum of parameter settings. We compared *SWAM* with *naïve average, best tissue* and *single tissue* approaches. For each simulation, we independently generate individual-level genotypes and expression multiple tissues. For the reference set, we simulated X_r, an $n_r \times m$ genotype where n_r is the number of individuals and m the number of SNPs. In our simple simulation, we assume that each SNP is independent, with non-reference allele frequency (AF) distributed with *Beta(1,3)*. The genotypes were simulated using a binomial distribution based off the AF. To simulate multi-tissue expressions, for each tissue $j \in (1, .., K)$ we specific effect sizes $\boldsymbol{\beta_j}$, to simulate expressions $\boldsymbol{y_j} = X_r\boldsymbol{\beta_j} + \varepsilon_j$. For reference tissue (i.e. $j = r$), we assume two causal SNPs with nonzero elements in $\boldsymbol{\beta_j}$, where one SNP is expected to explain tissue-specific heritability (h_r^2) for the reference tissue and the other SNP explains the cross-tissue heritability (h_c^2), summing up to total heritability ($h^2 = h_r^2 + h_c^2$). Other tissues (i.e. $j \neq r$) were divided into "related

tissues" and "independent tissues". For related issues, $\boldsymbol{\beta_j}$ had only one non-zero values corresponding to cross-tissue heritability (h_c^2). For independent tissues, all $\boldsymbol{\beta_j}$ had zero values. Finally, we generated another set of validation genotypes matrix X_v with size $n_v \times m$, and the validation expressions ($\boldsymbol{y_v} = X_v\boldsymbol{\beta_r} + \varepsilon_v$) of reference tissue using the same settings to use for evaluation.

We then trained tissue-specific imputation models $f_j(.), j \in (1, \dots, K)$ by applying an elastic-net model (using *glmnet* R package [142]) for each pair of X_r and $\boldsymbol{y_j}$**.** The tuning parameters for elastic net were determined via a five-fold cross-validation technique. Using $\boldsymbol{y_r}, X_r$ and $f_i(.)$, we obtained *naïve average, best tissue* and SWAM models as detailed in the framework and weights section. To calculate the proportion of imputable genes, we performed linear regression between $\boldsymbol{y_v}$ and the imputed expression from genotypes X_v using the different methods to obtain a p-value.

Each simulation was repeated for 1,000 times in each setting. We varied parameters to evaluate their impact on the performance of each method. We varied $h^2 \in \{0, 0.1, \cdots, 1\}$ (default 0.1), $h_c^2/h^2 \in \{0, 0.1, \cdots, 1\}$ (default 0.5), $K \in \{2, 4, 6, 8, 10, 20, 30, 40, 50\}$ (default 10), fraction of independent tissues ranging $\{0, 0.1, \cdots, 0.8\}$ (default 0.5), $n_r \in \{50, 100, \cdots, 500\}$ (default 200), and the p-value threshold ranging $\{10^{-6}, 10^{-5}, \cdots, 0.01, 0.05, 0.1\}$ (default 0.05). Throughout all simulations, $m = 35, n_v = 200$ were used.

2.5.5 Input Datasets: Genotypes, Expressions, and Imputation Models

In our experiments with real datasets, we leveraged multiple published datasets where genotypes, expressions, and imputation models are available to evaluate the performance of SWAM and other methods in various settings. Specifically, we used the GEUVADIS LCL [47] genotypes and expressions as a validation dataset. We used GTEx data [14] [24] and PredictDB [55] to build multi-tissue imputation models. To demonstrate the ability to SWAM to incorporate multiple datasets, we used DGN [50] dataset as well as multiple versions of GTEx datasets.

2.5.5.1 Multi-tissue transcriptomic profiles and imputation models from the GTEx project

To build multi-tissue imputation models using *SWAM, UTMOST, naïve average, and best tissue methods*, we used single-tissue imputation models, individual-level genotypes, and expressions obtained from the GTEx consortium. Single-tissue imputation models were downloaded from the PredictDB (http://predictdb.org/) repository for GTEx versions 6, 7 and 8 (44, 48 and 49 tissues respectively) [3] [14] [24], which were trained using PrediXcan's elastic net methods. Individual-level genotypes and expression levels were only used for the reference tissue (e.g. EBV-transformed lymphocytes) which is deemed to be the closest to the validation data (e.g. GEUVADIS LCL), using GTEx version 6.

When evaluating multi-tissue imputation models within a single dataset, we used GTEx version 6. When evaluating imputation models across multiple tissues and multiple datasets, we used various combinations of GTEx versions to evaluate the benefit of multiple imputation models trained from overlapping datasets. When training across different datasets, genes were matched by ensemble ID, ignoring version numbers. In addition to training SWAM, we also used the *single tissue* PredictDB imputation models as a basis for comparison with our method.

2.5.5.2 Validation dataset from the GEUVADIS study

We used individual-level genotypes and expression levels from lymphoblastoid cell lines (LCL) from the GEUVADIS consortium only to evaluate various methods after imputing expression levels with models built from other datasets. Each imputation model was evaluated by applying the model to GEUVADIS genotypes to impute individual expression levels, and by calculating the correlation between the imputed and measured expressions. We focused on 344 European individuals where genotypes and normalized expressions (from RNA-seq) are available, with comparable linkage disequilibrium (LD) structure to GTEx and DGN datasets.

2.5.5.3 Imputation models from Depression Genes Network

We also downloaded the imputation model trained using the 922 whole blood transcriptomes from the Depression Genes Network (DGN) via PredictDB. DGN was evaluated as a single-tissue

imputation model. It was also used in the evaluation of multi-dataset imputation models when DGN is combined with various versions of GTEx imputation models.

2.5.5.4 Imputation models from UTMOST

We compared our methods to *UTMOST*, another multi-tissue approach for expression imputation[93]. The *UTMOST* imputation models were jointly trained across 44 tissues from GTEx version 6 and were downloaded from their published online repository (https://github.com/Joker-Jerome/UTMOST). We applied the imputation model targeted for EBV-transformed lymphocytes when evaluating the imputation accuracy with the GEUVADIS LCL expression.

2.5.6 Experimental Evaluation with Real Datasets

2.5.6.1 Evaluating imputation accuracy with GEUVADIS measured expression

We evaluated the accuracy of various imputation models by comparing imputed expressions from individual-level genotypes with the measured expression from GEUVADIS LCLs. Individual-level expression were imputed across 344 European GEUVADIS samples using various single-tissue, multi-tissue/multi-dataset methods to calculate the correlation with the normalized measured expression from GEUVADIS LCL. The correlation between imputed and measured expressions were calculated using spearman correlation and a one-sided p-value was evaluated by converting the correlation coefficients into t-statistics. Genes were considered "significantly predictable" if the Benjamini-Hochberg false discovery rate (FDR) was less than 0.05. This procedure was applied across all genes within each method, with the counts being tabulated.

2.5.6.2 Comparing single-tissue and multi-tissue imputation models within a single dataset.

With these results, we first focused on comparing the imputation accuracy of SWAM with other methods using GTEx v6. We compared *SWAM-LCL* (SWAM using GTEx EBV-transformed lymphocytes as reference), every *single tissue* imputation model from PredictDB, *UTMOST-LCL* (*UTMOST* using GTEx EBV-transformed lymphocytes as reference), *naïve average*, and *best tissue* methods. We focused on evaluation using GTEx v6 models where *UTMOST* models were

available. We also focused on genes included in the Consensus Coding Sequence Project (CCDS) [143] to minimize the discrepancy between imputation models.

To keep a fair comparison with *UTMOST* and the *single tissue* methods, we restricted the set of genes to those that have at least one eQTL in any *single tissue* models from PredictDB and also in any *UTMOST* models across all reference tissues.

2.5.6.3 Evaluating multi-tissue imputation models across multiple datasets.

Our second comparison was conducted to examine the effect of integrating multiple imputation models trained from heterogeneous datasets into SWAM. Here, we used various combinations of GTEx and DGN resources to derive multi-tissue/multi-dataset models, such as combining GTEx v6 with DGN data, or combining GTEx v6, v7 and v8 altogether. For this analysis, the gene list was restricted to genes that were included in all three of the v6, v7 and v8 datasets in terms of Ensemble IDs.

2.5.7 Evaluation of SWAM in transcriptome-wide association studies (TWAS)

To evaluate our method in the context of TWAS, we used MetaXcan [137], which infers TWAS results from GWAS summary statistics. We focused on the HDL and LDL traits from Global Lipids Genetics Consortium (GLGC) [144] and Type-2 Diabetes (T2D) from the DIAGRAM consortium [145]. For this analysis, we generated SWAM imputation models targeting each of the 44 tissues from GTEx version 6. We used MetaXcan to infer the TWAS results for each of these tissues and applied a Bonferroni correction with false-positive rate of 0.05 based on the number of genes tested. We repeated this with all 44 UTMOST models as well as all 44 PrediXcan *single tissue* models.

We also compared our method with S-MultiXcan [94], a recently published extension of MetaXcan which uses a principal components regression to conduct trait-expression association with multiple tissues.

2.6 Figures

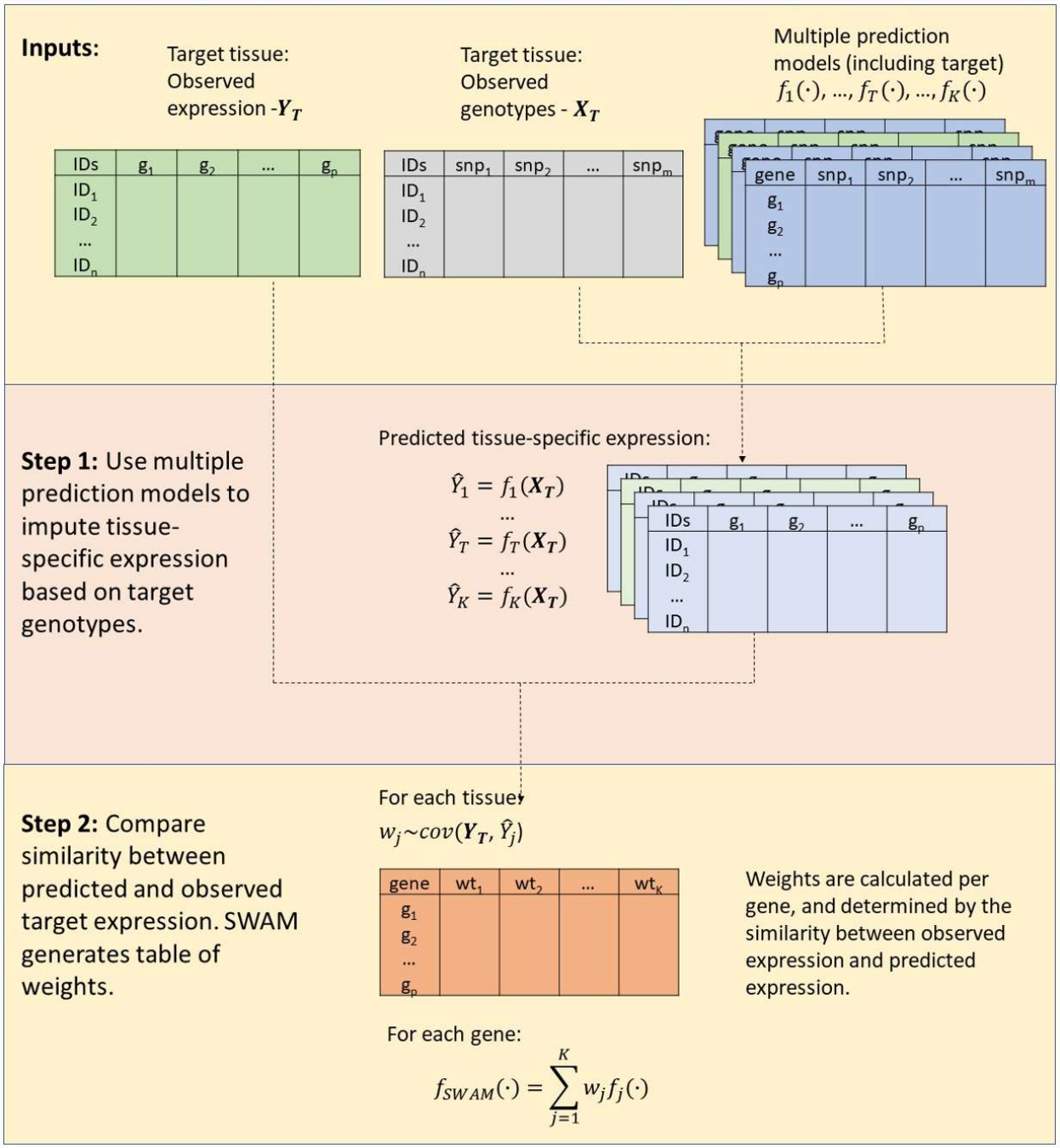


Figure 2.1 – overview of SWAM method.

This figure demonstrates the training of the imputation model using the reference data. The inputs required for SWAM are a set of reference genotypes with sample matched measured expression, and the multiple prediction models to be included. The list of multiple prediction models must also include a model derived from the reference data, which can be done via prediXcan. SWAM uses these models to impute tissue-specific expression levels from the reference genotypes. These imputed expression sets are then compared with the measured expression of the reference set. The weights are calculated based on the similarity between the measured and predicted expression and the covariance structure of tissues. For full details, see the methods section.

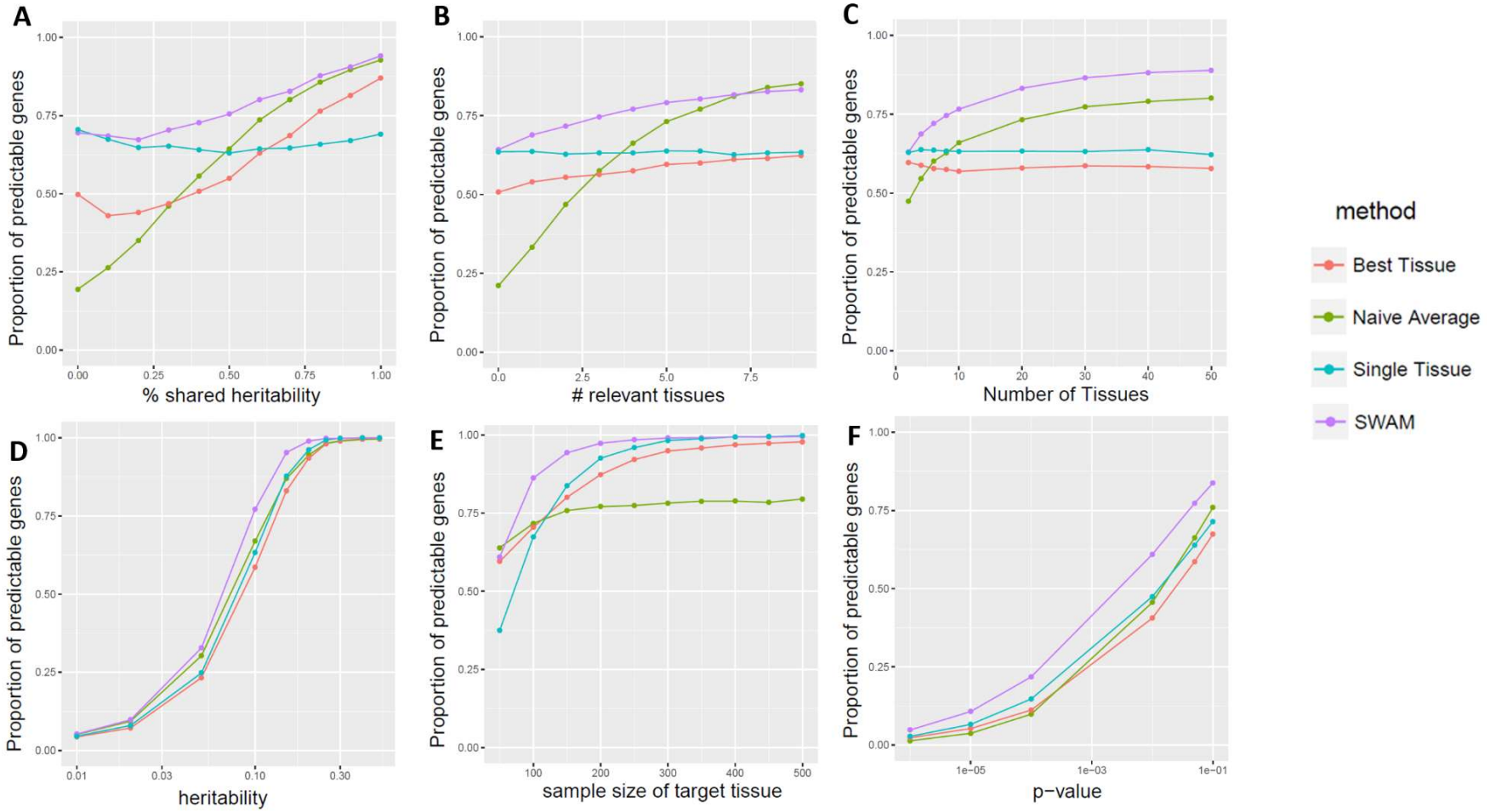


Figure 2.2– simulation study comparing SWAM with naïve average, best tissue and single tissue methods.

We ran each simulation 10,000 times, with the following default settings: 10 total tissues (1 target, 4 relevant, 5 irrelevant), 100 SNPs (2 per tissue), 10% genetic heritability, 50% shared heritability between relevant tissues. In addition, the sample size of the target tissue was 100 individuals, and the remaining tissues had 200 individuals. This was done to emphasize the importance of integrating information from other tissues when the quality of the target tissue model is limited. In panel (A), we varied the number of relevant tissues, from 0 to 10. Panel (B) shows the improvement when the total number of tissues is increased, with the number of irrelevant tissues fixed at 50% of the total. Panel (C) shows the effects of changing the shared heritability for the relevant tissues. We note here, that each tissue has 2 causal SNPS – for the relevant tissues, 1 of these causal SNPS is shared with the target tissue while the other is independent of all simulated tissues. Panel (D) shows the performance of the approaches for different levels of genetic heritability. This simulation demonstrates the range of heritability that we would expect to see the most improvement. Empirically, we do notice the same trend seen here, as SWAM performs similarly the single tissue model when the cross-validated R-squared is high. Panel (E) shows the effects of target tissue sample size. The x-axis pertains to the sample size of the target tissue only, and all other tissues were fixed at 200 individuals. Finally, panel (F) shows the performance of the methods at different p-value thresholds, using the default simulation settings.

Figure 2.3 – Empirical validation of SWAM using lymphoblastoid-cell line data from GEUVADIS consortium.

We used our LCL-targeted SWAM model to predict expression levels based on the genotypes of 344 European samples. We then calculated the concordance between imputed expression and measured LCL expression. We repeated this for all of the other methods mentioned here. (A) shows the performance of SWAM against the single-tissue models from 44 tissue-specific predictDB models derived from GTEx version 6. In (B), we derived various SWAM models using every combination of the following: 1) all GTEx v6 tissues, 2) all GTEx v7 tissues, 3) all GTEx v8 tissues, and 4) Depression Gene Network (DGN) single tissue whole blood model from predictDB. Here, we also included the UTMOST LCL model, naïve average and best tissue models, all derived from GTEx v6.

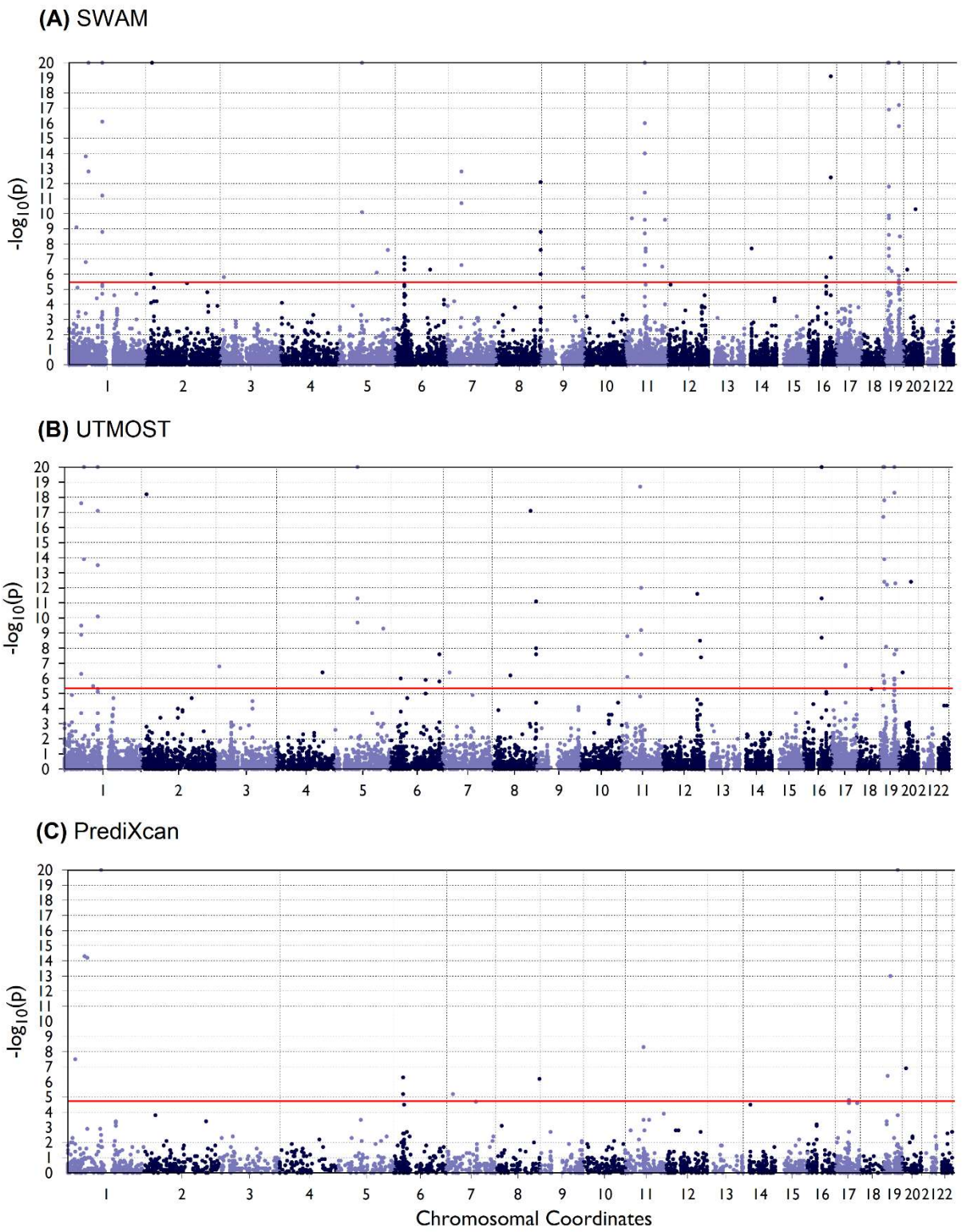


Figure 2.4 – TWAS on LDL trait targeting liver using SWAM, UTMOST and PrediXcan models

TWAS was performed using metaXcan on the LDL trait from the Global Lipids Genetics Consortium (GLGC) GWA analysis. For a consistent comparison, the SWAM and UTMOST models were derived from GTEx version 6 tissues, and the prediXcan model used was GTEx v6 liver. The number of associations were: 74, 69 and 19 for SWAM, UTMOST and prediXcan respectively. P-values were truncated at 10^{-20} in these plots.

2.7 Supplementary Materials

2.7.1 Derivation of weights for SWAM

In this section we derive the equation for the weights in SWAM. We wish to impute expression for a reference sample of N individuals with genotypes X_r and measured tissue-specific expression $\boldsymbol{y_r}$. Suppose we have single tissue prediction models for K tissues, with $r \in \{1, \dots, K\}$. For each gene g, we obtain a set of K predicted expression levels $\hat{\boldsymbol{s}}_{\boldsymbol{j}}^{\boldsymbol{g}} = X_T \hat{\beta}_j^g$, with $j \in (1, \dots, K)$. Dropping the superscript g for convenience, we define $w = (w_1, w_2, \dots, w_K)'$ be the set of weights corresponding to each of the tissues. The SWAM estimator is thus:

$$\widehat{\boldsymbol{m}}_{\boldsymbol{SWAM}} = \sum_{j=1}^{k} w_j \hat{\boldsymbol{s}}_{\boldsymbol{j}}$$

For further convenience, we denote $\widehat{\boldsymbol{m}}_{SWAM}$ as $\widehat{\boldsymbol{m}}$. Then, for each gene separately, the values for w are determined by minimizing the expression:

$$E\left[\| \widehat{\boldsymbol{m}} - y_r \|_2^2\right] = E\left[\left\| \sum_{i=1}^{k} (w_i \hat{\boldsymbol{s}}_i) - \boldsymbol{y_r} \right\|_2^2\right] = E\left[\sum_{i=1}^{k} w_i^2 \, (\hat{\boldsymbol{s}}_{\boldsymbol{i}} - \boldsymbol{y_r})^2\right]$$

Without loss of generality, we set the constraint $\sum_{i=1}^{k} w_i = 1$. The objective function to be minimized is

$$\begin{aligned} \mathcal{L}(\boldsymbol{w}, \lambda) &= E\left[\|\widehat{\boldsymbol{m}} - \boldsymbol{y_r}\|_2^2\right] + \gamma \left(\sum_{i=1}^{k} w_i - 1 \right) \\ &= \sum_{i=1}^{k} w_i^2 E[(\hat{\boldsymbol{s}}_{\boldsymbol{i}} - \boldsymbol{y_r})^2] + 2 \sum_{i=1}^{k} \sum_{j=1}^{i-1} w_i w_j E\left[(\hat{\boldsymbol{s}}_{\boldsymbol{i}} - \boldsymbol{y_r}) \, ' (\hat{\boldsymbol{s}}_{\boldsymbol{j}} - \boldsymbol{y_r})\right] \\ &\quad + \gamma \left(\sum_{i=1}^{k} w_i - 1 \right) \end{aligned}$$

The gradient of $\mathcal{L}(\boldsymbol{w}, \lambda)$ is

$$\nabla\mathcal{L}(\boldsymbol{w},\lambda)=\begin{bmatrix} w_i E[(\hat{\boldsymbol{s}}_{\boldsymbol{i}}-\boldsymbol{y}_{\boldsymbol{r}})^2]+2\sum_{j\neq i} w_j E\left[(\hat{\boldsymbol{s}}_{\boldsymbol{i}}-\boldsymbol{y}_{\boldsymbol{r}})'(\hat{\boldsymbol{s}}_{\boldsymbol{j}}-\boldsymbol{y}_{\boldsymbol{r}})\right]+\gamma \\ \sum_{j=1}^{k} w_j-1 \end{bmatrix}, i\in\{1,\cdots,K\}$$

Solving this system of equations, we obtain the optimal weighting minimizing the expected MSE across single tissue imputed expressions as

$$w_i=\frac{[S^{-1}\mathbf{1}]_i}{\sum_{j=1}^{K}[S^{-1}\mathbf{1}]_j}$$

Where $S=\begin{bmatrix} E[(\hat{\boldsymbol{s}}_{\mathbf{1}}-\boldsymbol{y}_{\boldsymbol{r}})^2] & \cdots & E[(\hat{\boldsymbol{s}}_{\mathbf{1}}-\boldsymbol{y}_{\boldsymbol{r}})'(\hat{\boldsymbol{s}}_{\boldsymbol{K}}-\boldsymbol{y}_{\boldsymbol{r}})] \\ \vdots & \ddots & \vdots \\ E[(\hat{\boldsymbol{s}}_{\boldsymbol{K}}-\boldsymbol{y}_{\boldsymbol{r}})'(\hat{\boldsymbol{s}}_{\mathbf{1}}-\boldsymbol{y}_{\boldsymbol{r}})] & \cdots & E[(\hat{\boldsymbol{s}}_{\boldsymbol{K}}-\boldsymbol{y}_{\boldsymbol{r}})^2] \end{bmatrix}$ and $\mathbf{1}$ = (1,…,1)'

2.7.2 Regularization of weights

The weights derived in the previous section provide an optimal solution to the expression $\underset{w^g}{\operatorname{argmin}} E\left[\left\|\hat{Y}_{mt}^g-Y_T^g\right\|_2^2\right]$. In the scenario in which the tissues are highly correlated with each other, the matrix $cov\left(\hat{Y}_{mt}^g\right)^{-1}=\begin{bmatrix} \langle\hat{Y}_1^g,\hat{Y}_1^g\rangle & \cdots & \langle\hat{Y}_K^g,\hat{Y}_1^g\rangle \\ \vdots & \ddots & \vdots \\ \langle\hat{Y}_K^g,\hat{Y}_1^g\rangle & \cdots & \langle\hat{Y}_K^g,\hat{Y}_K^g\rangle \end{bmatrix}^{-1}$ is numerically unstable as the columns of $cov\left(\hat{Y}_{mt}^g\right)$ are no longer linearly independent. This can lead to high weights assigned to irrelevant tissues and lower weights for relevant tissues. Furthermore, this may result in weights that are over-fitted to the noise of the data.

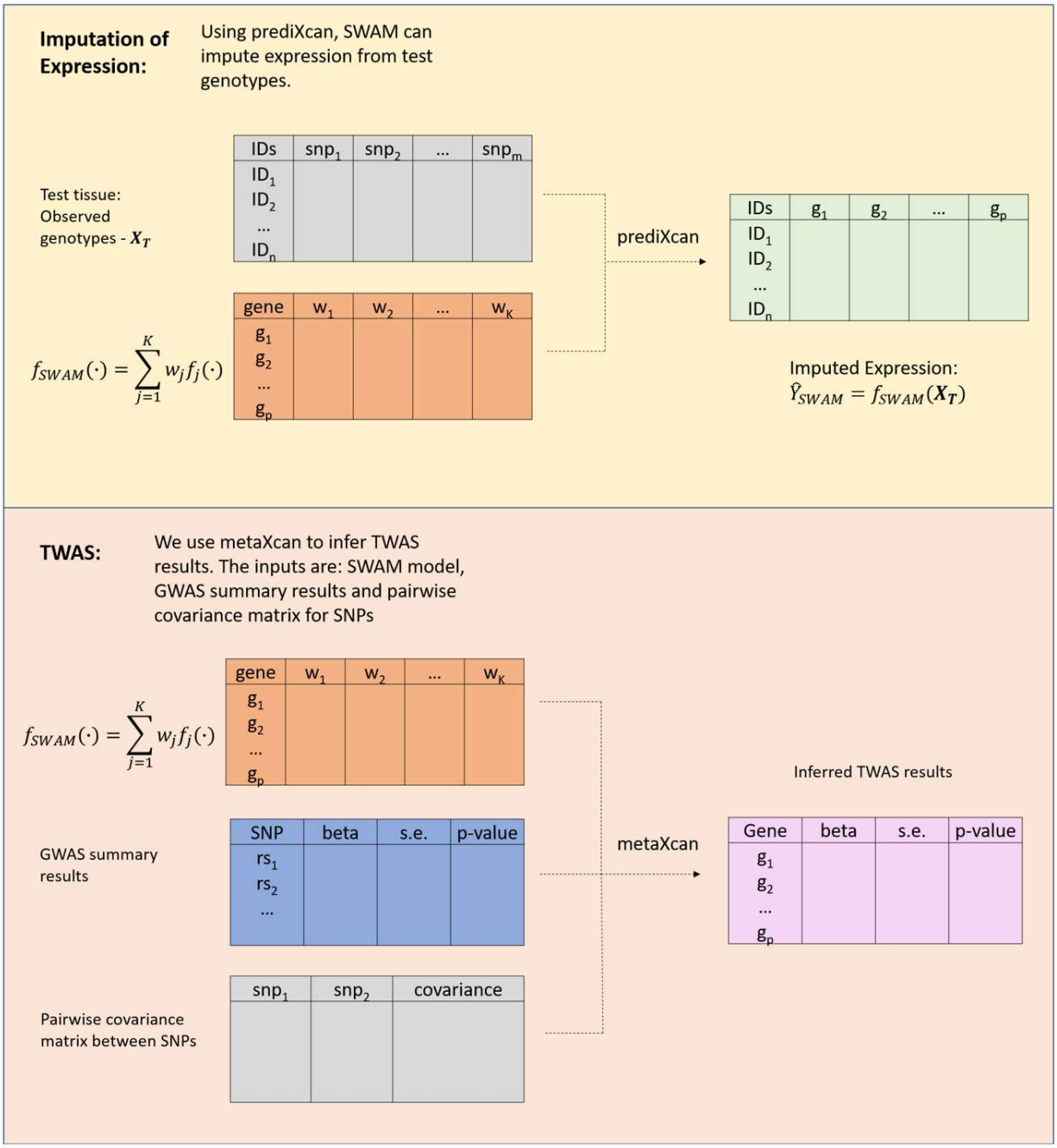
To correct for this, we added a diagonal matrix, λI prior to inverting the matrix $cov\left(\hat{Y}_{mt}^g\right)$, giving us the solution $w^g=\left[cov\left(\hat{Y}_{mt}^g\right)+\lambda I\right]^{-1}\widehat{cor}(\hat{Y}_{mt}^g,Y_T^g)$. To choose the correct value of λ, we tested the prediction accuracy of $\hat{Y}_{mt}^g$ in our validation test set for a large range of λ. We found that prediction accuracy was low when $\lambda=0$, likely due to overfitted and the amplification of noise. Larger values of λ yielded better results but ignored the correlation structure between

tissues. We found empirically that $\lambda = 3$ provided the best results (this value depends highly on the scale and normalization of the data).

2.7.3 Application of SWAM to other target tissues

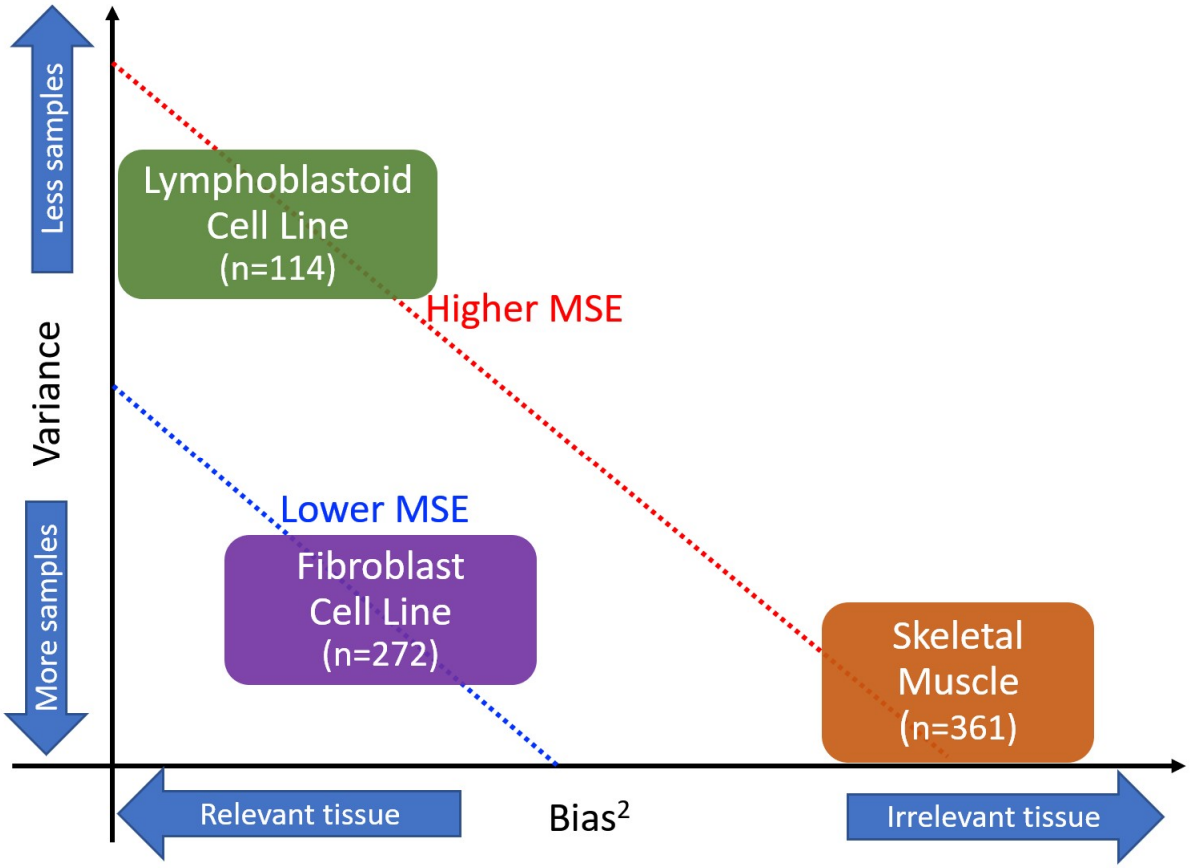
Throughout our work we primarily used the LCL tissue from GTEx version 6 as our target tissue for application of SWAM. In addition to producing SWAM-LCL models, we also generated models targeting each of the 44 GTEx v6 tissues. Supplementary Figure 2.3 displays the heatmap of weight contribution towards each of the tissues. The rows correspond to the SWAM model for each tissue type, and the color intensity of the columns show the contribution of each tissue towards the targeted tissue (number of times the tissue contributed the highest weight). Overall, we observe clustering that appears to separate the tissue types quite well. For example, brain tissues are primarily getting high weights from other brain tissues while receiving low weights from all other tissue types. This heatmap provides evidence of SWAM being able to capture tissue-specific signals.

2.8 Supplementary Figures and Tables



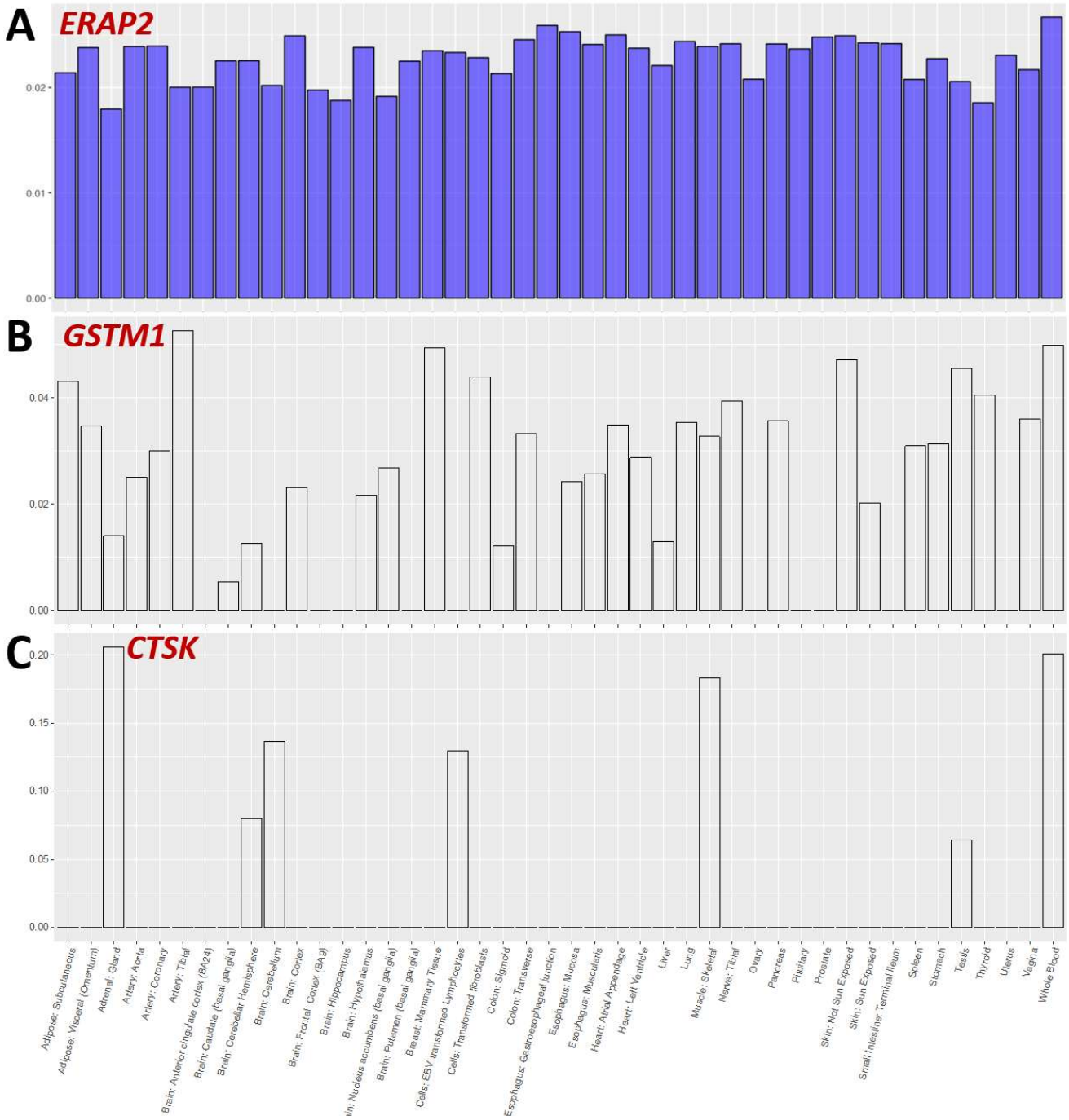
Supplementary Figure 2.1 – Using SWAM to impute expression and conduct TWAS

The first panel shows how SWAM can be used to predict expression levels via prediXcan, while the second panel shows the required inputs to conduct TWAS via metaXcan.



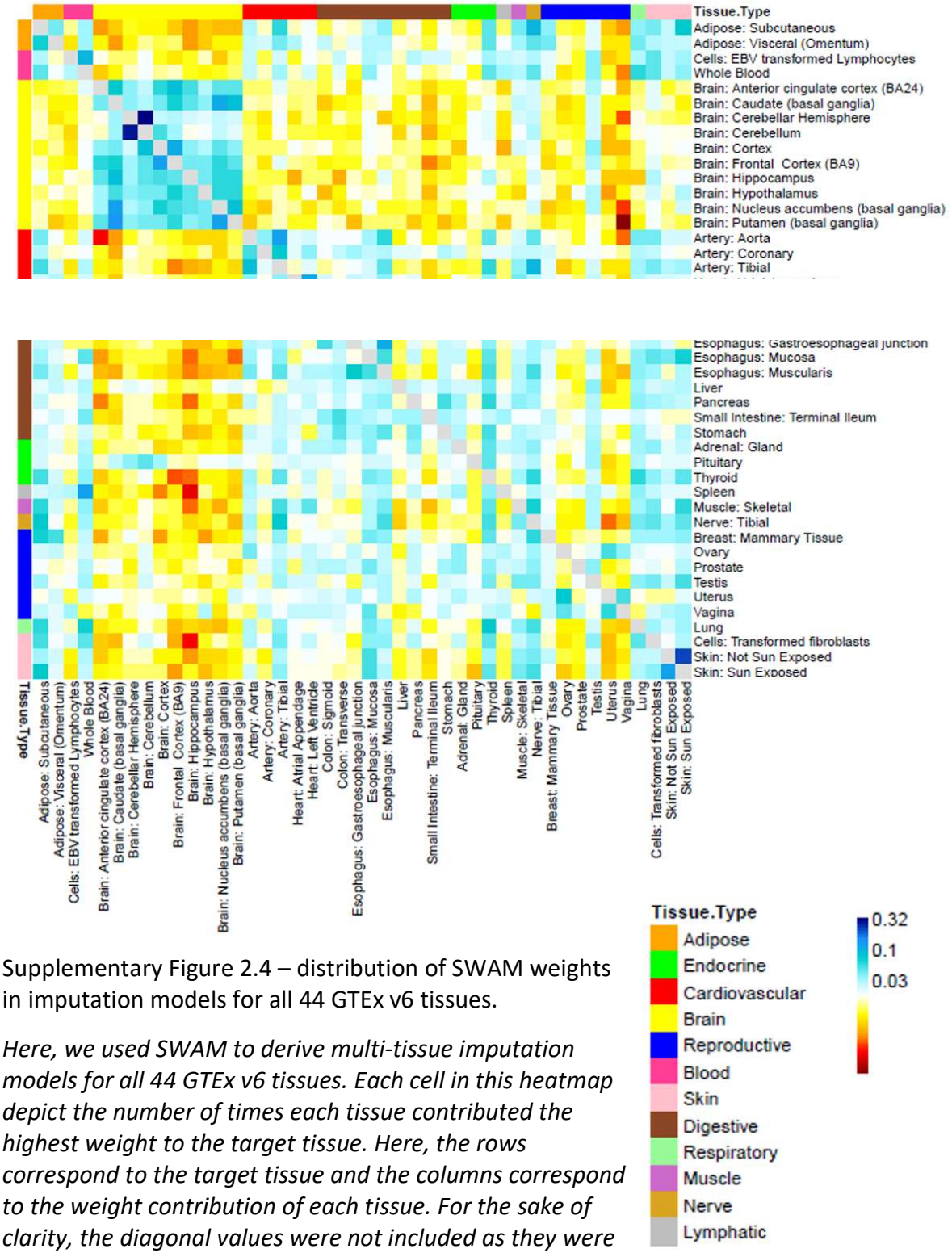
Supplementary Figure 2.2 – Bias-variance tradeoff for other tissues

The principal behind SWAM is it considers the bias-variance tradeoff for each tissue, and assigns higher weights to tissues that reduce MSE. In this example, tissues such as Skeletal Muscle have a high sample size (and therefore lower variance) but may be biased as they are not the relevant tissue to the tissue of interest (in this case LCL). Other tissues such as Fibroblasts may have a lower sample size but compensate by having low bias (high relevance to tissue of interest) and will contribute more weight.



Supplementary Figure 2.3– The distribution of weights for SWAM for three selected genes.

(2A) shows the ERAP 2 gene, which had a single tissue $r^2 = 0.801$, while the SWAM model had $r^2 = 0.795$. (2B) depicts are scenario where SWAM is able to leverage information from other tissues to make up for the relatively lower quality of the target tissue – here the single tissue model gave $r^2 = 0.368$ while SWAM increased the accuracy to $r^2 = 0.741$. (2C) shows an example where the eQTLs are highly tissue specific. Here, SWAM improved the single tissue accuracy from $r^2 = 0.111$ to $r^2 = 0.447$.



Supplementary Figure 2.4 – distribution of SWAM weights in imputation models for all 44 GTEx v6 tissues.

Here, we used SWAM to derive multi-tissue imputation models for all 44 GTEx v6 tissues. Each cell in this heatmap depict the number of times each tissue contributed the highest weight to the target tissue. Here, the rows correspond to the target tissue and the columns correspond to the weight contribution of each tissue. For the sake of clarity, the diagonal values were not included as they were consistently much higher than the remaining elements of the matrix.

Method	Total # genes	Genes with FDR < 0.05	P-value threshold for FDR=0.05	Genes with p-value < 0.05
DGN Whole Blood	13213	2390	0.009018	3450
Adipose Subcutaneous	13213	1500	0.005667	2203
Adipose Visceral Omentum	13213	963	0.003619	1432
Adrenal Gland	13213	735	0.002755	1151
Artery Aorta	13213	1177	0.004426	1783
Artery Coronary	13213	616	0.00231	926
Artery Tibial	13213	1439	0.005399	2124
Brain Anterior cingulate cortex BA24	13213	322	0.00117	580
Brain Caudate basal ganglia	13213	528	0.001979	895
Brain Cerebellar Hemisphere	13213	575	0.002172	1002
Brain Cerebellum	13213	661	0.002498	1144
Brain Cortex	13213	511	0.00193	878
Brain Frontal Cortex BA9	13213	456	0.001699	770
Brain Hippocampus	13213	336	0.001254	572
Brain Hypothalamus	13213	339	0.001264	554
Brain Nucleus accumbens basal ganglia	13213	454	0.001709	755
Brain Putamen basal ganglia	13213	391	0.001461	655
Breast Mammary Tissue	13213	933	0.003455	1430
Cells EBV-transformed lymphocytes	13213	1552	0.005845	1943
Cells Transformed fibroblasts	13213	1690	0.00635	2451
Colon Sigmoid	13213	680	0.002391	1052
Colon Transverse	13213	1013	0.003803	1536
Esophagus Gastroesophageal Junction	13213	716	0.002689	1101
Esophagus Mucosa	13213	1469	0.00553	2148
Esophagus Muscularis	13213	1272	0.004808	1959
Heart Atrial Appendage	13213	849	0.003167	1323
Heart Left Ventricle	13213	942	0.003537	1451
Liver	13213	427	0.001547	719
Lung	13213	1355	0.005122	1985
Muscle Skeletal	13213	1197	0.004454	1876
Nerve Tibial	13213	1450	0.005483	2186
Ovary	13213	417	0.001566	689
Pancreas	13213	911	0.003426	1406
Pituitary	13213	496	0.001875	808
Prostate	13213	398	0.001421	629
Skin Not Sun Exposed Suprapubic	13213	1063	0.003997	1619
Skin Sun Exposed Lower leg	13213	1463	0.005489	2145
Small Intestine Terminal Ileum	13213	450	0.001694	735
Spleen	13213	759	0.002821	1166

Stomach	13213	878	0.003291	1350
Testis	13213	956	0.003607	1563
Thyroid	13213	1454	0.005469	2208
Uterus	13213	297	0.001071	521
Vagina	13213	307	0.001149	511
Whole Blood	13213	1427	0.005343	2106
SWAM-LCL v6	13213	3040	0.01145	4148
NAIVE AVERAGE V6	13213	2666	0.010066	3830
BEST TISSUE V6	13213	2493	0.009394	3663
UTMOST-LCL v6	13213	2238	0.008466	3185

Supplementary Table 2.1 – GTEx version 6 comparisons of single-tissue and multi-tissue imputation models using GEUVADIS LCL RNA-Seq expression as validation.

Counts (B-H counts) are based on Benjamini-Hochberg procedure false discovery rate of 0.05. The last column displays the number of counts at p-value threshold 0.05 (without any corrections)

Method	Total # genes	Genes with FDR < 0.05	P-value threshold for FDR=0.05	Genes with p-value < 0.05
SWAM-LCL (GTEx v6)	13213	3040	0.01145	4148
SWAM-LCL (GTEx v6 + DGN)	13213	3192	0.012077	4301
SWAM-LCL (GTEx v7)	13213	3060	0.011463	4215
SWAM-LCL (GTEx v7 + DGN)	13213	3411	0.012903	4469
SWAM-LCL (GTEx v8)	13213	3203	0.01212	4236
SWAM-LCL (GTEx v8 + DGN)	13213	3449	0.013046	4460
SWAM-LCL (GTEx v6 + v7)	13213	3283	0.012383	4385
SWAM-LCL (GTEx v6 + v7 + DGN)	13213	3361	0.012674	4480
SWAM-LCL (GTEx v6 + v8)	13213	3134	0.01185	4274
SWAM-LCL (GTEx v6 + v8 + DGN)	13213	3275	0.012389	4384
SWAM-LCL (GTEx v7 + v8)	13213	3259	0.01227	4326
SWAM-LCL (GTEx v7 + v8 + DGN)	13213	3368	0.012737	4478
SWAM-LCL (GTEx v6 + v7 + v8)	13213	3342	0.012641	4448
SWAM-LCL (GTEx v6 + v7 + v8 + DGN)	13213	3413	0.012878	4526
UTMOST-LCL	13213	2238	0.008466	3185
NAIVE AVERAGE	13213	2666	0.010066	3830
BEST TISSUE	13213	2493	0.009394	3663

Supplementary Table 2.2– Comparison of all multi-tissue methods

We applied SWAM to all combinations of GTEx and DGN resources. For the GTEx resources, we always used every tissue available. In version 6, this comprised of 44 tissues. For version 7, there were 48 tissues and version 8 contained 49 tissues. For the sake of consistency, our target tissue for each of these combinations was GTEx v6 LCL.

Tissue	GTEx v7			GTEx v8		
	sample size	Total # genes	Genes with FDR < 0.05	sample size	Total # genes	Genes with FDR < 0.05
Adipose Subcutaneous	328	6689	2259	581	6847	2114
Adipose Visceral Omentum	273	5286	1905	469	5769	1957
Adrenal Gland	146	3784	1322	233	3816	1279
Artery Aorta	236	5488	1850	387	6090	1844
Artery Coronary	128	2838	1040	213	3177	1119
Artery Tibial	329	6836	2131	584	6955	2023
Brain Amygdala	81	1876	581	129	2081	682
Brain Anterior cingulate cortex BA24	102	2628	838	147	2662	840
Brain Caudate basal ganglia	126	3272	1035	194	3795	1146
Brain Cerebellar Hemisphere	113	3810	1050	175	4482	1127
Brain Cerebellum	137	4899	1310	209	5254	1299
Brain Cortex	119	3422	1073	205	4169	1203
Brain Frontal Cortex BA9	104	2812	879	175	3424	1019
Brain Hippocampus	99	2217	708	165	2806	886
Brain Hypothalamus	98	2219	710	170	2742	911
Brain Nucleus accumbens basal ganglia	114	2820	921	202	3629	1076
Brain Putamen basal ganglia	98	2542	806	170	3365	1035
Brain Spinal cord cervical c-1	76	2003	600	126	2455	744
Brain Substantia nigra	70	1609	496	114	1892	570
Breast Mammary Tissue	211	4280	1622	396	5076	1756
Cells EBV-transformed lymphocytes	96	2777	1731	147	2537	1620
Cells Transformed fibroblasts	256	6297	2336	483	7421	2428
Colon Sigmoid	185	4257	1556	318	4847	1633
Colon Transverse	210	4457	1771	368	4923	1781
Esophagus Gastroesophageal Junction	185	4325	1603	330	4964	1675
Esophagus Mucosa	307	6744	2305	497	6872	2167
Esophagus Muscularis	287	6354	2119	465	6554	2030
Heart Atrial Appendage	231	4866	1675	372	5262	1696
Heart Left Ventricle	233	4449	1491	386	4902	1569
Kidney Cortex	NA	NA	NA	73	1205	344
Liver	134	2746	883	208	2983	976
Lung	333	6251	2190	515	6173	2071
Minor Salivary Gland	74	1785	652	144	2161	842
Muscle Skeletal	421	6323	1901	706	6261	1762
Nerve Tibial	305	7512	2179	532	7764	2051
Ovary	99	2406	818	167	2751	909

Pancreas	180	4369	1504	305	4710	1532
Pituitary	143	3691	1228	237	4262	1381
Prostate	114	2487	926	221	3205	1125
Skin Not Sun Exposed Suprapubic	285	6092	1997	517	6802	2011
Skin Sun Exposed Lower leg	359	7221	2201	605	7203	2071
Small Intestine Terminal Ileum	103	2482	991	174	2844	1145
Spleen	119	3753	1514	227	4527	1720
Stomach	200	3899	1518	324	4064	1542
Testis	191	5919	1451	322	6470	1518
Thyroid	344	7556	2249	574	7468	2130
Uterus	82	1957	681	129	1944	699
Vagina	91	1889	690	141	1919	725
Whole Blood	315	5432	1915	670	6195	2082

Supplementary Table 2.3 – comparison of GTEx v7/v8 single tissue models versus GEUVADIS LCL

We also compared every prediXcan model derived from GTEx version 7 and version 8 tissues, and tested prediction accuracy against GEUVADIS LCL measured expression levels. Surprisingly, despite the increase in sample size, the LCL tissue from v8 performed worse than its version 7 counterpart. The number of tissues outperforming LCL in both v7 and v8 highlight the opportunity to leverage information from other tissues to improve prediction accuracy for under-powered tissues.

Tissue	HDL			LDL			T2D		
	# sig genes	p-value threshold	total genes	# sig genes	p-value threshold	total genes	# sig genes	p-value threshold	total genes
Adipose Subcutaneous	78	3.23E-06	15501	79	3.22E-06	15507	8	3.19E-06	15669
Adipose Visceral Omentum	80	3.26E-06	15326	71	3.26E-06	15333	8	3.23E-06	15476
Adrenal Gland	70	3.34E-06	14961	78	3.34E-06	14973	8	3.31E-06	15110
Artery Aorta	74	3.34E-06	14990	78	3.33E-06	15000	10	3.31E-06	15125
Artery Coronary	76	3.33E-06	15001	69	3.33E-06	15009	7	3.30E-06	15147
Artery Tibial	84	3.33E-06	14994	81	3.33E-06	15001	11	3.30E-06	15145
Brain Anterior cingulate cortex BA24	87	3.36E-06	14892	70	3.36E-06	14901	6	3.32E-06	15061
Brain Caudate basal ganglia	88	3.29E-06	15216	75	3.28E-06	15221	11	3.25E-06	15372
Brain Cerebellar Hemisphere	75	3.39E-06	14742	74	3.39E-06	14755	4	3.36E-06	14891
Brain Cerebellum	71	3.34E-06	14991	81	3.33E-06	15006	5	3.30E-06	15150
Brain Cortex	78	3.29E-06	15198	91	3.29E-06	15208	6	3.25E-06	15366
Brain Frontal Cortex BA9	74	3.32E-06	15061	65	3.32E-06	15073	5	3.28E-06	15239
Brain Hippocampus	79	3.31E-06	15098	71	3.31E-06	15106	9	3.27E-06	15269
Brain Hypothalamus	67	3.28E-06	15265	72	3.27E-06	15274	7	3.24E-06	15419
Brain Nucleus accumbens basal ganglia	73	3.30E-06	15152	83	3.30E-06	15166	7	3.27E-06	15313
Brain Putamen basal ganglia	86	3.35E-06	14926	79	3.35E-06	14935	9	3.32E-06	15066
Breast Mammary Tissue	77	3.19E-06	15682	79	3.19E-06	15687	11	3.16E-06	15838
Cells EBV-transformed lymphocytes	75	3.75E-06	13344	76	3.75E-06	13347	9	3.70E-06	13504
Cells Transformed fibroblasts	74	3.55E-06	14091	81	3.55E-06	14098	10	3.51E-06	14253
Colon Sigmoid	73	3.31E-06	15124	79	3.30E-06	15138	11	3.27E-06	15291
Colon Transverse	78	3.23E-06	15500	73	3.22E-06	15507	7	3.19E-06	15681
Esophagus Gastroesophageal Junction	79	3.34E-06	14988	78	3.33E-06	14993	8	3.30E-06	15140
Esophagus Mucosa	84	3.27E-06	15268	81	3.27E-06	15275	7	3.24E-06	15437
Esophagus Muscularis	77	3.29E-06	15199	93	3.29E-06	15205	7	3.26E-06	15343
Heart Atrial Appendage	88	3.36E-06	14879	90	3.36E-06	14890	12	3.33E-06	15030
Heart Left Ventricle	73	3.43E-06	14558	82	3.43E-06	14569	8	3.40E-06	14696

Liver	84	3.49E-06	14325	74	3.49E-06	14337	10	3.45E-06	14497
Lung	92	3.16E-06	15813	73	3.16E-06	15822	7	3.13E-06	15974
Muscle Skeletal	84	3.43E-06	14560	71	3.43E-06	14564	6	3.40E-06	14696
Nerve Tibial	88	3.22E-06	15548	87	3.21E-06	15559	9	3.18E-06	15706
Ovary	94	3.39E-06	14754	91	3.39E-06	14760	12	3.36E-06	14898
Pancreas	73	3.36E-06	14891	66	3.36E-06	14900	11	3.33E-06	15026
Pituitary	82	3.22E-06	15517	87	3.22E-06	15530	7	3.19E-06	15694
Prostate	84	3.24E-06	15420	79	3.24E-06	15429	12	3.21E-06	15588
Skin Not Sun Exposed Suprapubic	83	3.22E-06	15545	81	3.21E-06	15555	6	3.18E-06	15735
Skin Sun Exposed Lower leg	85	3.18E-06	15729	75	3.18E-06	15738	7	3.15E-06	15891
Small Intestine Terminal Ileum	78	3.28E-06	15265	67	3.27E-06	15281	7	3.23E-06	15462
Spleen	78	3.36E-06	14873	69	3.36E-06	14884	18	3.33E-06	15037
Stomach	80	3.23E-06	15495	74	3.22E-06	15506	7	3.19E-06	15658
Testis	80	3.03E-06	16520	83	3.03E-06	16528	9	2.98E-06	16764
Thyroid	78	3.18E-06	15705	70	3.18E-06	15714	10	3.15E-06	15876
Uterus	84	3.42E-06	14641	90	3.41E-06	14654	8	3.38E-06	14803
Vagina	84	3.30E-06	15157	84	3.30E-06	15167	7	3.26E-06	15328
Whole Blood	78	3.49E-06	14331	71	3.49E-06	14340	7	3.45E-06	14505
Average	79.71			77.75			8.43		

Supplementary Table 2.4– TWAS association signals for SWAM

We used SWAM to derive an tissue-specific model for every GTEx version 6 tissue, and used these models as inputs to metaXcan to infer TWAS results. As mentioned in the methods section, the HDL and LDL traits were from Global Lipids Genetics Consortium (GLGC) [130] and Type-2 Diabetes (T2D) from the DIAGRAM consortium [131].

Tissue	HDL			LDL			T2D		
	# sig genes	p-value threshold	total genes	# sig genes	p-value threshold	total genes	# sig genes	p-value threshold	total genes
Adipose Subcutaneous	82	4.18E-06	11964	61	4.18E-06	11969	11	3.94E-06	12688
Adipose Visceral Omentum	69	4.26E-06	11741	68	4.26E-06	11743	10	4.01E-06	12475
Adrenal Gland	72	4.55E-06	10998	66	4.54E-06	11004	7	4.26E-06	11737
Artery Aorta	75	4.47E-06	11180	55	4.47E-06	11184	7	4.19E-06	11926
Artery Coronary	61	4.39E-06	11391	58	4.39E-06	11397	7	4.13E-06	12114
Artery Tibial	74	4.43E-06	11292	60	4.43E-06	11295	9	4.15E-06	12044
Brain Anterior cingulate cortex BA24	53	5.00E-06	10002	62	5.00E-06	10003	8	4.64E-06	10779
Brain Caudate basal ganglia	61	4.61E-06	10850	55	4.61E-06	10850	11	4.30E-06	11627
Brain Cerebellar Hemisphere	65	4.94E-06	10113	64	4.94E-06	10115	11	4.60E-06	10864
Brain Cerebellum	62	4.78E-06	10457	67	4.78E-06	10457	7	4.46E-06	11205
Brain Cortex	61	4.69E-06	10650	57	4.70E-06	10647	8	4.37E-06	11445
Brain Frontal Cortex BA9	66	4.67E-06	10717	57	4.66E-06	10719	11	4.34E-06	11519
Brain Hippocampus	58	4.67E-06	10701	54	4.67E-06	10701	8	4.36E-06	11469
Brain Hypothalamus	75	4.55E-06	10986	65	4.55E-06	10987	11	4.25E-06	11764
Brain Nucleus accumbens basal ganglia	74	4.64E-06	10781	59	4.64E-06	10783	9	4.33E-06	11541
Brain Putamen basal ganglia	71	4.84E-06	10323	55	4.84E-06	10327	11	4.49E-06	11129
Breast Mammary Tissue	74	4.12E-06	12141	62	4.12E-06	12143	8	3.88E-06	12899
Cells EBV-transformed lymphocytes	65	5.20E-06	9610	50	5.20E-06	9615	5	4.87E-06	10267
Cells Transformed fibroblasts	61	4.80E-06	10406	60	4.80E-06	10409	10	4.51E-06	11089
Colon Sigmoid	72	4.44E-06	11257	68	4.44E-06	11261	7	4.16E-06	12010
Colon Transverse	74	4.25E-06	11759	61	4.25E-06	11765	10	3.99E-06	12545
Esophagus Gastroesophageal Junction	58	4.50E-06	11117	52	4.50E-06	11119	8	4.23E-06	11823
Esophagus Mucosa	77	4.31E-06	11595	71	4.31E-06	11603	7	4.06E-06	12316
Esophagus Muscularis	67	4.43E-06	11285	67	4.43E-06	11289	7	4.16E-06	12029
Heart Atrial Appendage	69	4.58E-06	10922	64	4.58E-06	10923	9	4.28E-06	11679
Heart Left Ventricle	77	4.75E-06	10519	55	4.75E-06	10522	10	4.45E-06	11242
Liver	63	4.97E-06	10062	69	4.97E-06	10068	7	4.63E-06	10808

Lung	81	4.02E-06	12428	62	4.02E-06	12433	8	3.80E-06	13152
Muscle Skeletal	81	4.70E-06	10629	65	4.70E-06	10631	6	4.41E-06	11326
Nerve Tibial	78	4.24E-06	11784	63	4.24E-06	11787	7	4.00E-06	12511
Ovary	57	4.71E-06	10608	60	4.71E-06	10607	8	4.42E-06	11323
Pancreas	60	4.71E-06	10619	63	4.71E-06	10625	7	4.38E-06	11410
Pituitary	75	4.41E-06	11336	69	4.41E-06	11337	8	4.13E-06	12117
Prostate	73	4.28E-06	11685	60	4.28E-06	11688	8	4.02E-06	12450
Skin Not Sun Exposed Suprapubic	75	4.24E-06	11789	69	4.24E-06	11797	7	4.00E-06	12505
Skin Sun Exposed Lower leg	72	4.12E-06	12144	62	4.12E-06	12149	8	3.87E-06	12904
Small Intestine Terminal Ileum	68	4.48E-06	11150	64	4.48E-06	11151	10	4.19E-06	11938
Spleen	70	4.66E-06	10736	55	4.66E-06	10739	13	4.36E-06	11474
Stomach	78	4.23E-06	11833	68	4.22E-06	11838	14	3.97E-06	12580
Testis	73	4.03E-06	12411	70	4.03E-06	12412	7	3.78E-06	13222
Thyroid	82	4.13E-06	12106	68	4.13E-06	12113	17	3.88E-06	12873
Uterus	60	4.93E-06	10148	61	4.93E-06	10151	8	4.62E-06	10813
Vagina	70	4.43E-06	11285	59	4.43E-06	11288	7	4.16E-06	12018
Whole Blood	59	4.76E-06	10511	52	4.75E-06	10518	12	4.48E-06	11168
Average	69.27			61.64			8.84		

Supplementary Table 2.5 – TWAS association signals for UTMOST

These models were also derived from GTEx version 6 tissues using the UTMOST method. Models were downloaded from https://github.com/Joker-Jerome/UTMOST

Tissue	HDL			LDL			T2D		
	# sig genes	p-value threshold	total genes	# sig genes	p-value threshold	total genes	# sig genes	p-value threshold	total genes
TW Adipose Subcutaneous	46	7.54E-06	6634	28	7.54E-06	6628	10	6.91E-06	7234
TW Adipose Visceral Omentum	32	1.20E-05	4174	21	1.20E-05	4173	6	1.10E-05	4542
TW Adrenal Gland	23	1.33E-05	3760	20	1.33E-05	3758	3	1.24E-05	4048
TW Artery Aorta	33	8.76E-06	5706	44	8.76E-06	5705	5	8.11E-06	6163
TW Artery Coronary	13	1.65E-05	3025	21	1.65E-05	3024	3	1.54E-05	3242
TW Artery Tibial	36	7.50E-06	6666	32	7.51E-06	6657	6	6.89E-06	7259
TW Brain Anterior cingulate cortex BA24	6	2.03E-05	2466	14	2.03E-05	2466	2	1.88E-05	2654
TW Brain Caudate basal ganglia	21	1.48E-05	3375	25	1.48E-05	3372	2	1.38E-05	3616
TW Brain Cerebellar Hemisphere	12	1.26E-05	3955	18	1.26E-05	3954	3	1.18E-05	4228
TW Brain Cerebellum	22	1.10E-05	4543	28	1.10E-05	4542	4	1.04E-05	4830
TW Brain Cortex	19	1.49E-05	3351	33	1.49E-05	3347	3	1.40E-05	3583
TW Brain Frontal Cortex BA9	12	1.66E-05	3013	16	1.66E-05	3013	1	1.56E-05	3211
TW Brain Hippocampus	12	2.12E-05	2362	10	2.12E-05	2362	1	1.97E-05	2534
TW Brain Hypothalamus	6	2.20E-05	2269	11	2.20E-05	2268	1	2.04E-05	2455
TW Brain Nucleus accumbens basal ganglia	16	1.70E-05	2935	17	1.70E-05	2935	1	1.60E-05	3131
TW Brain Putamen basal ganglia	14	1.91E-05	2620	17	1.91E-05	2620	5	1.78E-05	2807
TW Breast Mammary Tissue	23	1.16E-05	4292	20	1.17E-05	4288	4	1.07E-05	4655
TW Cells EBV-transformed lymphocytes	22	1.43E-05	3493	17	1.43E-05	3491	3	1.33E-05	3751
TW Cells Transformed fibroblasts	52	7.02E-06	7120	35	7.03E-06	7114	7	6.50E-06	7692
TW Colon Sigmoid	16	1.40E-05	3561	16	1.40E-05	3559	4	1.30E-05	3858
TW Colon Transverse	24	1.11E-05	4485	24	1.12E-05	4480	4	1.03E-05	4874
TW Esophagus Gastroesophageal Junction	14	1.45E-05	3456	17	1.45E-05	3454	3	1.34E-05	3727
TW Esophagus Mucosa	38	7.78E-06	6425	32	7.78E-06	6423	3	7.18E-06	6961
TW Esophagus Muscularis	31	8.37E-06	5971	33	8.38E-06	5966	4	7.70E-06	6493
TW Heart Atrial Appendage	28	1.19E-05	4187	21	1.19E-05	4185	6	1.11E-05	4501
TW Heart Left Ventricle	25	1.11E-05	4517	33	1.11E-05	4513	4	1.02E-05	4885
TW Liver	16	1.86E-05	2695	19	1.86E-05	2692	4	1.72E-05	2909

TW Lung	42	8.32E-06	6008	16	8.33E-06	6002	5	7.56E-06	6611
TW Muscle Skeletal	30	8.32E-06	6009	30	8.33E-06	6003	8	7.56E-06	6614
TW Nerve Tibial	41	6.60E-06	7577	39	6.61E-06	7570	6	6.13E-06	8157
TW Ovary	13	1.94E-05	2576	17	1.94E-05	2575	3	1.81E-05	2762
TW Pancreas	26	1.12E-05	4449	26	1.12E-05	4447	4	1.05E-05	4771
TW Pituitary	12	1.59E-05	3145	13	1.59E-05	3144	5	1.48E-05	3382
TW Prostate	10	2.09E-05	2389	20	2.09E-05	2389	3	1.92E-05	2609
TW Skin Not Sun Exposed Suprapubic	27	9.37E-06	5336	31	9.38E-06	5333	4	8.62E-06	5798
TW Skin Sun Exposed Lower leg	42	7.10E-06	7041	35	7.11E-06	7037	6	6.55E-06	7628
TW Small Intestine Terminal Ileum	15	1.97E-05	2538	16	1.97E-05	2536	3	1.83E-05	2729
TW Spleen	19	1.45E-05	3456	18	1.45E-05	3456	5	1.35E-05	3698
TW Stomach	16	1.27E-05	3945	23	1.27E-05	3943	4	1.17E-05	4271
TW Testis	30	7.37E-06	6780	28	7.38E-06	6778	7	6.91E-06	7234
TW Thyroid	39	6.69E-06	7478	40	6.69E-06	7469	7	6.14E-06	8140
TW Uterus	11	2.51E-05	1991	8	2.51E-05	1991	1	2.34E-05	2139
TW Vagina	15	2.57E-05	1946	10	2.57E-05	1944	1	2.40E-05	2082
TW Whole Blood	42	8.23E-06	6077	30	8.23E-06	6073	3	7.42E-06	6743
Average	23.68			23.23			4.02		

Supplementary Table 2.6 – *TWAS association signals for prediXcan (single-tissue)*

TWAS results via metaXcan using prediXcan single tissue models derived from GTEx version 6 tissues

Chapter 3 Revisiting Microarray Hybridization Biases in the Whole Genome Sequencing Era

3.1 Abstract

Traditional expression quantitative trait loci (eQTL) studies based on microarrays have successfully identified tremendous numbers of cis-acting associations between genetic variants and expression levels. However, microarray probes that contain a genetic variant can have weakened hybridization to RNA molecules due to base-pair mismatches, which artificially reduces the gene expression levels for individuals with non-reference alleles. This bias can lead to significant statistical associations in eQTL studies that are technical false positives. While existing publications have developed methods to address this issue by inferring and removing problematic probes (via reference panels or contrasting between groups with and without affected probes), it is impossible to fully correct the bias completely without knowing all genetic polymorphisms within the samples. In this chapter, we demonstrate that the availability of deep genome sequence data can be used to empower and refine existing eQTL studies by allowing us to correct for the reference-bias in variant-overlapping probes.

Here, we leveraged whole genome sequence data from the Pima diabetic nephropathy cohort to identify variant-overlapping probes from their corresponding microarray expression levels. Using all variants, we found 27,767 affected probes in the Affymetrix HuGene 2.1 ST array, corresponding to 13,219 genes. At the probe level, >99% of strong associations (p-value < 10^{-3}) between affected probe and corresponding variant were negative in effect size. At the probeset (gene) level, we found a 2.0-3.5x odds ratio of having negative effect sizes compared to RNA sequencing expression data as a baseline. We then corrected expression levels using three approaches: (1) by removing affected probes identified by WGS, (2) removing affected probes

identified by 1000 Genomes reference panel, and (3) by adjusting affected probe levels using unaffected probes as a baseline. We compared these expression datasets with uncorrected expression in a comprehensive eQTL scan. Before correction, effect size balance was skewed in a negative direction for two tissues tested – (53.8% and 54.7%). This was no longer the case after applying correction methods, with negative ratios ranging from 45%-50% after correction. However, probe removal using 1000G all variants showed a large reduction in power and is inadvisable. We found that probe removal using Pima common variants and probe adjustment using Pima variants performed consistently well in terms of effect size imbalance resolution, as well as identifying likely false positives and false negatives. When whole genome sequence data are not available, removing probes using common variants from reference panels such as 1000G can be a reasonable approach to correct for hybridization bias.

3.2 Introduction

Expression Quantitative Trait Loci (eQTL) studies have identified a tremendous number of cis- and trans- associations between genetic variants and gene expression levels and have provided many biological insights into the regulatory aspect of complex traits [146]. These studies have become increasingly viable in the past twenty years due to advances in technologies that facilitate high throughput measurements of gene expression. Microarrays were one of the first technologies that allowed researchers to assay expression levels of genes in a massively parallel manner, allowing for large-scale analyses of the transcriptome. Microarray platforms typically measure transcript abundance by using synthetic complementary DNA (cDNA) probes to bind with RNA molecules, which are quantified via color-coded dyes [60]. There are then various post-processing statistical methods which are used to generate gene-level expression estimates from these color intensities. Unfortunately, there are limitations with this technology that arise from the fact that cDNA probe sequences must be known a priori when conducting experiments. One well known shortcoming is that sequence variation within the probe regions can result in weakened hybridization between said probe and corresponding RNA molecules (due to base-pair mismatches) [100,101]. This can artificially reduce the estimated expression levels, creating statistically significant, but technically false associations between genetic

variant and gene expression primarily in the direction of apparent downregulation of non-reference alleles. Even though these biases typically affect a small proportion of microarray probes, it has been shown that a large proportion of eQTL signals can be affected, leading to many potential false positives and false negatives [102].

Because of this limitation (as well as others), microarrays have given way to next generation technologies such as RNA sequencing, which overcome many shortcomings by sequencing RNA transcripts directly. However, microarrays still have been used in recent eQTL studies [6] [7], and can also arise in meta-analyses or aggregate cohort studies that examine older datasets in which microarrays were the cutting edge technology at the time [8]. As such, there is merit to addressing limitations from this aging platform in the modern day.

Recent publications have sought to address the hybridization bias by identifying problem probes and removing them from the analysis. Dannemann and colleagues implemented the algorithm 'maskBAD' which uses a statistical model to estimate differential binding affinity of probes between two experimental groups, where one group has the SNP and the other does not [104]. While the approach was able to identify most problem probes, some probes overlapping with SNPs could potentially be missed, and probes that did not overlap with SNPs were removed. Furthermore, the two groups (with one not affected by differential binding) must be identified prior to the analysis, which is not always possible in an experimental setting. Quigley's equalizer algorithm detects all probes that overlap with SNPs given from a Variant Call Format (VCF) file, allowing for exact identification and removal of affected probes [103]. The author mentions that many human eQTL studies are conducted without exome or whole genome sequences of the individuals. To demonstrate the probe removal algorithm, common polymorphisms obtained from European and African sequences in the 1000 Genomes Project were used to identify these probes. When reviewing the overall imbalance effect of eQTLs, the author noted that the downstream effect of removing these probes led to a much improved but still incomplete resolution of the negative hybridization bias. Furthermore, if the study population of interest is genetically distant from both the European and African reference genomes, 1000 Genomes polymorphisms may not accurately identify biased probes. Knowing the exact genomes corresponding to the study cohort could improve expression correction

methods by identifying all the correct probes when compared to using reference panels such as 1000G. With next generation sequencing, the fine-resolution genotyping of individuals has become more accessible to many study cohorts, which presents the opportunity to re-visit the issue of negative hybridization with relevant genotype information.

In this chapter, we examine the benefits of using whole genome sequencing to identify biased probes more accurately. We demonstrate that this could improve estimation of expression levels, which in turn empower and refine existing eQTL studies. To highlight the improvement, we apply this approach to a population-specific cohort of Pima Native Americans, whose genomes differ from known reference panels for many sites. With whole genome sequence data, we characterize the extent of the negative hybridization bias at both the probe level and probeset (gene) level. We further demonstrate that using the population-specific genetic loci to identify and remove biased probes can result in more accurate eQTL discovery and fewer false positives, compared to using the 1000 Genomes reference panel. However, we also found that removing all affected probes might negatively affect eQTL power and accuracy of measured expression levels. Therefore, we also present a probe-adjustment method where, instead of removing all affected probes, we impute their intensity levels using information from non-affected probes within the same probeset. We demonstrate that this approach may – in some situations – resolve the negative-hybridization bias without paying the price of losing power from removing probes.

3.3 Materials and Methods

3.3.1 Data Source

To empirically demonstrate the different bias-correction strategies, we use data from a Pima Native American diabetic nephropathy cohort. In this study, deep whole genome sequencing was performed on 97 individuals, and renal expression data were obtained from microdissected biopsies of glomerular and tubular tissue compartments within the kidney, using both microarray and RNA sequencing techniques on the same tissue samples. The microarray

platform used was the Affymetrix HuGene 2.1 array, which consisted of 25-mer probe sequences specifically designed to target individual exons. To convert probe intensities into probeset-level expression estimates, we used a custom (customCDF) probe-to-probeset mapping provided by the Microarray Lab from the Molecular and Behavioral Neuroscience Institute at the University of Michigan [148], which maps probes to genes more accurately than the default method [149]. Under this mapping the HuGene 2.1 platform contained 25,583 probesets over 466,204 probes.

We also compared the microarray data with RNA sequencing data, which was performed on the same tissues and overlapping samples when we need evaluations with different technologies. RNA-seq reads were aligned with TopHat [150] software tool and the transcript counts were quantified with Cufflinks [151] and normalized via log-transformation of FPKM (fragments per kilobase of transcript per million mapped reads).

Deep whole genome sequencing was done for all study individuals via HiSeq X at the Macrogen Lab. SNPs were detected using the GotCloud SNP caller [152] and SNPs/indels were detected using HaplotypeCaller [153]. Affymetrix SNP arrays were used to check for genotype concordance with our whole genome sequence variants, and concordance was found to be very high (>99.9% across all samples). For 1000 Genomes variants, we used variant sites and allele frequencies from the phase 3 release [14].

3.3.2 Identification and removal of Probes overlapping with variants

To identify probes overlapping with variants, we first downloaded a Browser Extensible Data (BED) file from the Affymetrix website (http://www.affymetrix.com/support/technical/byproduct.affx?product=HuGene-1_1-st-v1) containing information on the start and end positions of each probe. Conveniently, the genomic coordinates in this BED file were based on the Genome Reference Consortium human genome build 37 [154], which matched with the sequence alignment for the Pima cohort. We developed an in-house software which uses BED and VCF files as input, and outputs the full list of probes that contain any sequence variation. We defined a variant-overlapping probe as any probe whose start and stop positions contained a genetic variant. In the case of SNPs, this would

mean the genomic coordinates of the SNP lies between the start and stop position of the probe. We also included insertions and deletions (indels) in the output, using the criteria that the indel must be completely contained within the boundaries of the probe.

We repeated this using different VCF files, obtaining several lists of variant-overlapping probes. Here, we used the Pima VCF obtained from the deep whole-genome sequencing, and a VCF file obtained from the 1000 Genomes project. Because of the diverse populations within the 1000 Genomes cohort, there were many genetic variants and thus for too many probes were identified as problem probes. To remedy this issue, we filtered out variants with minor allele frequency lower than 5%.

With the list of affected probes identified, our in-house software then directly modifies the cel definition file (CDF), removing the affected probes from their corresponding probesets. Using the new corrected CDFs, we then created an R-package compatible with Bioconductor's oligo library, which is used to calculate gene expression.

3.3.3 Probe adjustment approach

We also implemented a probe adjustment approach where instead of removing affected probes, we re-calculated their values based on other probes within the probeset. Suppose a probeset P has k probes, $p_1, p_2, \ldots, p_k \in P$, and that probe p_j overlaps with a genetic marker(s) with genotypes denoted as x_j (in the case of multiple, we pick one since they are almost always in high or perfect LD with each other). Let the $y_1, y_2, \ldots, y_k$ be vectors to denote the probe intensities for all samples, corresponding to each probe. For genetic variant x_j, we use a linear regression model to estimate a regression coefficient for each probe within the probeset:

$$\log(y_i) = \hat{\beta}_{0i} + \hat{\beta}_{1i} x_j + \hat{\varepsilon}_i, \qquad i \in 1, \ldots, k$$

We then average the coefficients for all the probes not overlapping the variant, that is:

$$\hat{\beta}_{adj0} = \frac{\sum_{i \neq j} \hat{\beta}_{0i}}{k-1}$$

$$\hat{\beta}_{adj1} = \frac{\sum_{i \neq j} \hat{\beta}_{1i}}{k-1}$$

The estimated expression is as follows:

$$\log(y_i) = \hat{\beta}_{adj} \ + \hat{\beta}_{adj1} x_j + \hat{\varepsilon}_j$$

This approach uses the non-problem probes to estimate the "true" effect the genetic marker has on expression levels, averaging across all unaffected probes. Here, the residuals $\hat{\varepsilon}_j$ are from the regression between the variant and its corresponding affected probe. These residuals provide an estimate of the individual-level probe intensities correcting for the technical effect induced by the overlapping variant.

3.3.4 Normalization of Expression Data

We calculated gene expression using the Bioconductor's oligo R package [155], with sample-level CEL files containing probe intensity information as input and using the aforementioned custom CDF to map the probes to probesets. We then applied Bioconductor's Robust Multi-Array Average (RMA) normalization algorithm to the probe level expression [67,156]. This was applied to each tissue and biopsy separately. Finally, we applied an inverse-normalization transformation across subjects for each probeset. Probesets were mapped to genes using Ensembl GRCh37 genes [154].

3.3.5 Quantification of probe- and probeset-level biases

The overall bias quantification of identified probes was done by comparing the estimated enrichment of probes with both positive and negative hybridizations. To do this, we performed a regression analysis between every variant-in-probe (VIP: a genetic variant that overlaps with a probe) and the probe intensity level for its corresponding affected probe (VIP probe). We then repeated this for all other probes within the same probeset (probes that do not overlap with the variant). We used these unaffected probes to establish a baseline for comparison with the VIP probes.

In addition to our probe-level regression analysis, we performed a regression analysis to quantify the negative hybridization effect at the gene (probeset) level. Here, we included the 13,219 genes that were affected by at least one VIP. We regressed every VIP against the gene expression level of its corresponding affected gene. To serve as a baseline for comparison, we used two approaches. First, we leveraged the RNA sequencing data that was also assayed from the same tissue sample as the microarray experiment. Since the RNA sequencing platform should not be susceptible to the hybridization effect, this allowed us to establish a relatively accurate estimate of the true gene-level enrichment of negative effect sizes. Secondly, for each affected gene (probeset containing an affected probe), we performed association analysis with all variants within the exon region of that gene. In this analysis, we excluded VIPs corresponding to each gene. This second approach allowed us to perform a high number of tests, while also allowing us to examine the effects of linkage-disequilibrium with VIPs on the hybridization bias.

3.3.6 cis-eQTL Analysis

Expression quantitative trait locus (eQTL) analysis was performed using mixed model association via the EMMAX software package [157]. A separate analysis was performed for each tissue and each expression correction method (uncorrected, correction using probe removal via Pima WGS and 1000G, and correction using probe imputation via Pima WGS). For every SNP and indel identified from our whole genome sequencing, we tested for association against each of the genes with measured expression. To account for potential confounders, we adjusted for age and sex as covariates. In addition, we calculated pairwise-kinship coefficients for all samples, using it as the fixed-effects component of the mixed model.

We defined an eQTL as cis-acting if it was located within 1 Mb of the transcription start site of the associated gene. Otherwise, the eQTL signal was defined as being trans-acting. To account for multiple testing, p-values were adjusted by the false-discovery rate (FDR) correction approach, using the trans-eQTL signals to determine the false discovery rates. The unadjusted p-values corresponding to an FDR of 0.05 was approximately $\alpha=10^{-5}$ for all our analyses, and thus we used this threshold as our significance cutoff. Because of linkage disequilibrium (LD)

between neighboring SNPs, only the SNP with the lowest p-value was taken as the true cis-eQTL signal for each gene.

3.3.7 Identification of technical false positives and technical false negatives

Using our new corrected eQTL analyses, we revisited the uncorrected analyses and determine which of its signals were potentially false positives and false negatives due to the hybridization bias. In essence, we sought to identify signals that were lost (or gained) due to expression correction and determine if they were lost (or gained) because of VIPs. To do this, we performed pairwise comparisons between all gene signals from our uncorrected and corrected analyses, examining the effect sizes of peak eQTL variants, correlation between peak variants and VIP variants, and the association between VIP variant and gene (and probe) expression levels.

For each gene, we defined it to be a false positive using the following criteria:

- Uncorrected expression p-value < $1x10^{-5}$ and corrected expression p-value > $1x10^{-5}$ **AND;**
- Peak eQTL is in high-LD with VIP ($r^2 > 0.1$) **AND;**
    - Peak eQTL effect size diminishes after correction **OR;**
    - VIP is negatively associated with affected probe (p-value < 0.1) **OR;**
    - VIP is negatively associated with affected gene expression

We defined false negatives using a similar criteria:

- Uncorrected expression p-value > $1x10^{-5}$ and corrected expression p-value < $1x10^{-5}$ **AND;**
- Peak eQTL is in high-LD with VIP ($r^2 > 0.1$) **AND;**
    - Peak eQTL effect size increases after correction **OR;**
    - VIP is negatively associated with affected probe (p-value < 0.1) **OR;**
    - VIP is negatively associated with affected gene expression

Finally, we excluded all genes in which the genotypes of the VIP had minor allele frequency less than 5%.

3.4 Results

3.4.1 Comprehensive scan of VIPs using deep whole genome sequencing (WGS).

We comprehensively scanned for variant-in-probes (VIPs: any SNP or indel contained within a probe) by comparing deep whole genome sequence (WGS) data to the genomic coordinates of every microarray probe. We examined 55 Pima Native American individuals (50 with glomerular and 54 with tubular expression data) where both deep WGS and array-based transcriptomic profiles (based on the Affymetrix HuGene 2.1.ST platform) were available (ClinicalTrials.gov number, NCT00340678). Among the 8.8 million SNPs and 1.8 million indels identified across these Pima genomes, we found that 29,917 (0.0028%) variants overlapped with a probe (VIPs). Using Pima VCFs, we observed that 6% of all probes (27,767/466,204) were affected by a VIP (Table 3.1A). However, despite the relatively low percentage of affected probes, because each probeset contains 18.2 probes on average, more than half (51.1%) of the probesets (13,219 out of 25,583) contained at least one affected probe, suggesting that the majority of array-based expression levels can potentially be affected with a bias to some degree. Per individual sample, we flagged an average of 10,219 affected probes (2.2% of all probes), which in turn mapped to 6,210 affected probesets (24.3% of all probesets) on average.

Using all 1000 Genomes variants to identify VIP probes is overly sensitive.

We repeated this evaluation using ~80 million variants from the 1000 Genomes Project (1000G) [14]. As mentioned by Quigley, this approach could potentially be a viable alternative when individual-level WGS are not available, and it could identify affected probes with a reasonable accuracy when the study cohort is represented by the ancestries included in this reference panel. Using 1000G variants, we identified 51.1% of all probes as affected (238,207/466,204), which mapped to 93.1% of all the probesets (23,827/25,583). We observe that among these 238,207 flagged probes, 212,455 were not flagged when using Pima WGS variants and are unlikely to have hybridization bias since they do not overlap with any Pima polymorphisms (Table 3.1B). We found that attempting to unnecessarily account for this high volume of probes led to inaccuracy in calculating expression for many genes, without providing benefit towards bias reduction.

Using common variants only achieves more reasonable sensitivity to identify VIP probes

A much more sensible approach was to limit our scan to common variants (minor allele frequency > 5%), which may have a higher impact on probes compared to rare variants. For the sake of completeness, we applied this approach to both Pima and 1000G, creating two additional list of probes that were flagged based on common variants only. For Pima, the fraction of flagged probes and corresponding probesets were reduced to 3.5% (16,423 probes) and 36.0% (9,202 probesets) respectively. For 1000 Genomes common variants, we flagged 4.6% of all probes (21,537 probes), mapping to 49.2% of all probesets (12,592 probesets). From We see that the 1000 Genomes common variant approach only flags 3,384 probes not identified by the Pima WGS (all variants) method, which is substantially fewer than the 212,455 extra probes flagged when using all variants (Table 3.1C). Correcting for these 3,384 extra probes will not likely affect the expression estimates to the same magnitude, and thus this method appears to be a much more reliable approach. For the rest of the comparison in this chapter, we focused primarily on the probes identified by Pima WGS all variants, and 1000G common variants.

3.4.2 Quantifying the effects of negative hybridization using probes identified by Pima WGS

The availability of Pima WGS data allows us to characterize the hybridization bias at a probe level. To do this, we performed a linear regression between every affected probe and the genotypes of their corresponding VIP(s). To serve as controls, we also regressed these (VIPs) genotypes against the unaffected probes within the same probeset of the affected probe.

Overall, we observe that a large proportion of probes are biased due to VIPs, and that the bias becomes more apparent as the regression effect size increases (Figure 3.1, Supplementary Table 3.1). For example, when examining all VIP probes ($\alpha = 1$), 62.9% and 64.0% have negative effect sizes for glomerular and tubular tissues, respectively (Figure 3.1A). Given an expected ratio of 50% (which is observed in the unaffected probes analysis), there are roughly 3,400 more negative probes than expected (out of a total of 26,435 total probes). Starting from $\alpha = 10^{-3}$ and onward, we observe >99% of effect sizes being in the negative direction. In contrast,

our unaffected probe analysis shows 45-60% negative effect sizes across the spectrum of p-value thresholds. This high imbalance of negative effect sizes provides evidence that strong associations between VIPs and their corresponding probes are almost all due to artificial artifacts. We also see an absence of strong positive signals, with only 1 positive signal at $\alpha = 10^{-5}$ for glomerular and 2 positive signals for tubular. This suggests that true associations with positive effect sizes may also be masked by this negative hybridization bias. This could potentially lead to a loss of power and false positives in downstream eQTL discovery.

We compared the effect size densities for affected and unaffected probes (Figure 3.1B). From here, we confirm the enrichment of negative effect sizes (from the heavy tail), as well as a mild shift in the median (median: -0.392 for glomerular, -0.433 for tubular) in the negative direction. The heavy left tail for highly significant associations (especially when t-statistic is less than -2) suggests that there may be a small number of probes that are greatly affected by the VIPs. The shift in the overall median for affected probes indicates that there are many probes that are mildly skewed in the negative direction – that is to say, most probes appear to have a small negative hybridization effect.

Finally, we found that the position in which the variant overlaps with the probe also affects the severity of hybridization reduction. To test this, we calculated the distance of each VIP variant from the center of the affected probe sequence and tested for association with effect size. We found that variants closer to the center of the affected probe had stronger negative effect sizes, while variants near the edge of the affected probe boundaries had milder negative affinity (p-value $< 10^{-15}$).

3.4.3 Quantifying the effects of negative hybridization at the gene level

Despite VIPs creating a strong negative bias for many of their corresponding probes, the impact of VIPs on gene expression at the probeset level is far more modest. Much like our probe-level analysis however, we still observe an enrichment of negative effect sizes which increases as the p-value threshold becomes more stringent. Upon examining all regressions (α=1) for the uncorrected expression, we see a slight imbalance of effect sizes across 18,109 tests (19,249 for tubular), with a negative ratio of 53.1% and 54.0% (confidence intervals) in glomerular and

tubular tissues respectively (Figure 3.2A, Supplementary Table 3.2). With more significant p-value thresholds, the negative ratio increases to roughly 70%. This relatively milder imbalance of negative signals compared to our probe regressions (70% versus >99%) can be explained with the knowledge that each probeset has an average of ~20 probes. Therefore, even though probes can be highly biased, the unaffected probes within the same probeset help mitigate the bias when calculating expression levels.

Because RNA sequencing is less susceptible to expression bias arising from sequence variation, we repeated our experiment using RNA sequencing expression data for the same set of genes as a baseline for comparison. Compared to RNA-seq, there is evidence of the microarray hybridization bias both enriching negative signals, while also suppressing positive direction signals (Figure 3.3). There appears to be a clear abundance of negative effect sizes within the microarray associations, whereas this does not appear from the RNA-seq analysis (Figure 3.3A). We calculated the odds-ratio of having a negative effect size using RNA-seq as a baseline – we divided the microarray relative-risk by the RNA-seq relative-risk (Figure 3.3B). For the glomerular tissue, we observe an OR of ~2 – 3.5, while the tubular tissue shows an OR of ~2.

We found that even non-VIP variants can be affected with hybridization bias through LD with VIPs. Our non-VIP regression analysis (Figure 3.2B) serves to characterize the relationship between affected/unaffected probesets and genetic variants within the corresponding exon regions of the gene. We separated this analysis into two components: 1) exon associations with genes containing an affected probe, and 2) exon associations with genes not containing an affected probe. Despite VIPs being excluded from this analysis entirely, we still observe a greater negative enrichment for affected genes versus non-affected genes. For example, at α=10-3, the glomerular tissue shows a 62.3% negative ratio for affected genes and a 54.5% negative ratio for unaffected genes. Similarly for the tubular tissue, there are 60.1% and 46.8% negative ratios for affected and unaffected genes respectively. This provides evidence that cis-variants that are not VIPs are still negatively skewed, and could potentially be technical false positives. This is likely due to the linkage-disequilibrium structure, where a VIP will cause other variants in high LD to also be associated with the gene expression levels.

3.4.4 Assessment of bias correction methods

Using our expression-correction methods, we generated ten new sets of expression data (5 sets for each tissue). These new datasets were derived from five correction methods: (1) Probe removal using Pima WGS (all variants) to identify probes, (2) Probe removal using Pima WGS (common variants) to identify probes, (3) Probe removal using 1000G (all variants) to identify probes, (4) Probe removal using 1000G (common variants) to identify probes and (5) Probe adjustment (see section 3.3) using Pima WGS (all variants) to identify probes.

We repeated the VIP-regression analysis performed in 3.4.3, where we calculated association effect sizes between every VIP and their corresponding affected (but now corrected) gene. Overall, the methods all appear to improve the effect size balance, with negative ratios close to 50% across all p-value thresholds (Figure 3.4, Supplementary Table 3.3). One exception would be the 1000G (all variants) probe removal method, which appears to have a higher negative ratio in the glomerular tissue. Based on effect size directions alone, we believe that all the correction methods other than 1000G (all variants) probe removal are viable. This is not surprising as the 1000G (common variants) method identifies many of the same probes that the Pima WGS method, and thus most of the expression estimates will be quite similar.

3.4.5 Impact of hybridization bias on eQTL analysis

Finally, we compared our expression correction approaches by performing a comprehensive eQTL scan across the entire genome. In contrast to our previous methods that tested only VIPs or non-VIP variants within exonic regions, our eQTL analysis tested every variant regardless of classification. Because VIPs are only a small portion of the entire genome, it is very unlikely that VIPs will be the peak eQTL for most genes. However, it is possible for the peak signals to be in linkage-disequilibrium with VIPs. As such, we broke down our results into further categories based on magnitude of LD between the eQTL peak and any VIPs that may affected the gene.

We observe that peak eQTLs from both tissues are biased in the negative direction, and that all methods can correct this (Figure 3.5, Supplementary Table 3.4). However, there may be a slight overcorrection as the effect sizes in general skew slightly in the positive direction. Overall, the correction methods based on common variants appear to perform better than those using all

variants. In the Glomerular tissue, both Pima and 1000G common variants probe removal have better balance in effect size directions compared to their counterparts, while also identifying more signals. The probe adjustment approach also performs quite well, particularly in the tubular tissue where we observe a 50% negative ratio as well as the highest number of significant signals. However, the 1000G common variants approach appears to overcorrect in the tubular tissue, leading to a 54.5% positive effect size ratio. The 1000G all variants method reduces power substantially, resulting in a 21.9% and 35.7% loss of significant eQTLs in the glomerular and tubular tissues respectively.

We examined the pairwise t-statistics between corrected and uncorrected genes (Supplementary Figure 3.1). Here, we observe a correlation of 0.88, 0.93 and 0.68 for the Pima probe removal, Pima probe adjustment and 1000G probe removal methods, respectively. This provides evidence that removing probes generates noise for gene expression estimates, which is particularly noticeable in the 1000G method. This in turn dilutes the effect sizes in eQTL analysis which reduces the number of significant signals detected.

3.4.6 Evaluation of technical false positives and false negative eQTLs.

We compared the list of signals lost and gained between the uncorrected and corrected analyses and identified false positives and false negatives using the criteria described in our methods section (3.3.7). For example, the *RPL9* gene was classified as a technical false-positive by both the probe removal and probe adjustment methods (Supplementary Figure 3.2). Here, we observe a strong correlation between uncorrected expression and VIP genotypes (t-statistic = -6.899) whereas the correlation between corrected expression and VIP genotypes is near zero. In this scenario, the peak eQTL was the VIP, with an association p-value of 2.94×10^{-8}. After expression correction, the peak eQTL was no longer the VIP and had a p-value of 5.37x10-^{4}.

However, not all signals gained or lost fulfilled the criteria of false positives/negatives. For example, the Pima probe removal method resulted in 66 and 85 signals lost for glomerular and tubular tissues, respectively (Supplementary Table 3.5). However, only 19 (28.8%) and 27 (31.7%) of those lost signals were classified as false positives. This indicates that many of the signals that were lost may have fallen beneath the significance threshold due to noise

generated in expression levels due to the removal of probes, rather than being true technical biases. We found that overall, using common variants (in both Pima and 1000G) performed better – with higher ratios of false positives/negatives among the list of genes lost/gained – supporting the notion that removing fewer probes is advantageous. Finally, we observe that the probe adjustment method has the highest concordance with the uncorrected results (fewest signals gained/lost), but also accounts for the highest proportion of false positives/negatives (relative to number of signals gained/lost) in both tissues. The 1000G all variants method on the other hand, flags the most potential false positives (20 in glomerular, 41 in tubular) but does so at the expense of too many lost signals (94 for glomerular, 137 for tubular).

3.5 Discussion

In this chapter we characterized the effects of weakened hybridization in microarrays at a probe and probeset level, while exploring the ramifications of this technical bias on downstream eQTL analysis. Although methods to account for this bias have been examined in previous publications, none to our knowledge have leveraged whole genome sequence data to refine and empower these studies. In our work we first used whole genome sequencing to accurately identify probes with weakened hybridization. Once we identified these probes, we were able to characterize the extent of the bias at both the probe and probeset level by directly calculating regression coefficients between the variant-in-probes (VIPs) and expression levels of their corresponding probes/probesets. Finally, we performed downstream eQTL analysis, examining the effect size direction of peak eQTLs as well as their correlation with VIPs.

Throughout our analyses, we compared correction methods that removed potentially biased probes from the calculation of the gene expression levels. Our hypothesis was that more accurate identification of probes would lead to improved analysis in terms of power, effect size distribution, and reduction of technical false positives and technical false negatives. We found in our analysis that overall, all methods – aside from probe removal via 1000G all variants – performed quite well. In terms of direction of effect size, these methods produced overall negative ratios close to 50%, which indicates that at the very least, the negative enrichment

from weakened hybridization has been resolved. In terms of power to discover eQTLs, we found that using common variants to remove probes performed better than using all variants. This is not surprising, as common variants are more likely to have an impact and negatively bias probe expression levels compared to rare variants. Furthermore, we found that removing probes introduced noise into the expression level estimates, which would then dilute the signals in an eQTL scan. Thus, limiting the list of probes removed to the most impactful ones would provide beneficial to the quality of the analysis. Finally, we examined the list of signals that were gained or lost after correcting expression levels and determined which of these were due to technical biases. From this, we found that methods that more accurately identified probes had higher success in determining false positives/negatives, while remaining relatively faithful to the original (uncorrected) analysis otherwise. In particular, the probe adjustment and probe removal with Pima common variants appeared to perform best from this angle.

From our various viewpoints, we believe that probe adjustment via Pima all variants is the best approach to use, as it resolves many of the shortcomings of weakened hybridization while retaining high power and fidelity towards the original analysis. For methods that remove probes, identification of affected probes using common variants from whole genome sequence data appears to be the next best solution. In the scenario in which whole genome sequence data is not available, using common variants from a reference panel would perform adequately, although noise may be introduced into expression levels for genetically distant populations, causing both spurious signals as well as unnecessary loss of signals in eQTL analysis.

Despite this recommendation, there are several limitations to our study and as such these findings should be approached with caution. First, our study population was an isolated and relatively homogeneous population of Pima Native Americans, and it is uncertain if the results from this chapter can be applied to other populations, or populations with genetic admixture. Next, we tested these methods using the glomerular and tubulointerstitial compartments in the Kidney, and findings here may not generalize to other tissues. We already noticed a slight difference in performance between these two tissues, with probe removal via 1000G common variants seemingly overcorrect – with more positive effect sizes – for the tubular tissue, as well as observing better performance in the tubular tissue for the probe adjustment method. In

addition to different tissues, our results may not extend to microarray platforms different from the HuGene 2.1 ST platform that was used in this study. Finally, the gene expression levels obtained from RNA sequence data were generated from transcript reads that may differ from the array-based probe sequences. Thus, the RNA-seq genes used as our baseline for comparison potentially interrogated different regions within the same gene compared to the microarray expression levels. As such, the RNA sequence gene expression levels do not provide exact comparisons between biased and unbiased transcript sequences, but rather serve as a general comparison, outlining the overall impact of negative hybridization bias at the gene level.

To conclude, we revisited a well-known limitation of microarrays in assaying gene expression levels, using whole genome sequence data – as opposed to reference panel data – to characterize the list of potentially biased probes. We found that overall, WGS can more accurately identify these probes and ultimately provides higher quality eQTL analysis.

3.6 Tables and Figures

		N individuals	Total	# Overlap with VIPs	% Overlap with VIPs
Probe	Per Individual		466,204	10,219	2.2%
	Pima-all	97	466,204	27,767	6.0%
	Pima-common	97	466,204	16,423	3.5%
	1000G-all	2504	466,204	238,207	51.1%
	1000G-common	2504	466,204	21,537	4.6%
Probeset	Per Individual		25,583	6,210	24.3%
	Pima-all	97	25,583	13,219	51.1%
	Pima-common	97	25,583	9,202	36.0%
	1000G-all	2504	25,583	23,827	93.1%
	1000G-common	2504	25,583	12,592	49.2%

Table 3.1A Counts of the number of probes and probesets affected by VIPs

Information on the VIPs (SNPs and insertions/deletions) identified using the following VCFs: Pima WGS (all), Pima WGS (common), 1000G variants (all), 1000G variants (common). The per individual counts are the average number of affected probes/probesets per Pima individual.

1KG-all only	1KG-all & Pima-all	Pima-all Only
212,455	25,752	2,015
1KG common only	1KG common & Pima-all	Pima-all Only
3,384	18,153	9,614
1KG common only	1KG common & Pima common	Pima common Only
6,947	14,590	1,833

Table 3.1B – Comparison between lists of affected probes as identified by Pima and 1000G variants.

It is evident that most of the probes identified by Pima WGS are captured when using the 1000G-all variants approach, although this method identifies far too many extra probes as being affected (212455). Using 1000G-common variants, this method does not capture all of the probes identified by Pima WGS (missing 9614 probes) but also does not flag an excessive number of extra probes (only 3384)

1KG-all only	1KG-all & Pima-all	Pima-all Only
10,617	13,210	9
1KG common only	1KG common & Pima-all	Pima-all Only
1,131	9,981	3,238
1KG common only	1KG common & Pima common	Pima common Only
2,611	8,501	701

Table 3.1C – Comparison between lists of affected probesets as identified by Pima and 1000G variants.

Only 9 probesets from the Pima-all approach are not covered using the 1000G approach, although the 1000G approach likely removes the wrong probes within the correct probeset.

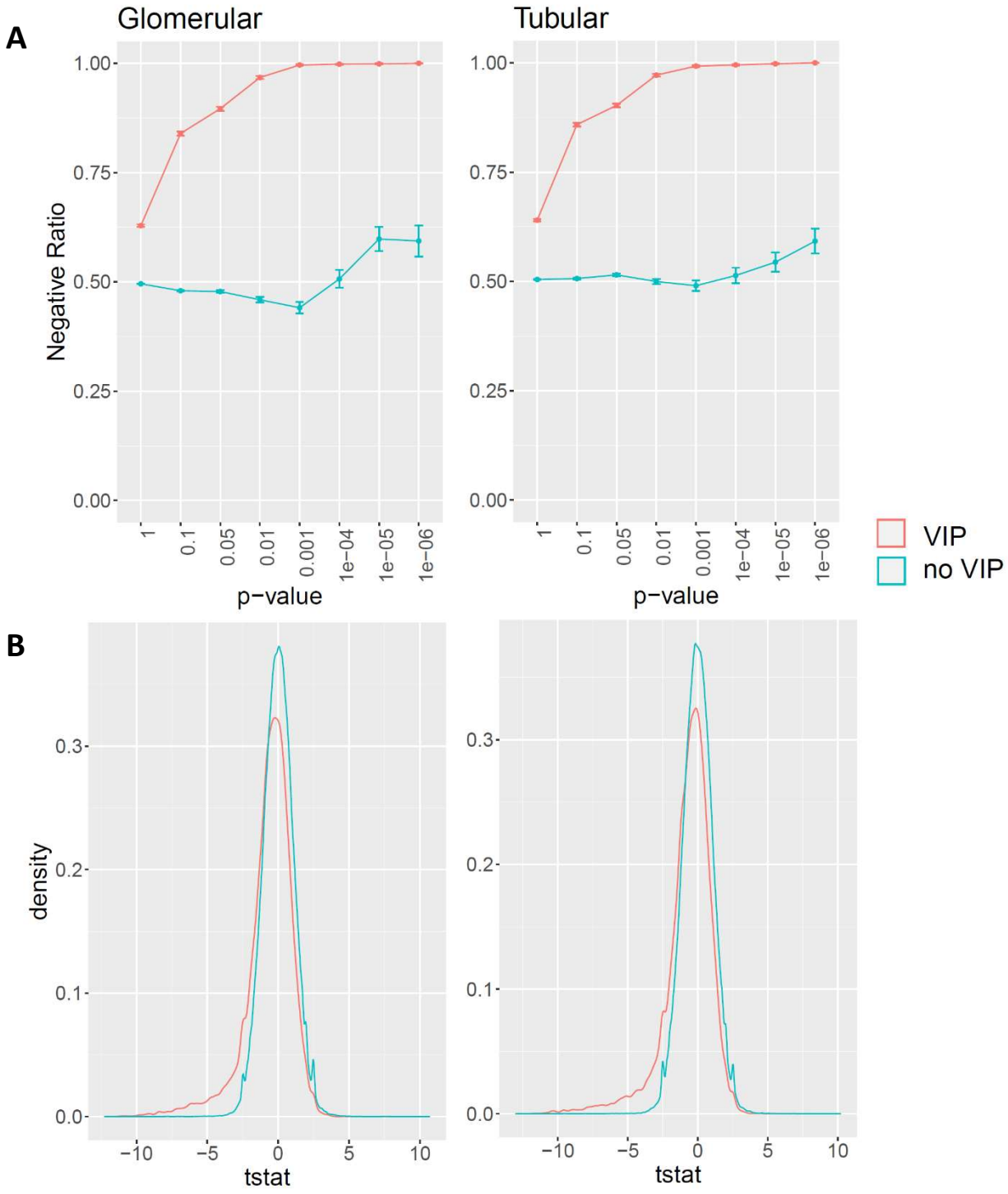


Figure 3.1 – regression between VIP and affected/unaffected probes

(A) shows the negative/positive ratio of effect sizes for the VIP regression against affected and unaffected probes identified using the Pima WGS approach. (B) shows the distribution of effect sizes for the same regression analysis. Medians are: -0.392/-0.433 for Glom/Tub for affected probes, and 0.01/-0.01 for Glom/Tub for unaffected probes.

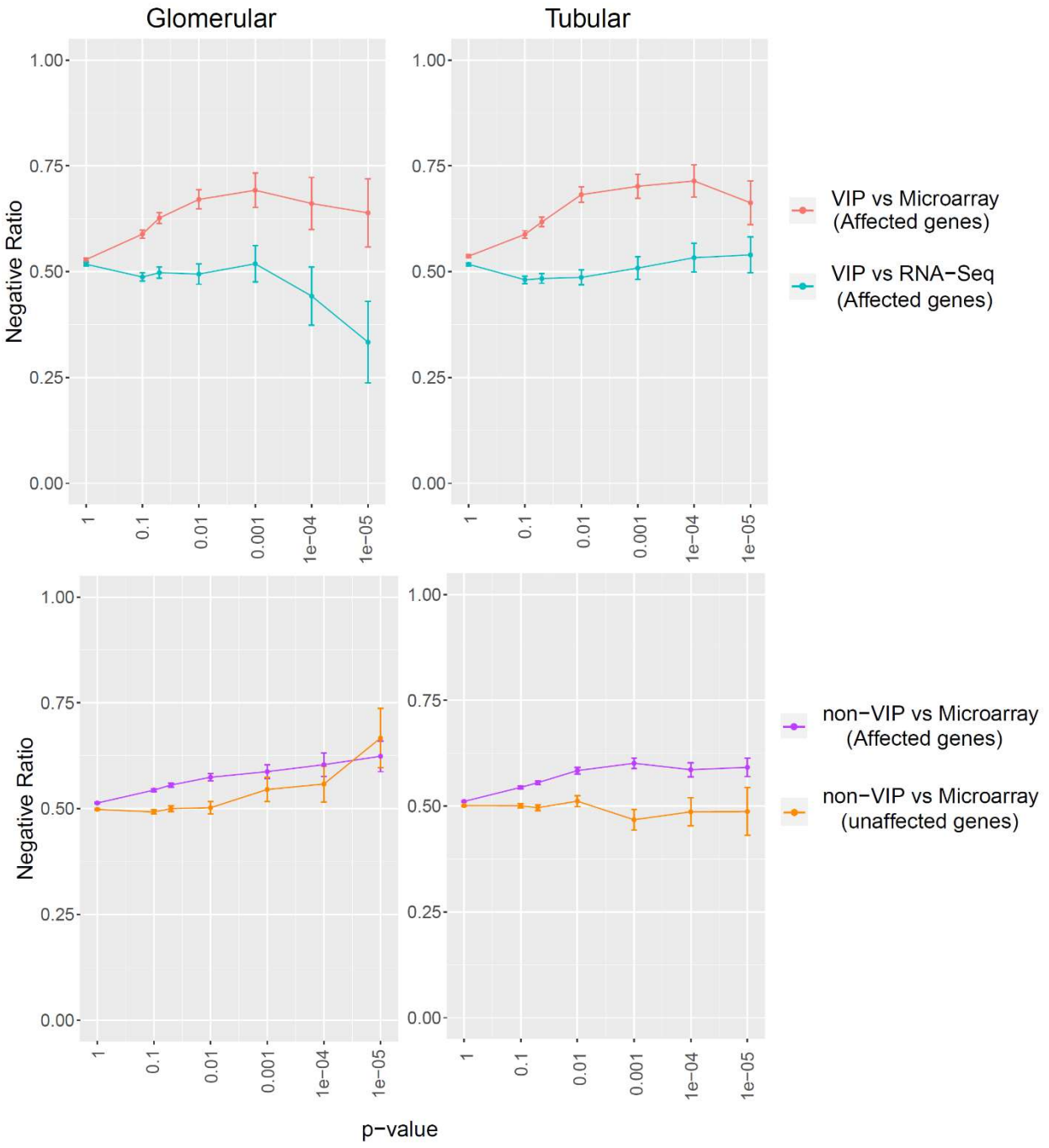


Figure 3.2 – *regression between VIPs and probesets*

(A) in our probeset analysis, we regressed every VIP against the probeset (gene) expression levels corresponding to the affected probe. 2A shows the negative ratio of effect sizes for microarray and RNA-Sequencing expression levels of affected genes. (B) For our exon analysis, we regressed all variants (excluding VIPs) within the exonic region of each gene with the microarray expression levels of that gene. 2B shows the negative ratio of effect sizes for hybridization affected and unaffected genes. Here, we observe an enrichment in negative effect sizes for affected genes and whereas the unaffected genes are overall better balanced (Glomerular tissue shows an imbalance here, but the number of signals is very low). This indicates a presence of LD between exonic variants and VIPs, leading to non-VIP variants being susceptible to technical biases.

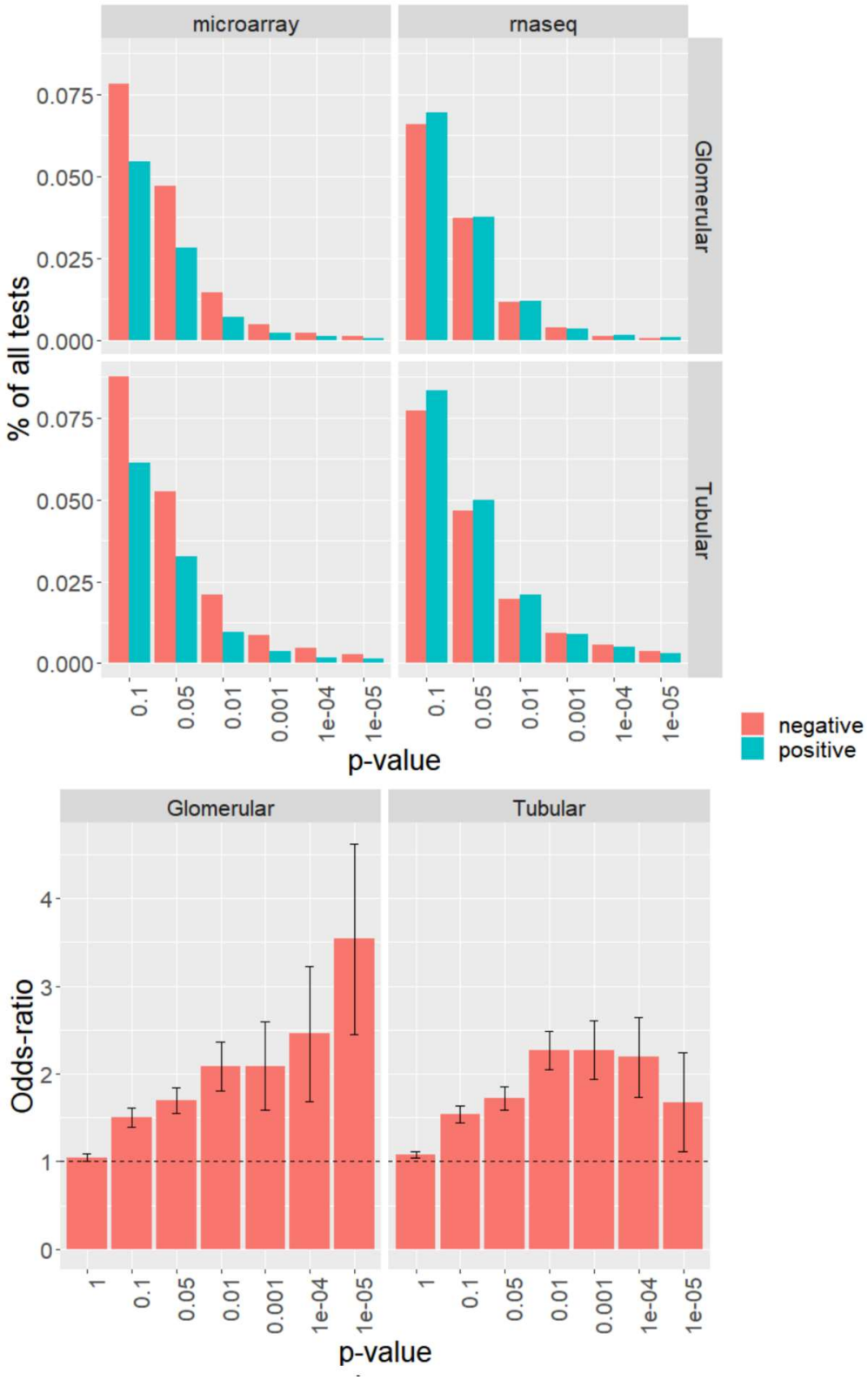


Figure 3.3 – *Comparison between Microarray and RNA-Seq effect sizes when performing regression between VIP and affected gene (probesets)*

(A) shows the counts of effect size directions for various p-value thresholds, with microarray counts on the left and RNA-Seq counts on the right. Overall, the microarray negative ratio ranged from 52% to 70%, whereas the RNA-Seq negative ratios ranged from 45%-55%. (B) shows the odds-ratio of having negative effect sizes for in Microarray genes relative to RNA-Seq genes. For the Glomerular tissue, we see a strong enrichment in negative effect sizes (up to 3.5 odds). In the Tubular tissue, there is milder enrichment in negative signals with an odds-ratio of 2.2.

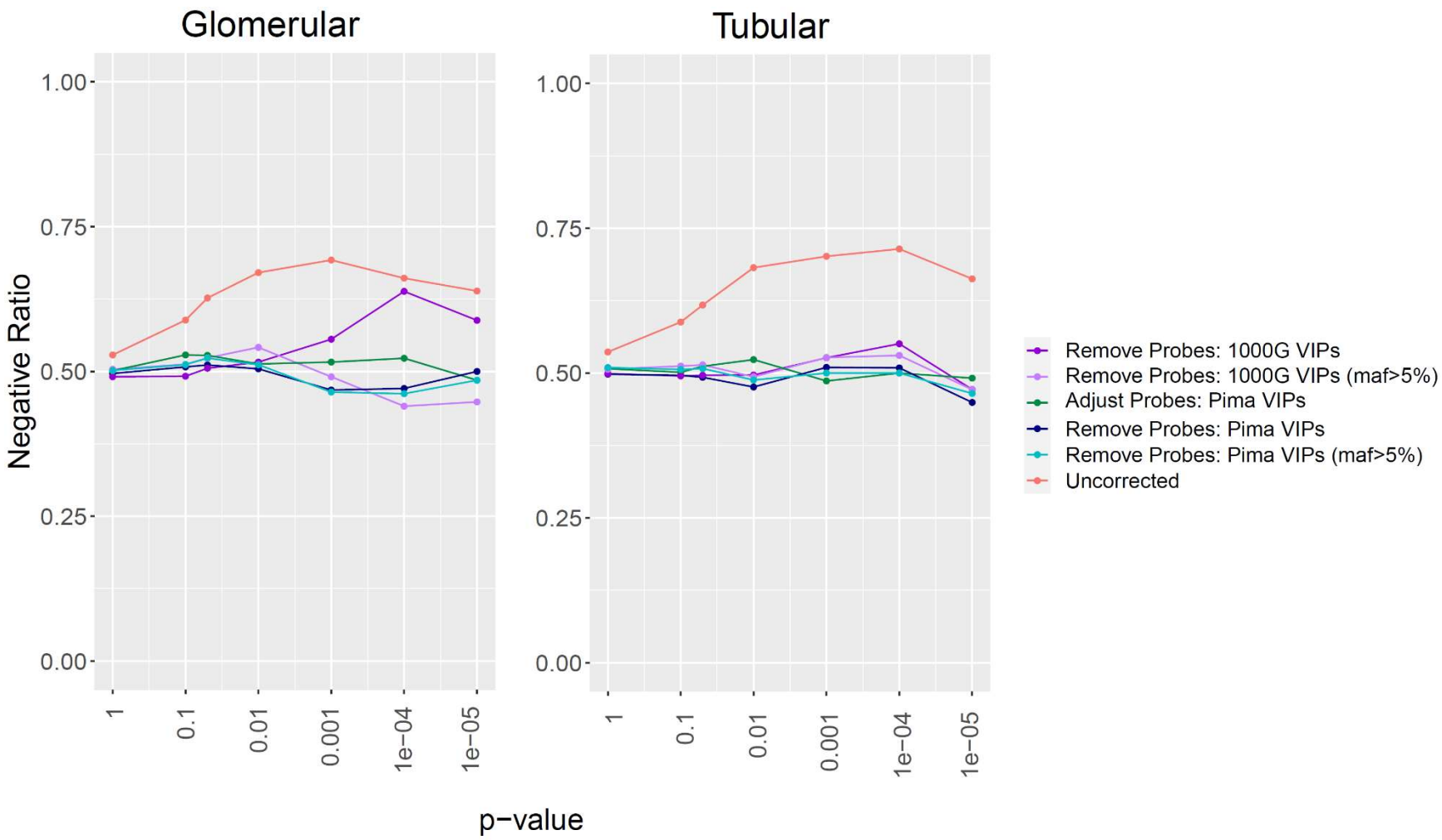


Figure 3.4 – Comparison of uncorrected and different corrected expression approaches at the probeset level.

Here, we regressed each VIP against their corresponding affected probeset. These plots show the negative ratio of effect sizes at various p-value thresholds. Overall, all correction methods appear to improve the effect size balance compared to the uncorrected approach, although the 1000G all variants probe removal method still displays a negative enrichment amongst signals with strong effect sizes within the Glomerular tissue.

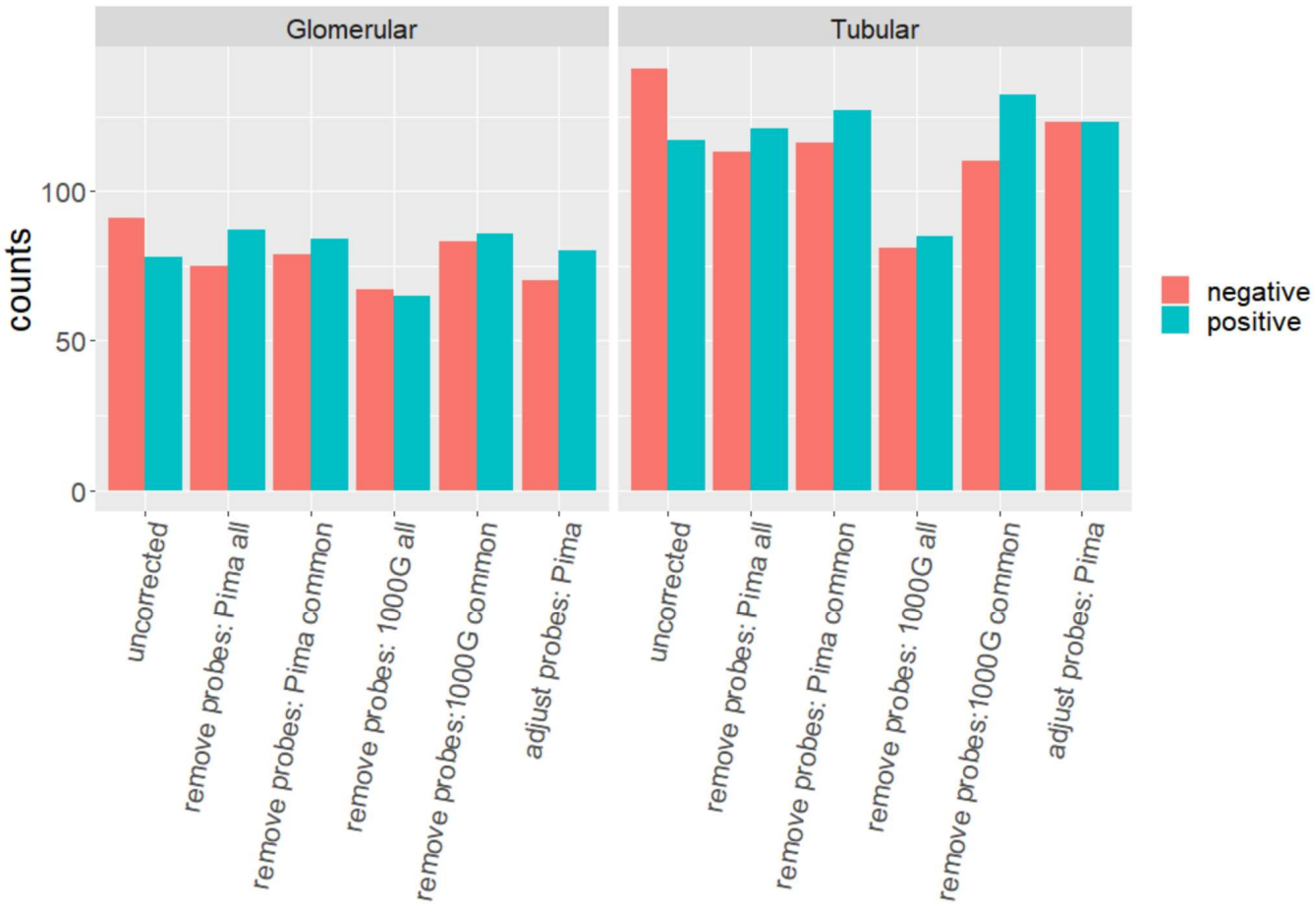


Figure 3.5 – Comparison of uncorrected and different corrected expression approaches in a full eQTL analysis.

Here, performed a comprehensive eQTL scan with the uncorrected expression, and each of our correction strategies. We noted the effect size direction for peak eQTL variants for each gene. Judging from the effect size balance, it appears that probe removal using 1000G common variants and Pima common variants performs well in the glomerular tissue, while the probe adjustment method performs best in the tubular tissue. As expected, using 1000G all variants to remove probes likely creates heterogeneity in expression levels by introducing white noise, which greatly reduces power in the eQTL analysis.

3.7 Supplementary Tables and Figures

Tissue	P-value Threshold	affected probes			unaffected probes		
		Positive	Negative	%Neg	Positive	Negative	%Neg
Glomerular	1	9,815	16,620	62.9	236,975	232,790	49.6
	0.05	430	3,711	89.6	14,534	13,308	47.8
	0.01	78	2,323	96.8	3,390	2,877	45.9
	1e-3	6	1,528	99.6	832	656	44.1
	1e-4	2	1,096	99.8	295	303	50.7
	1e-5	1	790	99.9	123	183	59.8
Tubular	1	9,893	17,601	64.0	241,988	246,237	50.4
	0.05	446	4,140	90.3	15,027	15,954	51.5
	0.01	76	2,623	97.2	3,615	3,612	50.0
	1e-3	13	1,718	99.2	913	878	49.0
	1e-4	6	1,259	99.5	396	418	51.4
	1e-5	2	955	99.8	233	278	54.4

Supplementary Table 3.1– Direction of regression effect sizes between VIPs and probe-level intensities.

For affected probes, we regressed the expression levels against the genotypes of the corresponding VIP. For unaffected probes, we regressed every probe within the affected probeset corresponding to the VIP.

Tissue	P-value	VIP vs Microarray (affected probesets)			VIP vs RNA Sequencing (affected genes)			Non-VIP vs Microarray (affected genes)			Non-VIP vs Microarray (unaffected genes)		
		Positive	Negative	Ratio	Positive	Negative	Ratio	Positive	Negative	Ratio	Positive	Negative	Ratio
Glomerular	1	9,177	10,284	0.528	8,982	9,636	0.518	73,001	76,929	0.513	37,775	37,486	0.498
	0.1	1,062	1,520	0.589	1,290	1,226	0.487	9,440	11,247	0.544	5,010	4,853	0.492
	0.05	546	917	0.627	701	694	0.497	4,875	6,096	0.556	2,601	2,600	0.500
	0.01	138	281	0.671	221	216	0.494	1,395	1,880	0.574	609	614	0.502
	1e-3	40	90	0.692	65	70	0.519	397	656	0.623	136	163	0.545
	1e-4	20	39	0.661	29	23	0.442	126	192	0.604	61	77	0.558
	1e-5	13	23	0.639	16	8	0.333	67	111	0.624	15	30	0.667
Tubular	1	9,584	11,087	0.536	9,641	10,335	0.517	75,981	79,436	0.511	38,918	39,103	0.501
	0.1	1,268	1,810	0.588	1,664	1,541	0.481	10,255	12,248	0.544	5,108	5,125	0.501
	0.05	673	1,086	0.617	996	933	0.484	5,705	7,128	0.555	2,784	2,743	0.496
	0.01	201	431	0.682	417	395	0.486	1,779	2,497	0.584	709	743	0.512
	1e-3	77	181	0.702	176	182	0.508	636	959	0.601	223	196	0.468
	1e-4	40	100	0.714	99	113	0.533	363	514	0.586	114	108	0.486
	1e-5	28	55	0.663	64	75	0.540	205	297	0.592	40	38	0.487

Supplementary Table 3.2– Direction of regression effect sizes between VIPs (and non-VIPs) and probeset (gene) expression.

Here, we performed regression analysis between: (1) every VIP vs. microarray gene expression levels for every affected gene, (2) every VIP vs. RNA-seq expression levels for each affected gene, (3) variants within the exome region (excluding VIPs) vs. microarray expression levels for every affected gene and (4) variants within the exome region vs. microarray expression levels for every non-affected gene

Tissue	P-value	Corrected (probe removal Pima)			Corrected (probe removal 1KG)			Corrected (probe removal 1KG MAF5)			Corrected (probe adjustments Pima)		
		Positive	Negative	Ratio	Positive	Negative	Ratio	Positive	Negative	Ratio	Positive	Negative	Ratio
Glomerular	1	9,789	9,451	0.491	9,907	9,549	0.491	9,651	9,796	0.504	9,019	9,064	0.501
	0.1	1,192	1,230	0.508	1,136	1,099	0.492	1,173	1,232	0.512	1,027	1,151	0.528
	0.05	657	687	0.511	592	605	0.505	644	707	0.523	579	647	0.528
	0.01	167	170	0.504	135	144	0.516	160	189	0.542	151	159	0.513
	1e-3	58	51	0.468	36	45	0.556	55	53	0.491	45	48	0.516
	1e-4	27	24	0.471	17	30	0.638	28	22	0.440	21	23	0.523
	1e-5	17	17	0.500	14	20	0.588	21	17	0.447	17	16	0.485
Tubular	1	10,386	10,304	0.498	10,362	10,321	0.499	10,204	10,477	0.507	9,481	9,774	0.508
	0.1	1,456	1,433	0.496	1,374	1,344	0.494	1,428	1,498	0.512	1,334	1,341	0.501
	0.05	797	774	0.493	730	719	0.496	774	819	0.514	714	748	0.512
	0.01	272	247	0.476	221	218	0.497	275	268	0.494	226	248	0.523
	1e-3	100	104	0.510	72	80	0.526	97	108	0.527	94	89	0.486
	1e-4	55	57	0.509	40	49	0.551	54	61	0.530	47	47	0.500
	1e-5	38	31	0.449	28	25	0.472	37	33	0.471	29	28	0.491

Supplementary Table 3.3– Regression between VIPs and probesets after various correction methods

This table shows the regression between every VIP vs. corrected microarray expression levels for each correction method. The first three columns correspond to the probe removal correction methods, while the rightmost column is our probe adjustment method.

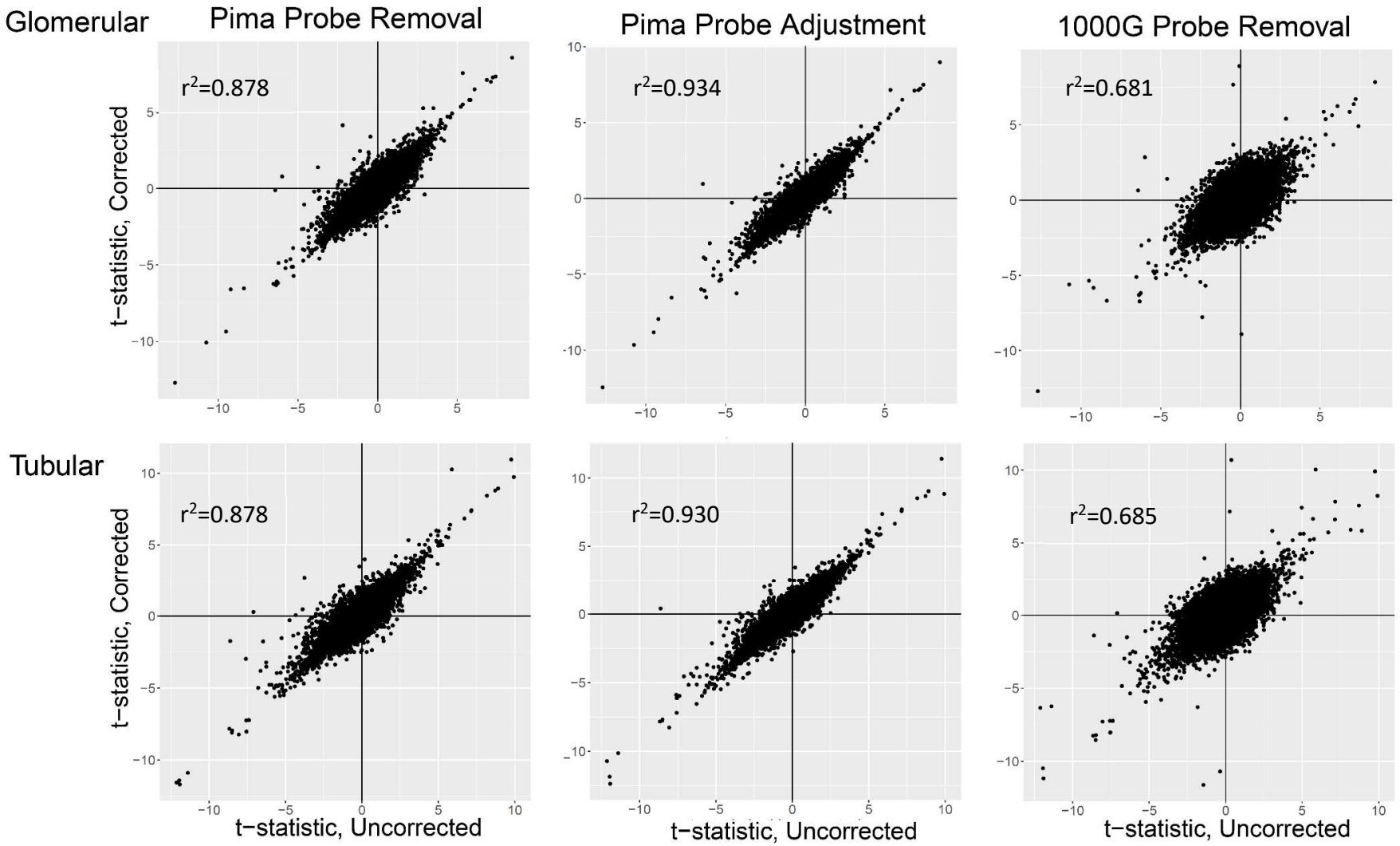
method	P-value	Total signals	overall		High LD		Low LD		
			negative	positive	negative	positive	negative	positive	No VIP
Uncorrected	1e-3	5,521	2,914	2,607	105	65	2,809	2,542	5,362
Probe Removal: Pima all		5,450	2,696	2,754	78	89	2,618	2,665	5,351
Probe Removal: Pima common		5,470	2,702	2,768	76	93	2,626	2,675	5,360
Probe Removal: 1000G common		5,468	2,739	2,729	80	85	2,659	2,644	5,359
Probe Removal: 1000G all		5,281	2,671	2,610	51	62	2,620	2,548	5,267
Probe Adjust: Pima		5,459	2,738	2,721	71	80	2,667	2,641	5,365
Uncorrected	1e-4	940	488	452	40	30	448	422	911
Probe Removal: Pima all		941	466	475	30	34	436	441	896
Probe Removal: Pima common		936	456	480	32	38	424	442	894
Probe Removal: 1000G common		912	451	461	27	33	424	428	898
Probe Removal: 1000G all		893	455	438	22	19	433	419	836
Probe Adjust: Pima		934	452	482	30	33	422	449	894
Uncorrected	1e-5	169	91	78	22	13	69	65	150
Probe Removal: Pima all		162	75	87	12	14	63	73	141
Probe Removal: Pima common		163	79	84	12	16	67	68	145
Probe Removal: 1000G common		169	83	86	15	15	68	71	145
Probe Removal: 1000G all		132	67	65	19	11	48	54	137
Probe Adjust: Pima		150	70	80	12	14	58	66	139

Supplementary Table 3.4A – comparison of different correction methods in peak eQTL analysis (Glomerular tissue)

This table shows the effect sizes at various thresholds for peak eQTLs (one per gene). The "overall" column includes all peak eQTLs, while the "high LD" column includes only variants with r^2>0.3 with a VIP, while the "low LD" column includes variants with r^2>0.1 with a VIP.

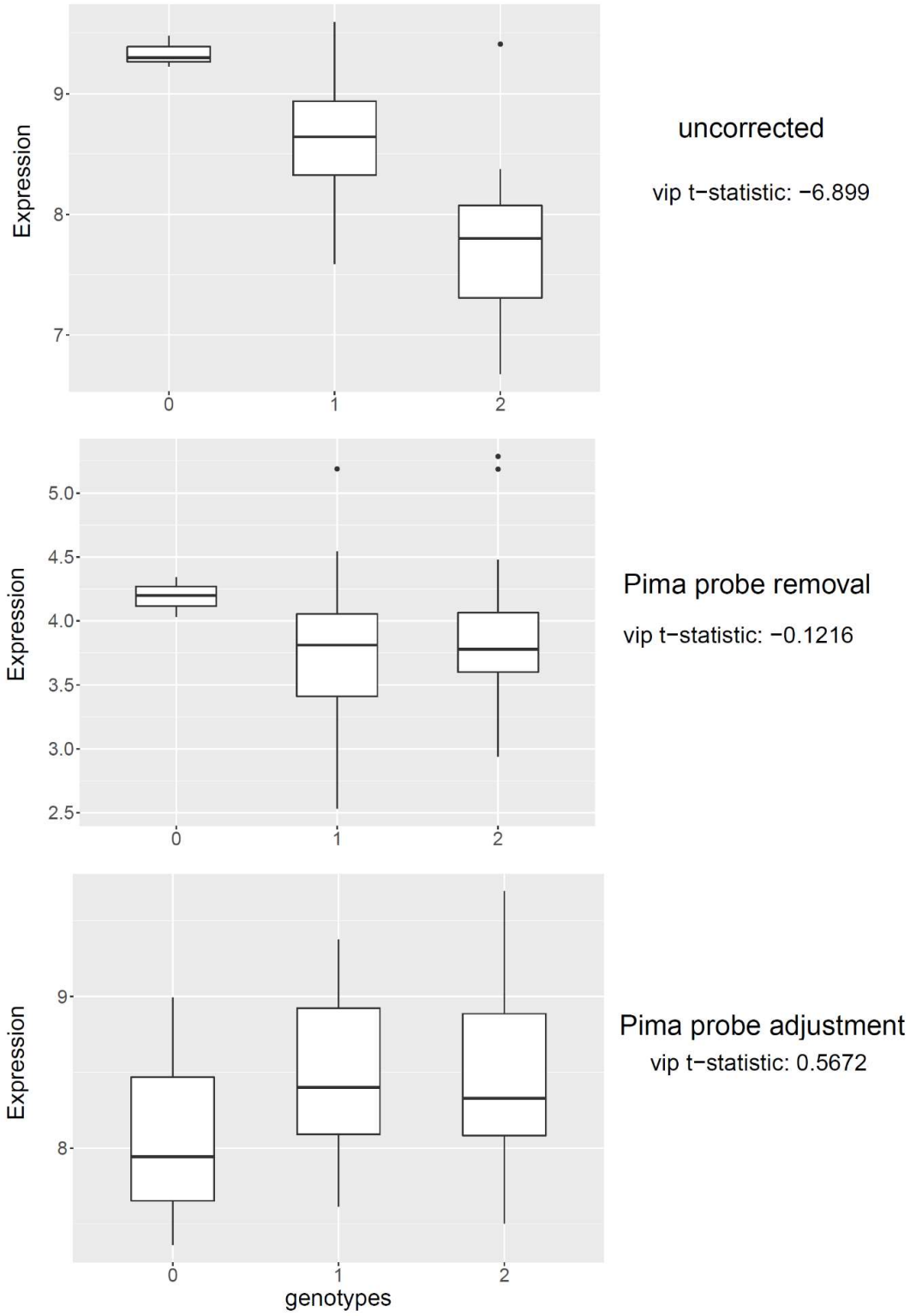
method	P-value	Total signals	overall		High LD		Low LD		No VIP
			negative	positive	negative	positive	negative	positive	
Uncorrected	1e-3	5,660	2,984	2,676	178	87	2,806	2,589	5,427
Probe Removal: Pima all		5,644	2,852	2,792	101	94	2,751	2,698	5,427
Probe Removal: Pima common		5,631	2,839	2,792	104	100	2,735	2,692	5,431
Probe Removal: 1000G common		5,660	2,888	2,772	110	100	2,778	2,672	5,399
Probe Removal: 1000G all		5,565	2,773	2,792	82	72	2,691	2,720	5,433
Probe Adjust: Pima		5,640	2,784	2,856	113	104	2,671	2,752	5,436
Uncorrected	1e-4	1,161	606	555	92	46	514	509	1,029
Probe Removal: Pima all		1,118	577	541	58	54	519	487	1,021
Probe Removal: Pima common		1,130	544	586	57	57	487	529	1,020
Probe Removal: 1000G common		1,101	545	556	60	56	485	500	1,018
Probe Removal: 1000G all		1,004	499	505	45	37	454	468	943
Probe Adjust: Pima		1,120	564	556	65	52	499	504	1,030
Uncorrected	1e-5	258	141	117	50	22	91	95	195
Probe Removal: Pima all		234	113	121	30	29	83	92	194
Probe Removal: Pima common		243	116	127	30	30	86	97	197
Probe Removal: 1000G common		242	110	132	30	30	80	102	195
Probe Removal: 1000G all		166	81	85	25	22	56	63	162
Probe Adjust: Pima		246	123	123	38	26	85	97	198

Supplementary Table 3.4B – comparison of different correction methods in peak eQTL analysis (Tubular tissue)



Supplementary Figure 3.1 – correlation between uncorrected and corrected expression

Here, we calculated the pairwise concordance between uncorrected and corrected expression levels for various correction methods. While some genes may be heavily biased if not corrected, many genes are only mildly affected by the hybridization bias. As such, we believe that it is important to avoid overcorrecting or unnecessarily removing probes for these mildly affected genes. From the 1000G probe removal approach, it is evident that removing too many probes adds noise to the expression levels, causing concordance to be low. Because our probe adjustment approach does not remove the probe, but rather estimates the value using other probes within the same probeset, we observe a higher concordance with the original uncorrected expression levels.



Supplementary Figure 3.2 – example of potential false positive gene (RPL9)

Here, we see an example of a likely false positive gene – RPL9 had a significant association with a VIP (p-value = 2.94x10^{-8}). After removing/adjusting the affected probe, we no longer observe a strong negative association, and the peak variant (which no longer was the VIP) was no longer statistically significant (p-value = 5.37x10-^{4})

Tissue	Correction Method	Likely False Positive eGenes when Uncorrected	eGenes Lost when corrected	Likely False Negative eGenes when Uncorrected	eGenes Gained when corrected
Glomerular	Probe Removal: Pima (all)	19 (28.8%)	66	19 (30.6%)	62
	Probe Removal: Pima (common)	18 (35.3%)	51	15 (34.1%)	44
	Probe Removal: 1000G (all)	20 (21.2%)	94	10 (15.6%)	64
	Probe Removal: 1000G (common)	15 (31.3%)	48	9 (17.0%)	53
	Probe Adjust: Pima	16 (40.0%)	40	9 (39.1%)	23
Tubular	Probe Removal: Pima	27 (31.7%)	85	10 (14.3%)	70
	Probe Removal: Pima (common)	27 (40.3%)	67	11 (19.3%)	57
	Probe Removal: 1000G (all)	41 (29.9%)	137	7 (10.4%)	67
	Probe Removal: 1000G (common)	27 (37.5%)	72	9 (14.3%)	63
	Probe Adjust: Pima	20 (48.8%)	41	9 (30%)	30

Supplementary Table 3.5 – False positive/negative candidates identified by each correction approach.

We compared the list of genes that no longer had significant eQTLs after expression correction and determined if they were likely false positives based on the criteria described in section 3.3.7. We repeated this for genes gained significant eQTLs after expression correction and determined if they were likely false negatives based on the criteria outlined in 3.3.7. Overall, we found that probe removal with 1000G all variants caused the most number of signals to change, but a relatively low proportion of these were actually false positives/negatives. As expected, methods that removed fewer probes changed fewer signals, with the probe adjustment method changing the least compared to the original analysis. However, despite changing the lowest amount of signals, the probe adjustment method had a high concordance with "likely" false positives and negatives.

Chapter 4 Systems Genetics Study in Pima Diabetic Nephropathy Cohort

4.1 Introduction

Systems genetics is an approach that seeks to holistically understand the biological mechanisms that drive complex traits and diseases [1]. These studies have the potential to unravel causal genes and pathways between endpoint phenotypes and DNA variation by examining intermediate phenotypes such as transcript, protein, or metabolite abundance. Using molecular phenotypes that are most relevant to the trait of interest can provide deeper insight than solely examining sequence variation. For example, gene expression studies such as eQTL (expression quantitative trait loci) mapping can be integrated with genome-wide association studies (GWAS) to predict causal genes and their functions. The advancement of high-throughput technologies and the increasing availability of large-scale molecular data have facilitated unprecedented studies which have furthered our understanding on various facets of systems genetics. For example, the UK biobank (UKBB) has collected genotypes on approximately 500,000 UK individuals with many phenotypes, including biological measurements, lifestyle indicators, blood and urine biomarkers, and brain/liver imaging [125]. The database of published associations discovered by UKBB is freely available online for researchers to query [158]. The TOPMed program has collected and sequenced 53,831 genomes with diverse ethnicities, including European, Asian, African, Latino/Hispanic, Native American and more [159]. In addition to genomes, the TOPMed program has also provided phenotypic measurements, such as transcriptomic profiles via RNA-seq from whole blood for hundreds of individuals [160]. Another example is the SCALLOP consortium, which has collected many phenotypes, proteomic measurements and genotypes to perform large-scale cardiovascular

pQTL (protein quantitative trait loci) mapping, which have identified potential drug therapy targets [161,162].

Despite the increasing availability of resources, not all tissues have been studied equally, and disease-relevant tissues can be difficult to obtain. Some easily accessible tissues, such as blood or cell lines have been well-studied with large sample sizes. For example, the eQTLgen consortium has meta-analyzed blood eQTLs of 31,684 samples and identified ~17,000 cis-eQTL genes [105]. The renal tissue on the other hand, is still largely underrepresented and is still an emerging area of research [106]. As of this date, only a handful of studies have assayed kidney expression and performed eQTL analysis. While the GTEx consortium (version 8) has assayed RNA-seq expression levels for many tissues (49 published), the sample size for the kidney tissue is the lowest among all tissues (n=73) and the number of discovered cis-eQTL genes (eGenes) is ~2.8-fold fewer than the next lowest, due to lower tissue quality and limited sample size [52]. Another eQTL scan was performed on 96 normal kidney samples from The Cancer Genome Atlas (TCGA), in which 1,886 eGenes were found [108]. One major limitation to both studies was that the kidney tissues were obtained from cortex in bulk, which does not represent key renal cell types - such as podocyte, glomerulus, tubular epithelial – that are relevant to the function of kidney. Furthermore, both GTEx and TCGA individuals were heavily biased towards European ancestry and does not represent genetic diversity of global populations.

Recently, the Nephrotic Syndrome Study Network (NEPTUNE) performed an eQTL mapping (nephQTL) with micro-dissected kidney compartments (glomeruluar and tubulointerstitial) on 187 individuals with nephrotic syndrome [106]. This resource substantially expanded our understandings of genetic regulation in renal tissues, even though it is focused on kidneys carrying rare diseases. There are increasing needs to profile the transcriptomes of each renal tissue compartment across diverse ancestries. The paucity of underrepresented populations in genomics studies is an important issue that, unless addressed properly, could contribute to further inequities in healthcare outcomes between different demographic groups [163,164]. Increasing the diversity of populations included in genomic studies could bridge this healthcare disparity by enriching our knowledge of genetic variation within different ancestry groups [165],

which will ultimately lead to improved and more accurate clinical care in a precision medicine setting [166].

Despite recent advances of eQTL studies focusing on renal tissues, systems genetics approach to understand the landscape of molecular and clinical profiles of renal tissues have not been explored yet. Molecular changes within renal tissues are more meaningful when they are also associated with clinical phenotypes, such as glomerular filtration rate (GFR) or albumin-creatin ratio (ACR). Morphometric measurements from imaging analysis of biopsy samples are also very important phenotypes to precisely quantify the function of individual kidneys. Multi-omics profiling integrated with morphometric and clinical phenotypes would make ideal resource for systems genetics studies to holistically understand the molecular basis of renal diseases.

In this chapter, we present a systems genetic study focusing on microdissected renal tissue compartments for an underrepresented indigenous population. In particular, we study the Pima diabetic nephropathy cohort, which followed 97 Pima Native Americans over 15 years (ClinicalTrials.gov number, NCT00340678). Among the many measurements taken include deep whole genome sequencing, transcriptomic profiles of microdissected kidney compartments at two timepoints, clinical measurements related to kidney function, and morphometric features obtained from biopsies. With these datasets, we performed eQTL mapping, as well as differential expression, genome-wide association on a number of traits, and transcriptome-wide association. Here, we discovered 805 glomerular eGenes and 1,118 tubular eGenes, with 129 novel tissue-specific and 64 novel population-specific eGenes not identified in previous studies. We also identified 4,605 genes differentially regulated by renal phenotypes, enriched in pathways specific to cytokine-cytokine receptor interactions, focal adhesion, cancer, and ECM-receptor interactions. We also report genome-wide significant associations with the VPC (volume of podocyte cell) morphometric trait for variants within the *C2CD4B* gene region – which has been identified as a potential risk variant for type 2 diabetes [167,168] – and a composite trait aggregated from multiple phenotypes. Our resource will help further our understanding of the molecular basis of diabetic renal diseases specific to cell-specific renal compartments for an understudied population.

4.2 Results

4.2.1 A Landscape of Native American Renal eQTLs with Deep Whole Genome Sequencing

The Pima diabetic nephropathy study provides numerous longitudinal clinical and multi-omics resources which can be used to help further our understanding of the kidney tissue from a systems genetics perspective. The depth and quality of these datasets provides an excellent opportunity to identify novel genes associated with renal diseases [169], potentially enriching our knowledge of this relatively understudied tissue. Among the many measurements gathered in this cohort include whole genome sequence data, transcriptomic profiles, kidney-related clinical traits, and kidney morphometry (Figure 4.1A). Of 77 participants received a biopsy in the first timeframe, 40 individuals received a follow-up biopsy in the second timeframe, and 37 dropped out of the study primarily due to poor disease progression or death. The second biopsy period introduced 20 new individuals to the study (Supplementary Table 4.1).

With the availability of deep, high quality datasets, we performed several analyses including identifying genetic determinants of gene regulation (eQTL mapping), identifying transcripts associated with phenotypes, identifying genetic variants associated with phenotypes (GWAS), identifying genes associated with phenotypes through genetic regulation (TWAS) (Figure 4.1B). In our eQTL analysis, we examined the association between genotypes versus each biopsy (B1 and B2) and tissue type - glomerulus (Glom) and tubulointerstitium (Tub) - across technologies (array and RNA-seq) separately, as well as jointly. Differential expression was performed using clinical and morphometric traits as explanatory variables under linear mixed model [170]. We also imputed gene expression levels using 44 tissues from Genotype-Tissue Expression (GTEx) Project, while applying a combined-tissue aggregate score to perform TWAS analysis on our morphometric and clinical traits (see Chapter 2). Finally, GWAS on morphometric, clinical, and composite phenotypes were performed using genotypes from whole genome sequencing (WGS).

For the sake of convenience, throughout this chapter we may sometimes abbreviate each gene expression assay as follows: biopsy 1 glomerular array expression (B1G-Array or B1GA), biopsy 2 glomerular array expression (B2G-Array or B2GA), biopsy 2 glomerular RNA-seq expression

(B2G-RNAseq or B2GR), biopsy 1 tubular array expression (B1T-Array or B1TA), biopsy 2 tubular array expression (B2-Array or B2TA), biopsy 2 tubular RNA-seq expression (B2T-RNAseq or B2TR).

4.2.2 Discovery of Pima cis-eQTLs

We sought to identify genetic determinants of gene regulation from the transcriptomic profiles of Pima Native Americans. Given the limited sample size, we focused on the cis-acting eQTLs identified within 1Mb of the transcription start site of each gene, from each of the 6 sets of transcriptomic profiles. We used linear mixed model eQTL analysis with EMMAX [157] and corrected for systematic variations after normalization to increase the power to detect cis-eQTL while correcting for false positives (see Methods). For more consistent and interpretable comparisons of cis-regulated genes, our results mainly focus on 19,612 protein-coding genes (GENCODE v27) unless indicated otherwise.

We identified a total of 805 glomerular and 1,118 tubular significant cis-eGenes across all biopsies and platforms (Table 4.1). Interestingly, despite that the first biopsy was based on older array-based technologies, in glomerular tissue, we identified more cis-eGenes (n=435) in the first biopsy than the second biopsies (178 from array, 351 from RNA-seq). We suspect that progression of glomerular damage increased heterogeneity between samples in biopsy 2 and the proportion of genetic variance among the overall variance of each gene (i.e. heritability) is higher for biopsy 1 compared biopsy 2. For tubulointerstitial compartment, array-based cis-eGenes were slightly smaller for biopsy 2 (n=285) compared to biopsy 1 (n=315). However, RNA-seq in biopsy had much higher power to detect significant cis-eGenes (n=814). These results suggest that RNA-seq improves the power to detect eQTLs but heterogeneity between samples also substantially affect the power of eQTL studies.

We also attempted joint eQTL analysis across three datasets using the APEX software tool [171], treating the overlapping sample between experiments as identical twins using linear mixed model. While this approach found more eQTLs compared to single array datasets, it did not always identify more eQTLs than our RNA-seq datasets, presumably due to large heterogeneity between platforms (Table 4.1, Supplementary Table 4.8).

4.2.2.1 Concordance of eQTLs between tissues and biopsies

To understand the similarity and differences in transcriptomic regulation between the glomerular and tubular tissues, we examined the concordance cis-eQTLs between the tissue types, biopsies, and assays. It has been shown that eQTL overlaps can be severely underestimated due to limited power [172], so we considered a cis-eQTL as 'replicated' in one of other two experiments for the same tissue if the Bonferroni-adjusted p-value is nominally significant (i.e. adjusted $p<0.05$, point-wise $p<0.025$) at the same SNP. With these criteria, we observed that 28-47% of glomerular cis-eQTLs and 35-60% of tubular cis-eQTLs are replicated in another experiment within the same tissue (Table 4.2). When comparing within the same experiment but different tissues, we note that replication rates are higher overall within the same biopsy than between biopsies, and overlaps are higher within biopsy 1 compared to biopsy 2. For example, 47.1% of B1GLOM-Array eQTLs are replicated by B1TUB-Array, whereas only 25.2% of B2GLOM-Array eQTLs are replicated by B2TUB-Array. This trend is also observed in the tubular tissue, where 58.7% of B1TUB-Array eQTLs are replicated by B1GLOM-Array, but only 36.5% of B2TUB-Array signals are replicated by B2GLOM-Array. This lower replication rate in the second biopsy suggests that glomerular and tubular eGenes become more tissue specific as time passes on.

4.2.2.2 Identification tissue-specific and population-specific cis-eQTLs novel to GTEx

Next, we sought to identify potential novel cis-eQTLs discoveries from our dataset compared to published eQTLs from GTEx project. For each peak cis-eQTLs from our study, we classified it as "novel" if it was located outside of the 95% credible set of any of the finely-mapped cis-eQTLs for each of the 48 GTEx version 8 tissues (except for kidney) [52,173] in which the deterministic approach of posteriors (DAP) algorithm was used to detect eQTLs [174]. We also evaluated overlap with kidney cis-eQTLs from NephQTL [106], to identify kidney-specific eQTLs shared with European ancestry. Overall observed that ~45%-55% of Pima cis-eQTLs overlapped with at least one of the GTEx cis-eQTLs, with the same proportion of eQTLs overlapping with nephQTL signals (Table 4.3).

To potentially understand and identify the mechanisms behind our novel signals, we further classified the novel signals into three coarse categories. We hypothesized that Pima signals could have been novel due to the signals being (1) highly tissue-specific (expressed in kidney, not expressed in GTEx tissues), (2) population-specific (low non-reference allele frequency in non-Pima individuals) or (3) other reasons, which may include lack of power in the GTEx tissues.

Among the genes novel to GTEx, we identified a total of 53 tissue-specific eQTL variants from the glomerular compartment and 83 tissue-specific eQTL variants from the tubular compartment, with a total of 129 tissue-specific eQTL variants between the two tissues (Figure 4.2A, Supplementary Table 4.5). To determine tissue-specificity of each eGene, we calculated the median TPM for each gene across each of the 49 GTEx tissues and compared this to the median TPM of the RNA-seq expression levels for both tissue compartments in the Pima cohort. Each gene was designated as tissue-specific if the Pima median TPM was higher than the top 5% of median TPMs among the GTEx tissues.

Next, we found a total of 64 peak eQTLs in to be population specific, with 26 coming from the glomerular tissue and 45 from the tubular tissue (Figure 4.2B, Supplementary Table 4.6). To determine the population-specificity of novel variants, we compared the minor allele frequency of each peak cis-eQTL between the Pima and European populations. To do this, we examined specifically the European individuals from the 1000 Genomes reference panel. We defined a variant as being Pima-specific if the minor allele frequency difference between the two populations were greater than 20%, and less than 5% overall for the 1000 Genomes population.

Finally, there were 23 genes that had both tissue-specific and population-specific eQTLs. These genes were involved in Acyl-CoA dehydrogenases (*ACOX2, ACAD10*), fatty acid metabolism (*D2HGDH, BPHL*), reactive oxygen process pathway (*NDUFA6, TNXRD2*), DNA binding (*MSH3*), and the glyoxylate metabolic process (*AGXT2*).

4.2.3 Association with Phenotypes and Measured Expression

We conducted association analysis between expression levels of each gene and phenotypes to identify genes differentially regulated based on phenotypes. Specifically, for each tissue (glomerular and tubular) and biopsy, we examined the pairwise association between each of the 28 phenotypes (25 morphometric and 3 clinical) and individual genes using linear model and identified genes significantly associated at FDR 0.05 [175]. Association between these measured gene expressions and phenotypes may implicate their relationships in either direction: genes affecting phenotypes or phenotypes affecting the genes. A total of 4,605 genes (2,650 in glomerular and 2,579 in tubular) were identified as significant in at least one of analysis, and 2,157 genes were significant in two or more analyses (Figure 4.4, Supplementary Table 4.12)

Gene set enrichment analysis using KEGG pathway [176–178] with Enrichr software [179] identified many significantly enriched biological pathways associated with morphometric and clinical phenotypes. For example, cytokine-cytokine receptor interactions, which was previously implicated for diabetic kidneys and renal carcinoma [180–182], were significantly enriched with Glomerular Filtration Rate (GFR), Urinary Albumin Creatin Ratio (uACR), and multiple morphometric traits (VVAT_NS, VVATTT_NS, VVMM) (See Table 4.4A, 4.4B). These results suggest that cytokine response to activating stimulus may be an important factor involved with the function of diabetic kidneys. Focal adhesion pathway was significantly enriched (adjusted $p=2.2x10^{-4}$) with numerical density of podocyte cell per glomerulus (NVPC), which is consistent to the previous studies capitalizing on the importance of focal adhesion in podocyte attachment within glomerular structure [183,184]. We also identified pathways related to cancer and ECM-receptor interaction, which had significantly enriched associations between glomerular gene expression and morphometric traits measuring glomerular basement membrane width (GBM) and mesangial fractional volume (VVMES), suggesting potentially shared factors between diabetic kidneys and renal carcinoma [185,186].

4.2.4 GWAS with Clinical and Morphometric Traits

We performed GWAS across 28 first biopsy and 25 second biopsy clinical and morphometric traits, as well as with the 13 first biopsy and 10 second biopsy composite traits (derived by principal components, Supplementary Materials 4.6.1). To correct for multiple traits, we used a conservative Bonferroni significance threshold of 2.0×10^{-9} (5×10^{-8}, corrected across 25 traits). With this p-value threshold, we did not find any statistically significant signals. However, we found some marginal signals that may be suggestive of true associations. For example, for the biopsy 2 VPC trait (Figure 4.3A), we identified an association with variant *rs11637089* (p-value = 2.01×10^{-8}, minor allele frequency = 38.8%), lying within the *C2CD4B* region, which has been previously identified as a risk-variant for type 2 diabetes [167,168]. Using the composite traits, we identified variant *rs1559274* to have suggestive association with the PC3 trait from biopsy 2 (p-value = 9.42×10^{-9}, minor allele frequency = 30.0%), which corresponds to SV, uACR, VVPC and NVPC (Figure 4.3B). This top locus lies within 500KB downstream of the *AGXT2* gene, which has been implicated in regulation of blood pressure [187]. *AGXT2* was also found to have tissue-specific and population-specific eGenes in Pima eQTLs. Given that PC3 is driven by uACR, which has been shown to be correlated with blood pressure [188–190] there may be some suggestive evidence that this genetic marker play a part in regulation of blood pressure as well.

Although power to detect associations in GWAS can depend on various factors including allele frequency and effect size or trait heterogeneity, sample size is an important factor for these studies [191–193]. Given that sample sizes for many GWA studies can number in the thousands of even tens or hundreds of thousands [194], the relatively scarcity of signals here is not surprising given the sample sizes (n=77, n=60 for biopsy 1 and 2 respectively). Despite the marginal levels of association overall, it appears that some of the top signals may provide meaningful insight into the biology of diabetes and kidney function.

4.2.5 TWAS Between Predicted Expression and Clinical and Morphometric Traits

We performed TWAS by associating predicted gene expression levels with clinical and morphometric traits, as well as with composite traits. Here, the predicted gene expression was based on a SWAM model (Chapter 2) derived from the GTEx version 8 whole blood tissue, for

13,907 protein coding genes (see methods). To correct for multiple testing, we used a Bonferroni corrected p-value threshold of $1.5x10^{-7}$ for regular traits and $3.0x10^{-7}$ for composite traits. While we did not discover any significant associations for the regular or composite traits, we found some marginal associations that may be of biological relevance. The top association among biopsy 2 traits was between VVPT_NS and the *TFDP2* gene (p-value = $7.85x10^{-7}$), which contains risk variants previously identified in chronic kidney disease [195].

4.3 Discussion

In this chapter, we presented a longitudinal systems genetics resource based on the diabetic nephropathy cohort of Pima Native Americans, encompassing deep whole genome sequencing, transcriptomic profiles of two microdissected renal compartments, clinical traits, and morphometric phenotypes. Given that there have been few kidney transcriptomic resources (GTEx, nephQTL) and fewer yet with microdissected tissue compartments to differentiate between cell types and function, our work on this underrepresented, population-specific cohort can potentially provide a unique perspective of biological mechanisms underlying kidney disease.

With the various analyses performed in this study, we were able to replicate numerous established or hypothesized pathways relevant to diabetes and kidney disease, as well as uncover potentially new signals. For example, our eQTL analysis replicated many signals from GTEx tissues, while also discovering tissue- and population-specific novel signals that we believe to be authentic. Indeed, among the 805 glomerular and 1,118 tubular eGenes, roughly 50% of the genes were also eGenes in one of the 48 GTEx version 8 tissues. On the other hand, we discovered 129 genes are likely novel due to tissue-specificity (genes expressed in Pima Kidney tissue but not GTEx tissues) and 64 genes due to population-specificity (novel signals due to studying a population that is distant from the GTEx samples). While the sample size of the cohort was sufficient to discover many eQTLs, GWAS are usually conducted using much larger cohorts, and often even meta-analyzed across multiple studies [196]. As such, it is not surprising that our GWAS provided only marginal associations for very few traits. Yet, our GWAS

with the VPC trait (volume of podocyte cell) identified genetic variants located within the *C2CD4B* gene region, which has been previously linked to diabetes susceptibility [167,168]. In addition to eQTL and GWAS, we characterized our clinical and morphometric traits in the context of transcriptome variation, testing for differentially expressed genes as well as conducting TWAS. With our measured expression association tests, we discovered 4,605 genes associated with morphometric traits, highlighting various relevant biological pathways such as cytokine-cytokine receptor interactions, focal adhesion, cancer, and ECM-receptor interactions. Finally, our TWAS also may have found a relevant association between VVPT_NS and *TFDP2,* a gene implicated in chronic kidney disease. All of these findings highlighted above will serve as a valuable resource to both validate and augment our current transcriptomic and genomic knowledge base for kidney structure and kidney disease progression, particular for the Native American population.

The study design and quality of this dataset yields several distinct benefits that make it an valuable kidney transcriptome resource. Firstly, studying microdissected compartments of renal tissues can be extremely important in understanding the cell-type-specific regulation of renal transcriptomes. Even though a large number of eQTLs are shared between glomerular and tubular compartments, it was clear that many eQTLs are shared within the same compartments, even at differing time periods. Differentiating cell-types within an organ has been shown to more accurately identify new disease pathways by capturing the signature of important cell types, such as podocytes [197,198]. We expect that single-cell transcriptomic profiling technologies will allow us to understand cell-type-specific nature of transcriptional regulation in renal tissues much more precisely [199]. Secondly, the longitudinal collection of data gives multiple snapshots of transcriptomic profiles, allowing us to observe change in potential biomarkers that may be associated with renal disease state. Overall, we observed that eQTLs were more highly detected in biopsy 1 microarrays compared to those from biopsy 2, which likely could be attributed to heterogeneity due to disease progression. However, identifying the exact biomarkers associated with this progression remains challenging, as the batch effect between the two biopsies are completely confounded with the time variable. A future direction of research could be to deconvolute this confounder by calibrating the biopsy 2

microarray expression with the RNA sequencing data, which could potentially be achieved with more sophisticated statistical or computational methods. Finally, while GTEx provides an excellent reference for transcriptomic profiles of many tissues, the Pima samples have the benefit of being biopsied from live individuals. Because many of the high quality GTEx tissues are donated as transplants, the condition of the remaining tissue samples may be subpar compared to those from the Pima cohort. As such, we believe that our study will serve as a valuable resource to complement the recently growing pool of information in the field of kidney transcriptomics, providing both tissue-specific and population-specific insights.

4.4 Materials and Methods

4.4.1 Data Source

The Pima diabetic nephropathy cohort provides a deep longitudinal catalogue of genomic, transcriptomic, morphometric, and clinical resource focusing on renal traits in Pima Native Americans. In this study, 97 individuals with minimum of 5 years of type 2 diabetes (T2D) were followed over a time period of 15 or more years. Many samples (n=68) had early onset of diabetic nephropathy. In this chapter, we focus on four different types of genomic and clinical measurements (Fig 4.1A) that were collected from the study. First, we deeply sequenced 97 Pima Native Americans to comprehensively identify genetic variation within the cohort. Second, we assayed transcriptomic profiles between two time points with multiple technologies (microarray and RNA-seq) across two micro-dissected kidney compartments (glomerulus and tubulointerstitial). The first biopsy was taken from 2003-2007 while the second biopsy was taken from 2014 onward. Individuals who remained healthy enough at the second time point underwent the second biopsy, while those whose disease progression prevented them from safely undergoing an additional biopsy were excluded. Third, clinical phenotypes relevant to kidney functions were collected several times a year, such as Glomerular Filtration Rate (GFR), Urinary Albumin-Creatine Ratio (uACR), and hemoglobin A1C (HbA1c). Finally, morphometric measurements including volume of podocyte (VPC) and messangial cells (VVMES) within the glomerulus as well as cortical interstitial fractional volume (VVint) (and many more) were taken with each biopsy.

4.4.2 Whole Genome Sequencing

Deep whole genome sequencing was performed on 97 individuals using Illumina HiSeq X-Ten at the MacrogenLab. A mean depth was 32x was achieved, with 99.3% coverage of the genome (98.77% covered with at least 10x depth). Overall, the quality metrics looked excellent after 2 potentially contaminated samples were identified and re-sequenced. We used the GotCloud [152] pipeline to produce SNP calls, using 1000G genotypes as cues. We also detected novel SNPs, which were included if there was strong evidence of being a true positive. We then used HaplotypeCaller [200] to detect both SNPs, as well as insertions and deletions (indels). We then generated variant call files (VCFs) using the SNPs from GotCloud and indels from HaplotypeCaller.

In addition to whole genome sequencing, a genotyping array was used for 54 of the samples. We checked the concordance for the overlapping variant calls between these two genotyping technologies. Excluding 5 samples, we found the concordance overall to be very high across shared sites (>99.3%). Of the 5 samples with low concordance, 4 of them appear due to be a quality issue with the genotyping array, and the other one a sample swap.

4.4.3 Measurements of Expression

The expression microarray platforms used for the first biopsy were the Affymetrix HGU-133A (glomerular n = 21, tubular n = 22) and HGU-133 Plus 2 arrays (glomerular n = 48, tubular n = 24). In the second biopsy, the Affymetrix HuGene 2.1 array was used (glomerular n = 50, tubular n = 54). All three platforms consisted of 25-mer probe sequences specifically designed to target individual exons. To harmonize between platform differences, we used a custom probe-to-probeset mapping provided by the Microarray Lab from the Molecular and Behavioral Neuroscience Institute at the University of Michigan [148,149]. Under this mapping, the HGU-133A platform contained 12075 probesets over 174129 probes, the HGU-133 Plus 2 platform contained 19703 probesets over 333134 probes, and the HuGene 2.1 platform contained 25583 probesets over 466204 probes.

RNA sequence data was available for both glomerular and tubular tissues in biopsy 2. Here, reads were aligned with TopHat [150] software tool and the transcript counts were quantified

with Cufflinks [151] and normalized via log-transformation of FPKM (fragments per kilobase of transcript per million mapped reads). *SVDiff* was also applied to RNA-seq data for eQTL analysis.

4.4.4 Clinical and Morphometric Measurements

Clinical traits were taken every six months, measuring kidney function metrics such as urinary albumin-to-creatinine ratio (uACR), glomerular filtration rate (GFR) and hemoglobin A1c (HbA1c). Morphometric traits were taken from the sampled biopsies with 25 features measured in the first biopsy and 22 features in the second biopsy. The full list of morphometry features is shown in Supplementary Figure 4.11.

4.4.5 Normalization of *Microarray* Gene Expression

To normalize expression, we applied Bioconductor's robust-multiarray averaging (RMA) to each microarray platform separately. This method also takes probe-level intensities and combines them into probeset-level expression, which are later converted to genes. To maintain a consistent set of genes across the different platforms, we used a custom probe-to-probeset mapping provided by the Microarray Lab from the Molecular and Behavioral Neuroscience Institute at the University of Michigan. Under this mapping, the HGU-133A platform contained 12075 probesets over 175,294 probes, the HGU-133 Plus 2 platform contained 19703 probests over 333134 probes, and the HuGene 2.1 platform contained 25583 probesets over 466204 probes.

Because the microarray experiment for biopsy 1 was done in two separate batches (both with different platforms), we used ComBat, an empirical Bayes batch correction method [68] to combine across platforms into a unified dataset for the first biopsy. To further deal with latent systematic technical effects in the array data, we used a singular-value decomposition method (which is outlined in the supplementary materials) similar to PEERS [201] in which we factor the expression matrix, identifying components of variation. We called this method *SVDiff* and found this approach worked very well in terms of removing these systematic biases, and in increasing the power of our eQTL analysis (see Supplementary Materials 4.6.2).

4.4.6 Variant-Aware Correction of Microarray Expression

It is well-known in microarray experiments that when target probe sequences contain a genetic variant for an individual, there can be a hybridization bias due to differential binding affinity caused by said variant [100,101]. Typically, this negative hybridization will result in systematically lower probe intensity levels and potentially create false association signals. To combat this, we used our in-house software to identify probes that overlapped with Pima variants and removed them from the analysis (See Chapter 3). For the HGU-133A platform, we identified 7,542 probes (4.3% of the total probes) affecting 4,375 probesets (36.0% of the total probesets). For the HGU-133 Plus platform, we identified 14,840 probes (4.4% of the total probes) affecting 7,992 probesets (40.4% of the total probesets). Finally, for the biopsy 2 platform, HuGene 2.1 ST, we identified 27,767 probes (6.0% of the total probes) affecting 13,219 probesets (51.1% of the total probesets).

4.4.7 eQTL Mapping

Expression quantitative trait locus (eQTL) analysis was performed using mixed model association via the EMMAX software package [157]. A separate analysis was performed for each tissue. For every SNP and indel identified from our whole genome sequencing, we tested for association against each of the genes with measured expression. To account for potential confounders, we adjusted for age and sex as covariates. In addition, to account for potential familial-relatedness, we calculated a pairwise kinship matrix for all samples, using it as the fixed-effects component of the mixed model.

We defined an eQTL as cis-acting if it was located within 1 Mb of the transcription start site of the associated gene. Otherwise, the eQTL signal was defined as being trans-acting. To account for multiple testing, p-values were adjusted by the false-discovery rate (FDR) correction approach, using the trans-eQTL signals to determine the false discovery rates. Even though the p-value thresholds at FDR < 0.05 are different between datasets (Supplementary Table 4.3), they were reasonably close to each other, so we used a fixed pointwise p-value threshold of $5x10^{-6}$ for straightforward comparisons between different eQTL datasets. Because of linkage

disequilibrium (LD) between neighboring SNPs, only the SNP with the lowest p-value was taken as the true cis-eQTL signal for each gene.

4.4.8 Combined Biopsy eQTL Mapping

In addition to eQTL mapping with each tissue/biopsy as a separate analysis, we performed a joint tissue analysis where we combined across biopsies for each tissue. To do this, we used APEX (All-in-one Package for Efficient Xqtl analysis) [171], inputting B1GA, B2GA, and B2GR simultaneously, as well as B1TA, B2TA, and B2TR simultaneously. To account for duplicate individuals, we specified kinship coefficients of 1 (monozygotic twins) for same individuals and 0 (unrelated) for all other pairwise relationships. The results of this analysis are shown in Table 4.1 and Supplementary Table 4.5 alongside our main eQTL analysis.

4.4.9 GWAS with Morphometric and Clinical Traits

We performed a GWAS analysis between the Pima genetic variants and the morphometry and clinical data. Here, we tested 3 clinical traits (uACR, GFR, and HBA1c) and 25 biopsy 1 morphometric traits, and 22 biopsy 2 morphometric traits (Supplementary Table 4.9). Like our eQTL analysis, adjusted for age and sex as covariates, and used the pairwise kindship matrix to account for familial structure. Because many of the morphometric traits are highly correlated, we also performed a PCA analysis on the traits, using the top 10 PCs for the analysis, in addition to all the original traits.

To determine the significance threshold, we used $5x10^{-8}$ for each trait, and applied a Bonferroni correction based on the number of traits analyzed. For the regular traits, we used a p-value threshold of $2x10^{-9}$ and for the composite traits, we used a threshold of $5x10^{-9}$.

4.4.10 Association Analysis between Measured Gene Expressions and Phenotypes

We performed association analysis between 28 phenotypes – 25 morphometric, 3 clinical – and the expression levels of each gene, per tissue, biopsy, and assay. For array-based expression, we used quantile-normalized expressions obtained from RMA. For expression levels from RNA-seq, we used FPKM values. The association analysis was performed using a linear model, with gene expression levels as the response variable and the phenotypes as predictor variables,

accounting for sex and age as covariates. For each dataset, we identified significantly associated genes using FDR < 0.05 threshold [175] and the significant genes were merged across biopsies and assays for each tissue and each phenotypes when summarized into Figure 4.4 and Supplementary Table 4.9 To perform pathway enrichment analysis, we used EnrichR [179] using KEGG 2019 Human database [176–178]. We reported only the significant enrichment with adjusted p-value < 0.001 to account for multiple comparisons. When a certain phenotype/tissue has many significantly enriched pathways, we only listed top 5 pathways in terms of p-value.

4.4.11 Transcriptome Wide Association Analysis

We performed transcriptome-wide association mapping by first using SWAM (chapter 2) to derive a whole blood model using 49 tissues from GTEx version 8 [52] and DGN whole blood tissue [50]. We then used PrediXcan [55] to calculate predicted expression levels based on the WGS genotypes of the 97 Pima individuals. From this prediction model, we successfully imputed 15,319 genes. We filtered these genes further using gencode v27 protein coding genes, resulting in 13,907 genes for the analysis. Next, we used EMMA [202] to perform mixed-model association using clinical and morphometric traits as the outcome variable, and predicted expression levels as explanatory variables, adjusting for age and sex. We repeated this using the composite traits derived from our PC analysis (Supplementary 4.6.2). To account for population substructure from predicted expression levels, we calculated a covariance matrix between individuals across all genes and modeled it as a random effect. We accounted for multiple testing by setting a Bonferroni-corrected p-value threshold of 1.5×10^{-7} (correcting across 13,219 genes and 28/25 traits) for original clinical/morphometric traits, and 3.5×10^{-7} for composite traits (correcting across 13,219 genes and 10 composite traits).

4.5 Figures and Tables

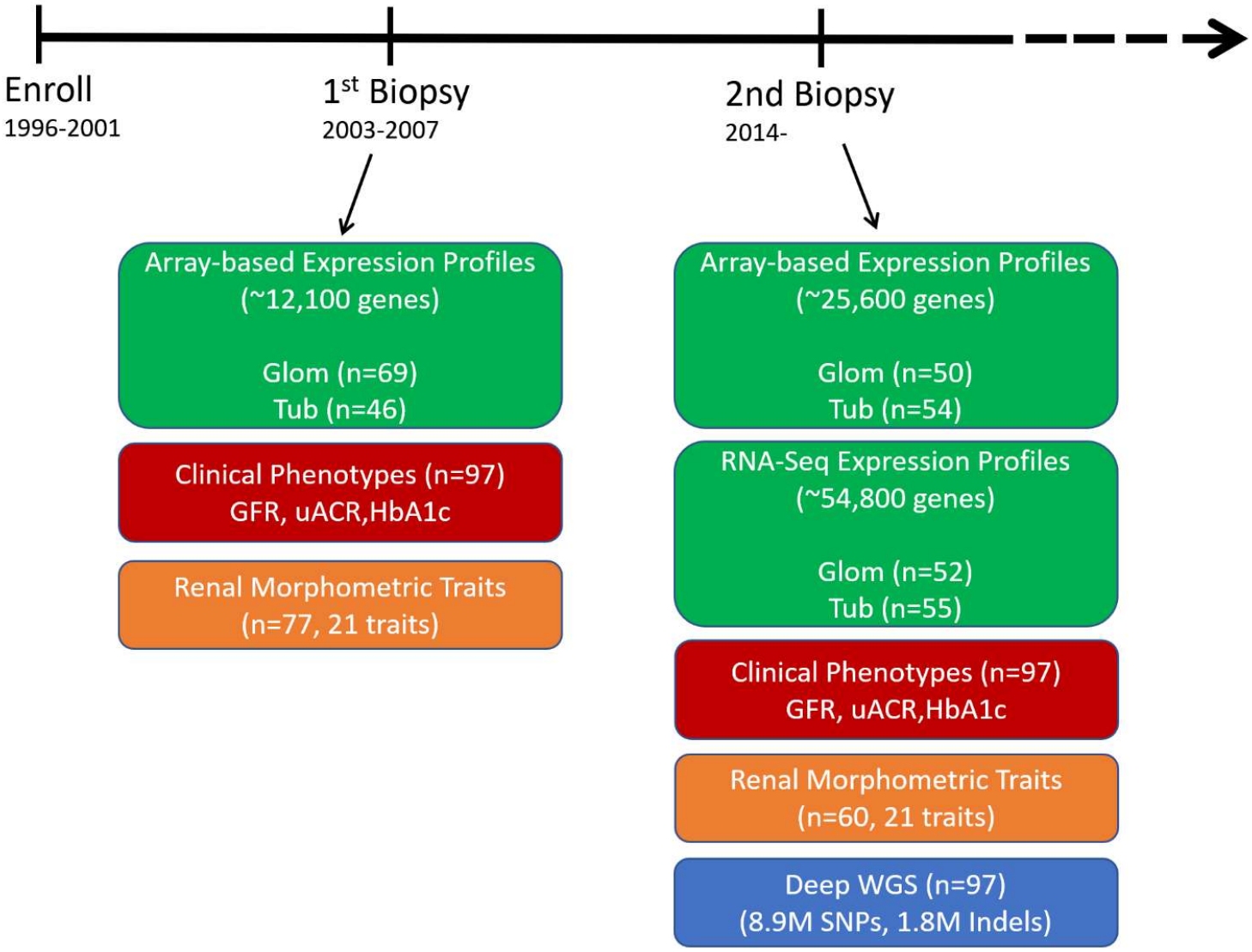


Figure 4.1A – Overview of the Pima study

The Pima diabetic nephropathy cohort is a longitudinal study focusing on microdissected renal tissue compartments – glomerular and tubulointerstitial. Among the many measurements taken from the study included deep whole genome sequencing, gene expression levels (microarray and RNA-seq), clinical phenotypes, and morphometry traits determined by biopsies performed at two time points. Second biopsies were only performed on individuals healthy enough to undergo the operation 8 years after the initial biopsy. Additional participants who were added to the study later on were classified in the 2nd biopsy groups.

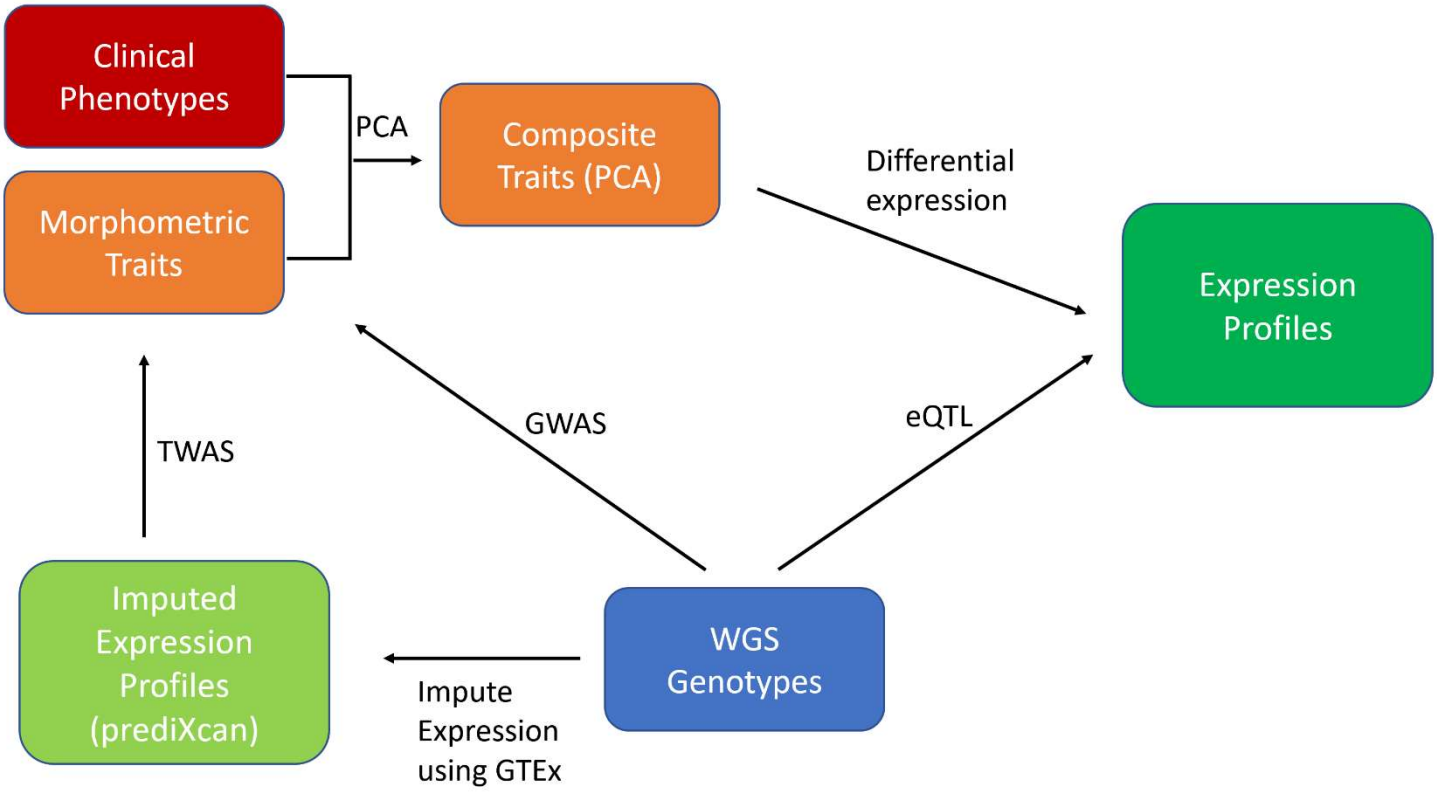


Figure 4.1B – Overview of analyses performed in this chapter

With the multitude of datasets available, we performed various analyses to provide a multi-faceted perspective of this unique cohort, including eQTL analysis, GWAS, TWAS, and trait-expression association with both measured expression levels and imputed expression levels.

Biopsy	Platform	N samples	# eGenes
B1 Glom	Microarray	69	435
B1 Tub	Microarray	46	315
B2 Glom	Microarray	50	178
B2 Tub	Microarray	54	285
B2 Glom	RNA-seq	52	351
B2 Tub	RNA-seq	55	814
Apex-Glomerular	Combined	93(171)	403
Apex-Tubular	Combined	77(155)	408

Table 4.1 – eGene discovery from cis-eQTL analysis

Our main cis-eQTL analyses were performed using mixed-model regression via EMMAX [156]*. We normalized each dataset with our SVDiff method, which is outlined in section 4.6.2. We also performed a combined analysis, where we aggregated across biopsies and platforms for each tissue type. To do this, we used the APEX software* [170] *to meta-analyze the multiple datasets. We filtered genes to only include protein coding genes according to gencode version 27.*

Platform	# eGenes	# Replicates					
		B1GA	B2GA	B2GR	B1TA	B2TA	B2TR
B1GA	435	-	144 (33.1%)	204 (46.9%)	205 (47.1%)	115 (26.4%)	182 (41.2%)
B2GA	178	64 (36.0%)	-	81 (45.5%)	38 (21.3%)	56 (25.2%)	60 (33.7%)
B2GR	351	115 (32.8%)	98 (27.9%)	-	65 (18.5%)	80 (22.7%)	170 (48.4%)
B1TA	315	185 (58.7%)	73 (22.5%)	88 (27.9%)	-	144 (45.7%)	177 (56.2%)
B2TA	285	85 (29.8%)	104 (36.5%)	82 (28.8%)	116 (40.7%)	-	172 (60.4%)
B2TR	814	229 (28.1%)	168 (20.6%)	331 (40.7%)	283 (34.8%)	341 (41.9%)	-

Table 4.2 – cis-eQTL replication across tissues and biopsies

For every peak cis-eQTL (p-value = $5x10^{-6}$), we searched for replication for that variant with other tissues/platforms. We considered the eQTL to be successfully replicated in another experiment using a p-value threshold of 0.025.

Platform	All eGenes	Novel eGenes*	Tissue-specific eGenes	Population-specific eGenes	Tissue- & Population-specific eGenes	eQTL is Replicated in NephQTL
B1G (Microarray)	435	199	33	13	3	206
B1T (Microarray)	315	151	35	19	7	160
B2G (Microarray)	178	96	20	9	3	86
B2T (Microarray)	285	142	25	12	1	155
B2G (RNA-seq)	351	173	22	20	2	164
B2T (RNA-seq)	814	358	69	37	8	438

Table 4.3 – eQTL breakdown compared to other datasets

We compared Pima eQTLs to other resources, namely GTEx and nephQTL. Genes were considered novel if they were not in the 95% credible set for any GTEx tissue, excluding kidney. From the list of novel genes, we further classified them as being tissue- or population-specific based on the criteria outlined in 4.2.2.2. For nephQTL replication, we considered a cis-eQTL replicated using a p-value threshold of 0.05.

(*) Novel eGenes compared to GTEx version 8 tissues, excluding GTEx-Kidney

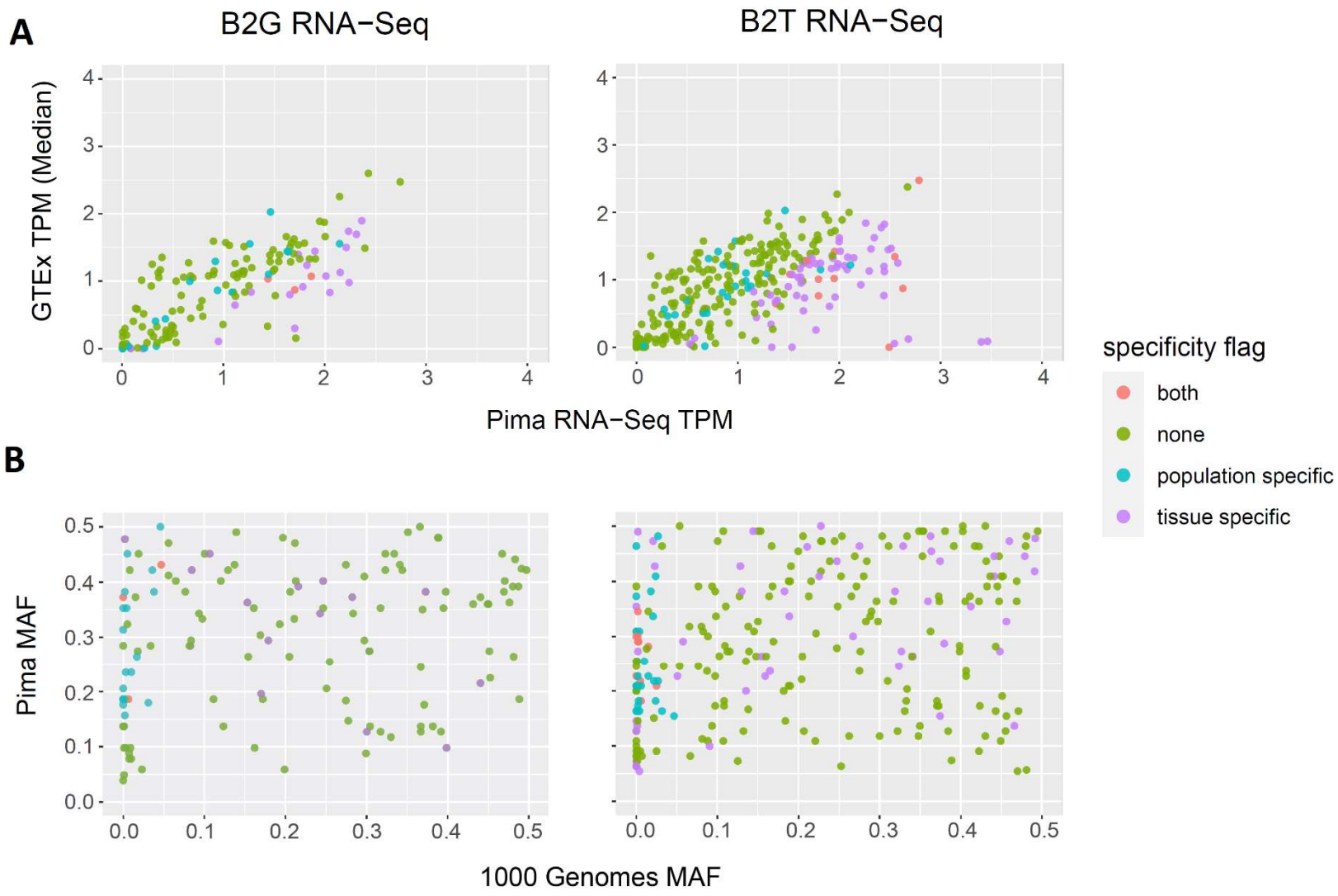


Figure 4.2 – breakdown of pima cis-eQTL variants compared to GTEx version 8

In these plots, each point is a novel gene(as defined by Table 4.3), color coded to show if they are tissue-specific, population-specific, or both. (A) shows the comparison of median TPM between Pima RNA-seq and GTEx tissues. Here, we approximated the TPM count in the Pima individuals by applying the formula: tpm = exp(log(fpkm)-log(sum(fpkm)) + log(10^{-6})). (B) shows the minor allele frequencies between the Pima population and European samples from the 1000 genomes reference panel.

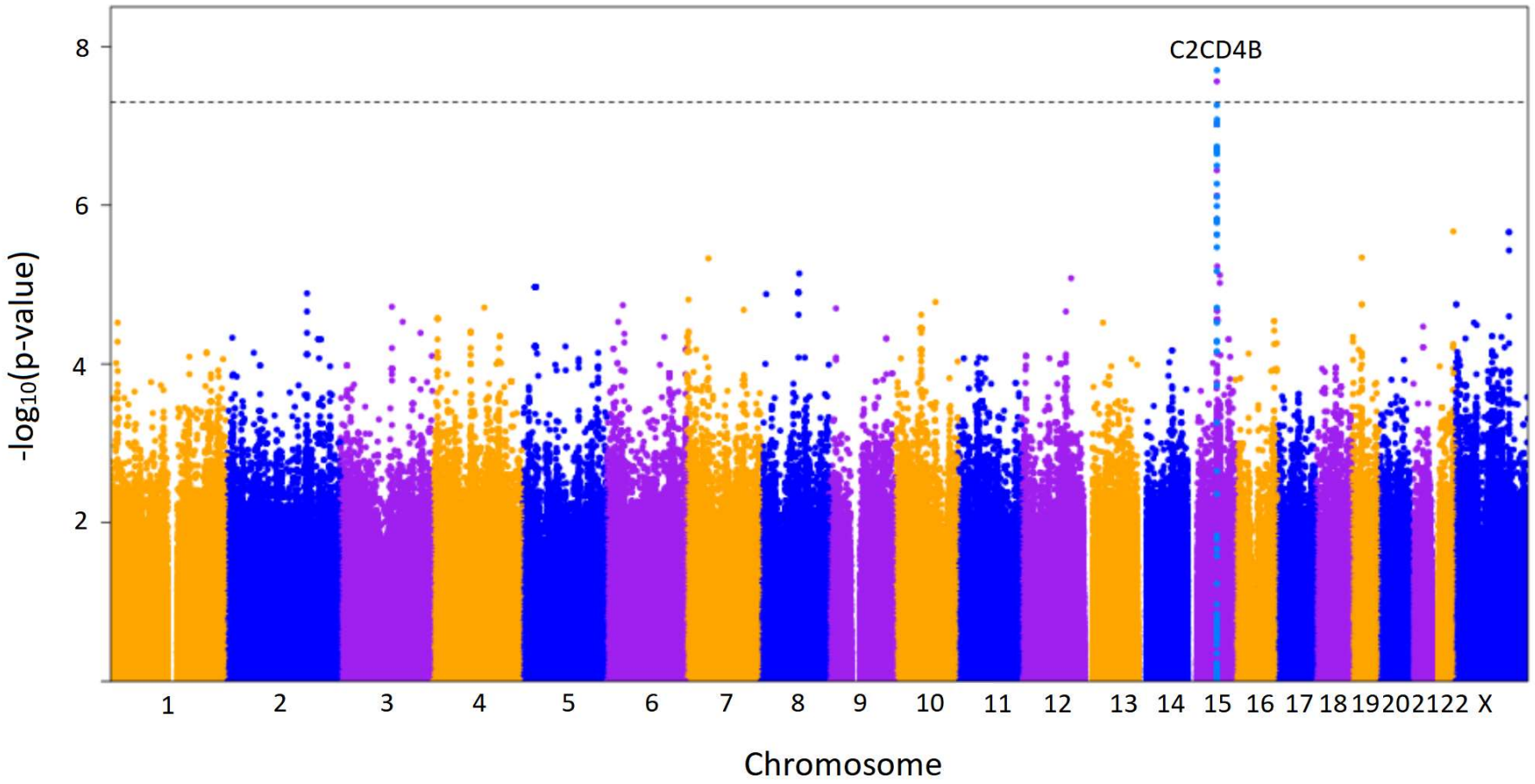


Figure 4.3A – GWAS results for VPC trait

Here we observe an association signal between the biopsy 2 VPC trait and a variant within the C2CD4B region (p-value = $2.01x10^{-8}$), which has been previously implicated in diabetes risk.

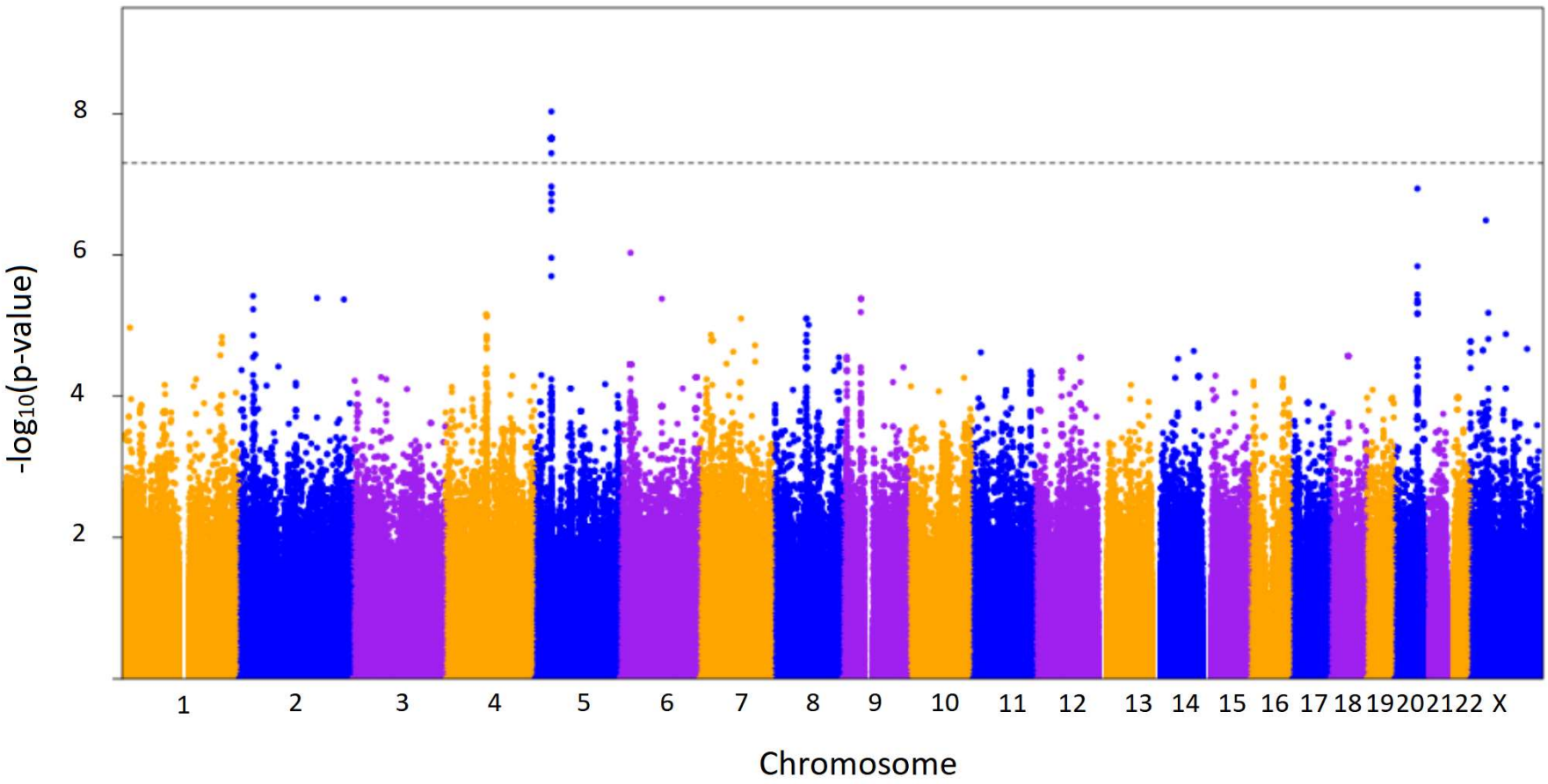


Figure 4.3B – GWAS results for PC3 composite trait

With our composite traits, we observe a significant association between biopsy 2 PC3, (which corresponds to SV, uACR,VVPC and NVPC) and a variant located nearby the AGXT2 gene, which may regulate blood pressure.

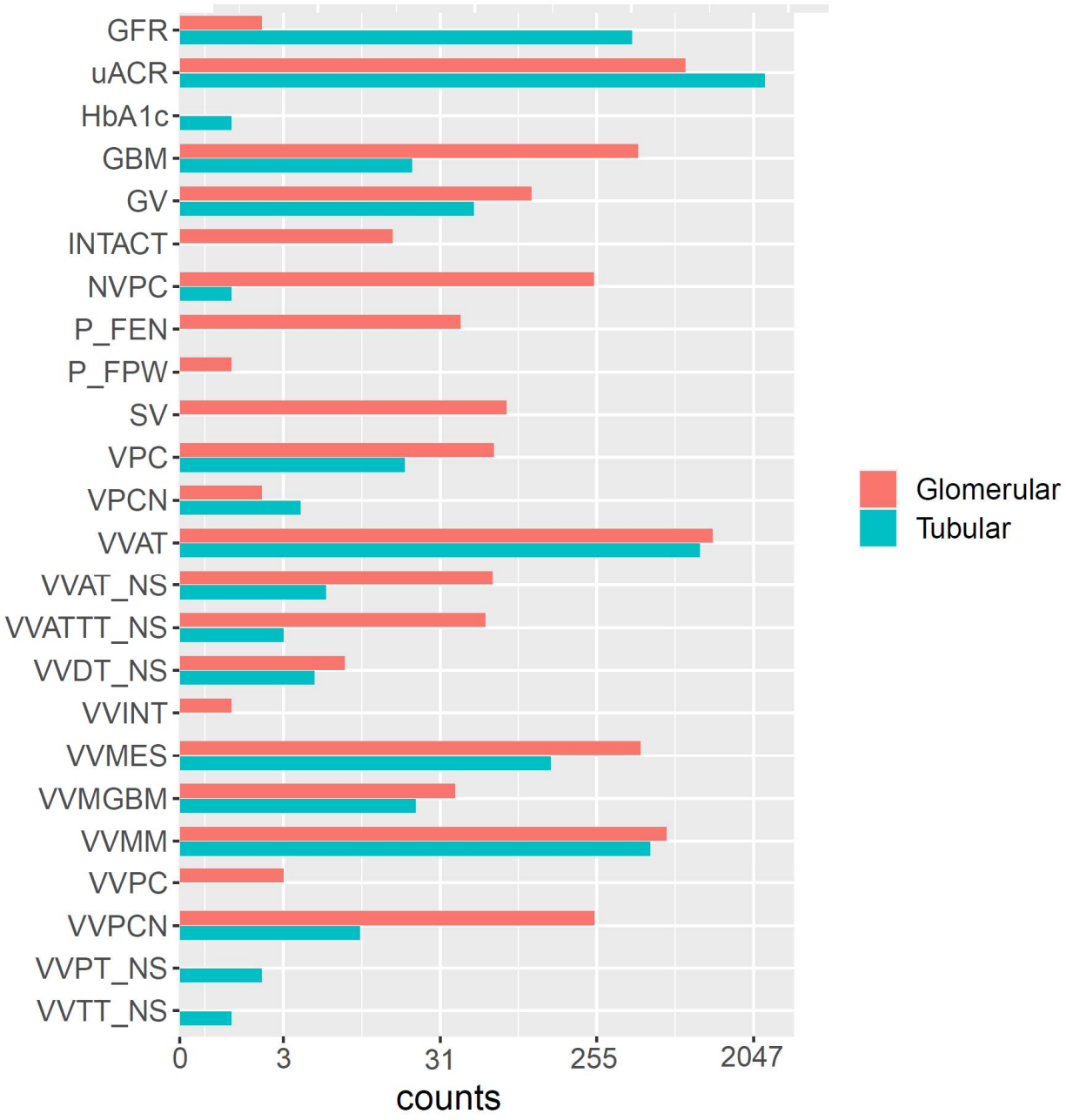


Figure 4.4 –genes associated with clinical/morphometric traits (exponential scale)

Here, we present the number of genes associated with clinical and morphometric traits. We performed this analysis on B1G, B2G, B2GR, B1T, B2T, and B2TR. Counts were pooled across platforms, and ascertained using false discovery rate of 0.05.

Tissue	Trait Name	Trait Description	# DE Genes	Pathway	Adjusted p-value	Fold-Enrichment
Glomerular	GBM	glomerular basement membrane width	365	Pathways in cancer	$1.0x10^{-06}$	3.4
				ECM-recepter interaction	$3.1x10^{-06}$	8.6
Glomerular	NVPC	Numerical density of podocyte cell per glomerulus	245	Focal adhesion	$2.2x10^{-04}$	5.9
Glomerular	uACR	urine albumin creatinine ratio	778	Cytokine-cytokine receptor interaction	$1.3x10^{-04}$	3.0
Glomerular	VVAT_NS	fraction of cortex that is atrophic tubules	63	T cell receptor signaling pathway	$1.0x10^{-03}$	17.8
				Chemokine signaling pathway	$1.0x10^{-03}$	11.3
				Cytokine-cytokine receptor interaction	$1.0x10^{-03}$	8.6
Glomerular	VVATT_NS	fraction of total tubules that are atrophic tubules	57	T cell receptor signaling pathway	$4.9x10^{-04}$	19.9
				Chemokine signaling pathway	$4.9x10^{-04}$	12.6
				Cytokine-cytokine receptor interaction	$4.9x10^{-04}$	9.6
Glomerular	VVMES	mesangial fractional volume	395	ECM-receptor interaction	$8.7x10^{-04}$	7.1
				Pathways in cancer	$8.7x10^{-04}$	2.8
Glomerular	VVMM	volume fraction of mesangial matrix per glomerulus	552	Complement and coagulation cascades	$9.1x10^{-07}$	8.5
Glomerular	VVPCN	volume fraction of podocyte nuclei per podocyte cell	246	Lysosome	$8.8x10^{-07}$	10.0
				Protein processing in endoplasmic reticulum	$1.9x10^{-06}$	7.8
				Ribosome	$4.1x10^{-06}$	7.8
				Prion diseases	$2.1x10^{-04}$	17.0
				Spliceosome	$2.8x10^{-04}$	6.7

Table 4.4A – pathway analysis of differentially expressed Glomerular genes

Tissue	Trait Name	Trait Description	# DE Genes	Pathway	Adjusted p-value	Fold-Enrichment
Tubular	uACR	urine albumin creatinine ratio	1,969	Cytokine-cytokine receptor interaction	$3.7x10^{-15}$	3.5
				Osteoclast differentiation	$4.5x10^{-15}$	5.7
				Pertussis	$4.2x10^{-09}$	5.7
				Hematopoietic cell lineage	$2.6x10^{-08}$	4.6
				TNF signaling pathway	$4.2x10^{-08}$	4.2
Tubular	GFR	glomerular filtration rate	383	Cytokine-cytokine receptor interaction	$4.2x10^{-09}$	5.5
				Fc gamma R-mediated phagocytosis	$1.5x10^{-04}$	7.2
				Leukocyte transendothelial migration	$7.6x10^{-05}$	5.7
				Pertussis	$8.0x10^{-04}$	7.0
				Chemokine signaling pathway	$8.3x10^{-04}$	4.2
Tubular	VVMM	volume fraction of mesangial cells per glomerulus	490	Cytokine-cytokine receptor interaction	$2.3x10^{-16}$	6.7
				Chemokine signaling pathway	$5.3x10^{-06}$	4.8
				Pertussis	$1.8x10^{-05}$	7.6
				Fc gamma R-mediated phagocytosis	$1.8x10^{-05}$	6.7
				Pathogenic Escherichia coli infection	$3.4x10^{-05}$	9.0

Table 4.4B – pathway analysis of differentially expressed Tubular genes

4.6 Supplementary Materials

4.6.1 PCA on Morphometric and Clinical Phenotypes

We performed principal components analysis jointly on clinical traits and morphometric phenotypes. Here, we used the *prcomp* function in R, where we first inverse-normalized the data, centering around 0. Supplementary Figure 4.2 shows the correlation structure between these phenotypes. The heatmap here suggests that performing GWAS on PCA composite traits may reduce some of the redundancy of the highly correlated traits. The PCA loadings are shown in Supplementary Figure 4.3. For our GWAS, we used the first 10 PCs as they accounted for 90% of the variation explained.

4.6.2 SVDiff normalization of gene expression levels

Our *SVDiff* procedure was designed to account for latent systematic technical effects in the array data. This approach was an extra normalization step applied to expression at the probeset level. Suppose Y is an $n \times m$ matrix with n individuals and m genes. We performed singular value decomposition on this matrix to factor it into the form,

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{*}$$

Where $\boldsymbol{U}$ is an $n \times n$ unitary matrix, $\boldsymbol{\Sigma}$ is an $n \times m$ rectangular diagonal matrix, and $\boldsymbol{V}$ is a $m \times m$ unitary matrix. Next, we subset each matrix, capturing the first four singular value components to re-estimate Y. We defined $\boldsymbol{U_{SV}}$ to be the first 4 columns of $\boldsymbol{U}$, $\boldsymbol{V_{SV}^{*}}$ to be the first 4 rows of $\boldsymbol{V}^{*}$, and $\boldsymbol{\Sigma_{SV}}$ to be a diagonal matrix with the first 4 elements of $\boldsymbol{\Sigma}$. Using these matrix subsets, we defined

$$\boldsymbol{Y_{SV}} = \boldsymbol{U_{SV}}\boldsymbol{\Sigma_{SV}}\boldsymbol{V_{SV}^{*}}$$

Our *SVDiff* corrected expression was defined to be: $\boldsymbol{Y_{SVDiff}} = \boldsymbol{Y} - \boldsymbol{Y_{SV}}$, which we used for our eQTL mapping.

4.7 Supplementary Figures and Tables

		Biopsy 1	Biopsy 2	Combined
Number of Samples		77	60	97
Age at time of biopsy (years)	*mean*	45.94	54.19	-
	sd	10.04	9.37	-
Sex	*males*	22	13	24
	females	55	47	73
BMI (kg/m^2)	*mean*	35.78	35.62	-
	sd	8.34	8.06	-
Diabetes Duration by time of enrolment (years)	*mean*	-	-	10
	sd	-	-	6.21
	min	-	-	2.28
	max	-	-	31.45
GFR (ml/min)	*mean*	145.83	128.39	-
	sd	51.23	47.4	-
HBA1C (%)	*mean*	9.36	9.71	-
	sd	2.05	2.01	-
uACR	*normal (<30mg/g)*	35	32	-
	microalbuminuria (30-299mg/g)	30	17	-
	macroalbuminuria (>300mg/g)	12	11	-

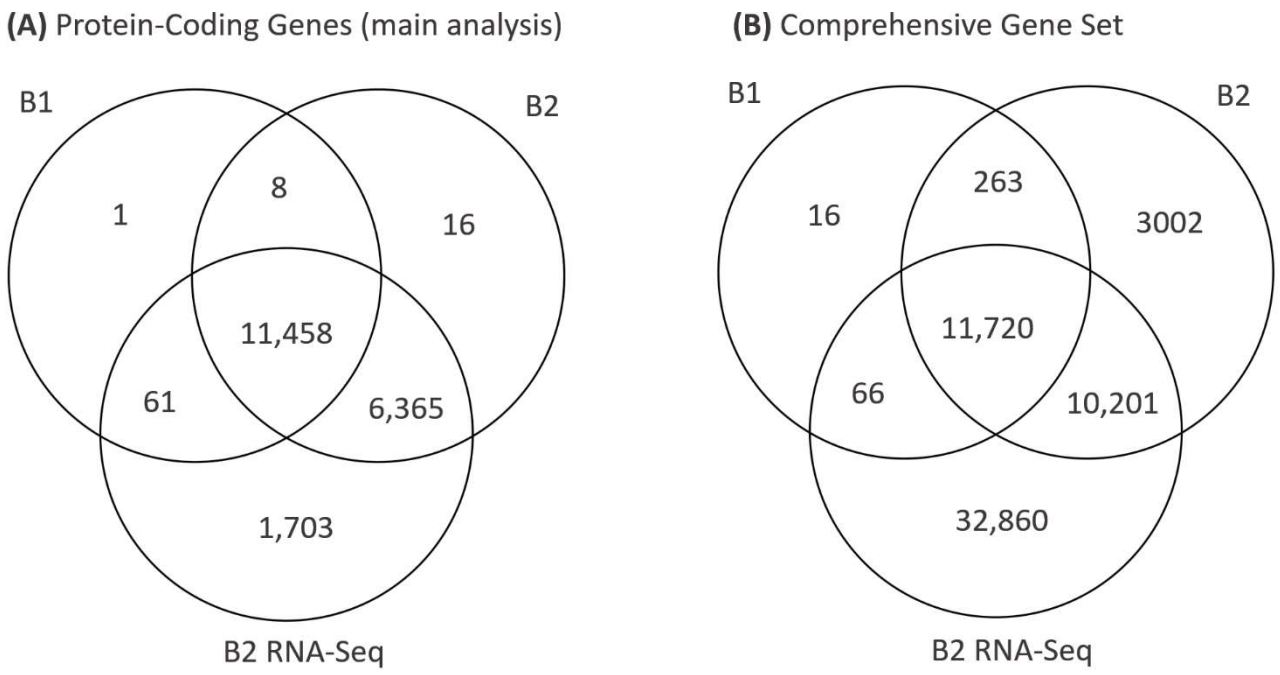
Supplementary Table 4.1 – Demographic information for Pima cohort

Platform	Biopsy	Genome Build	Total Number of Probes	Total Number of Probesets
HGU133A	1	GRCh36	175,294	12,142
HGU133Plus2	1	GRCh36	334,233	19,764
HuGene 2.1 ST	2	GRCh37	466,204	25,583

Supplementary Table 4.2 – Microarray probe information for gene expression measurements

Platform	Biopsy	Total Number of Genes
RNA-seq	2	54,847

Supplementary Table 4.3 – RNA-seq information for gene expression measurements



Supplementary Figure 4.1 – counts of genesets used for expression analysis

We filtered genes using gencode version 27 protein coding genes for our main analysis.

Platform	P-value threshold corresponding to FDR 0.05
B1 Glomerular (microarray)	$1.8x10^{-5}$
B2 Glomerular (microarray)	$1.0x10^{-5}$
B2 Glomerular (RNA-Seq)	$5.5x10^{-6}$
B1 Tubular (microarray)	$2.0x10^{-5}$
B2 Tubular (microarray)	$1.0x10^{-5}$
B2 Tubular (RNA-Seq)	$5.0x10^{-6}$

Supplementary Table 4.4 – P-value thresholds corresponding to FDR of 0.05 for eQTL analyses

For the sake of consistency and simplicity, we used p-value threshold of $5x10^{-6}$ for all of our eQTL analyses to declare significance. Thus, the numbers shown in the main sections for microarray platforms is slightly more conservative than compared to using FDR.

Assay	Tissue-specific genes
B1GLOM-Array	*AAMDC, ALG8, ARPC5, ATG12, ATP5S, BTD, C17orf75, CNN3, CPQ, CTBS, EFCAB2, FLT3LG, METTL22, MGST2, MIER2, NAAA, NAT8, NDUFA6, NUBP2, PPCDC, PRDX1, RAB8A, SKP1, SLC31A2, SLC9A3R2, SPAG7, TCF21, TEFM, TMEM50A, TMEM53, UBE2I, VPS51, WDR45B*
B2GLOM-Array	*C10orf107, DNAJC10, DNAJC15, GPRC5A, HEBP2, HPGD, LMBR1, LRP11, METTL5, NDUFA6, NQO2, ORMDL3, PAPOLA, PRPF40A, RABGEF1, RP1, ST6GALNAC3, TAPBP, VTI1A, ZNF880*
B2GLOM-RNAseq	*BPHL, C10orf107, CCND3, CMTM8, DNAJC15, FBXO25, GNGT1, MRPL34, MRPL53, OR2T6, PCMTD1, RSRC1, SMIM19, TCF4, TCTN3, TMEM150C, TMEM230, TXNRD2, WISP3, YIPF3, ZNF250, ZSWIM7*
B1TUB-Array	*ABCC6, ACOX2, ALDH2, ATP6V1D, BCL2L13, BDH2, BPHL, BTD, COX11, CRYL1, CTSH, DECR2, DNAJC15, DYNC2LI1, DYNLT1, EFCAB2, ENTPD5, GGH, GUSB, HSPBP1, IL17RB, IMPA2, LACTB2, MRPL2, NDUFA6, NUBP2, PBLD, PCBD1, PDSS2, PIGF, RITA1, SKP1, SLC33A1, ST3GAL1, UCHL5*
B2TUB-Array	*ADI1, AGMAT, AGXT2, ASRGL1, CNTNAP3B, DNAJC15, DPYS, EIF4E2, FAH, IL17RB, KL, MMAA, MSH3, NAT8, PRELID1, PTGR2, RNF130, RNF212B, RNF5, SH3YL1, SLC25A26, SLC28A1, SLC6A18, TSPAN33, ZSWIM7*
B2TUB-RNAseq	*ACAD10, ACBD4, ACOT2, ACOX2, AGXT2, ATAD3C, ATP6V0E2, ATXN7L1, BPHL, BTD, C4orf19, CARHSP1, CBWD1, CDPF1, CEP104, CNTNAP3B, COA6, COX7A2L, D2HGDH, DHRS4, DHRS4L2, DNAJC15, DPYS, DTD1, EFCAB2, FAHD1, FOLR1, FUCA2, GSTA1, HIBADH, HSPBP1, IL17RB, KIAA1191, MCCC1, MRPL36, MRPL42, MSRA, MTFMT, MTG2, MUC20, MYL12B, NAT8, NDUFA6, NTPCR, OPA3, PBLD, PCTP, PHYH, PPP2R5C, PSMG4, PSPH, PXMP2, RAB3IP, RABGEF1, RNF212B, SLC22A18AS, SLC44A3, SLC6A18, SMIM19, SMIM7, ST3GAL1, TBCD, THOC3, TIMM10, TIMMDC1, TPRKB, TSTD1, WNT9B, ZSWIM7*

Supplementary Table 4.5 – Genes with novel tissue-specific eQTLs from Pima analysis

Here, we see the list of tissue-specific genes that had Pima eQTLs novel to GTEx version 8. We checked if genes were tissue-specific by approximating their median TPM counts from Pima RNA-seq expression, and comparing it to GTEx median TPM counts. If the Pima median TPM was greater than the 95th percentile GTEx median TPM, then we defined the tissue as tissue-specific to the Pima population.

Assay	Genes with population-specific peak eQTLs
B1GLOM-Array	*ACOX2, BTD, DGCR8, IL17RB, LIPT1, MRPL39, NDUFA6, PGBD5, PPFIA4, SOWAHC, STAT6, VPS51, ZMYND10*
B2GLOM-Array	*ACOX2, CXCL12, DNAJC15, ESD, FXR2, GPATCH8, IL17RB, PPP6C, SLK*
B2GLOM-RNAseq	*BPHL, CTNNB1, CXCL12, DNAJC15, GLCCI1, LAIR2, LILRA6, LIN7B, MIS18BP1, NT5DC4, OSTN, PPP1R3C, SFTPD, SLCO1B3, SLK, SNX15, TXNRD2, UBE2N, UNC5D, ZSWIM1*
B1TUB-Array	*ABCC6, ACOX2, BTD, CNN3, DGCR8, DNAJC15, DYNLT1, ERAP1, EXTL1, GLIPR1, GLP1R, GUF1, IL17RB, MRPL2, NTRK2, PALM, RAPGEF3, TNFSF12, TNFSF15*
B2TUB-Array	*DGKB, DNAJC15, FAH, GSTP1, GUF1, IL17RB, LOXHD1, MSH3, ODF3B, PBX3, TRIM24, TSPAN33*
B2TUB-RNAseq	*AC011479.1, ACAD10, AFMID, AGXT2, ATP6V0E2, ATXN7L1, BPHL, BTD, C17orf58, CFB, CIB2, CLIP4, CPT1A, CTNNB1, D2HGDH, DMAC1, ENPP4, FAHD2A, FAM212B, GNG10, IL17RB, KCNK10, MRPL42, MYL12B, NUDT12, PROB1, RUNX3, SEMA4D, SMN2, TBCD, TMEM163, TRIM24, UTS2, VHL, WDR31, ZNF587B, ZNF787*

Supplementary Table 4.6 – Genes with novel population-specific eQTLs from Pima analysis

Here, we see the list of population-specific genes that had Pima eQTLs novel to GTEx version 8. These are genes corresponding to peak eQTL variants that had minor allele frequency > 20% in the Pima cohort, but minor allele frequency < 5% from Europeans within the 1000 genomes reference panel.

Platform	# eGenes	# Replicates					
		B1GA	B2GA	B2GR	B1TA	B2TA	B2TR
B1GA	435	-	28 (6.4%)	46 (10.6%)	69 (15.9%)	29 (6.7%)	68 (15.6%)
B2GA	178	41 (23.0%)	-	34 (19.1%)	14 (7.9%)	24 (13.5%)	21 (11.8%)
B2GR	351	63 (17.9%)	32 (9.1%)	-	27 (7.7%)	22 (6.3%)	99 (28.2%)
B1TA	315	69 (21.9%)	7 (2.2%)	21 (6.7%)	-	34 (10.8%)	90 (28.6%)
B2TA	285	37 (13.0%)	12 (4.2%)	20 (7.0%)	42 (14.7%)	-	95 (33.3%)
B2TR	814	75 (9.2%)	15 (1.8%)	72 (8.8%)	95 (11.7%)	84 (10.3%)	-

Supplementary Table 4.7 – cis-eQTL replicates using stringent p-value thresholds for both datasets (p-value < $5x10^{-6}$)

This list includes gencode v27 protein coding genes only.

Biopsy	Platform	N samples	# eGenes
B1 Glom	Microarray	69	457
B1 Tub	Microarray	46	325
B2 Glom	Microarray	50	219
B2 Tub	Microarray	54	328
B2 Glom	RNA-Seq	52	583
B2 Tub	RNA-Seq	55	1275
Apex-Glomerular	Combined	93(171)	422
Apex-Tubular	Combined	77(155)	420

Supplementary Table 4.8 – eQTL analysis with full list of genes

This list includes all genes that were assayed. We used p-value threshold of $5x10^{-6}$.

Platform	# eGenes	# Replicates					
		B1GA	B2GA	B2GR	B1TA	B2TA	B2TR
B1GA	457	-	156 (34.1%)	211 (46.2%)	218 (47.7%)	122 (26.7%)	190 (41.6%)
B2GA	219	67 (30.6%)	-	88 (40.2%)	39 (17.8%)	69 (31.5%)	65 (29.7%)
B2GR	583	123 (21.1%)	126 (21.6%)	-	70 (12.0%)	103 (17.7%)	195 (33.4%)
B1TA	325	191 (58.8%)	76 (23.3%)	92 (28.3%)	-	150 (46.2%)	182 (56.0%)
B2TA	328	90 (27.4%)	119 (36.3%)	87 (26.5%)	119 (36.3%)	-	179 (54.6%)
B2TR	1275	244 (19.1%)	207 (16.2%)	365 (28.6%)	300 (23.5%)	394 (30.9%)	-

Supplementary Table 4.9 – cis-eQTL replication with full list of genes (p-value 0.025)

This replication list includes all genes that were assayed.

Platform	# eGenes	# Replicates					
		B1GA	B2GA	B2GR	B1TA	B2TA	B2TR
B1GA	457	-	30 (6.6%)	49 (10.7%)	71 (15.5%)	30 (6.6%)	74 (16.2%)
B2GA	219	43 (19.6%)	-	39 (17.8%)	14 (6.4%)	31 (14.2%)	26 (19.6%)
B2GR	583	68 (11.7%)	37 (6.3%)	-	28 (4.8%)	28 (4.8%)	114 (19.6%)
B1TA	325	72 (22.2%)	7 (2.2%)	22 (6.8%)	-	35 (10.8%)	94 (28.9%)
B2TA	328	39 (11.9%)	17 (5.2%)	23 (7.0%)	43 (13.1%)	-	101 (30.8%)
B2TR	1275	83 (6.5%)	21 (1.6%)	87 (6.8%)	100 (7.8%)	94 (7.4%)	-

Supplementary Table 4.10 – cis-eQTL replication with full list of genes (p-value $5x10^{-6}$)

This replication list includes all genes that were assayed.

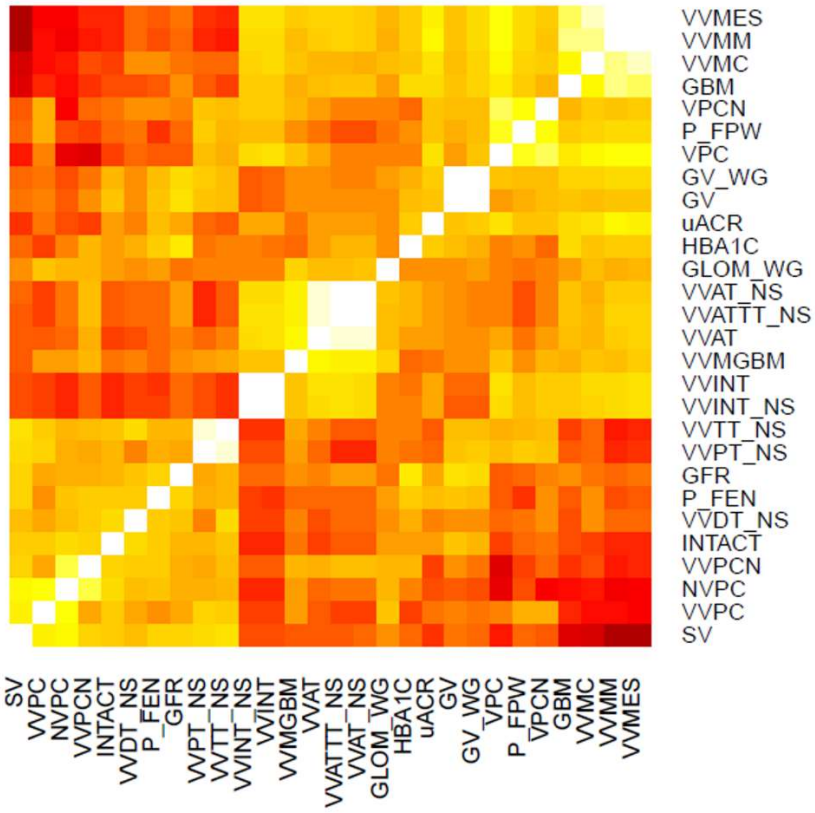
Variable Name	Definition	Unit
GBM	glomerular basement membrane width	nm
VVMES	mesangial fractional volume	%
VVMC	volume fraction of mesangial cells per glomerulus	%
VVMM	volume fraction of mesangial matrix per glomerulus	%
VVMGBM	volume fraction of mesangial glomerular basement membrane per glomerulus	%
SV	surface volume of peripheral glomerular basement membrane per glomerulus	$\mu m^2/\mu m^3$
FPW_UM	foot process width per biopsy from UMinn	nm
VPCN	volume of podocyte nuclei	μm^3
VVPCN	volume fraction of podocyte nuclei per podocyte cell	%
VPC	volume of podocyte cell	μm^3
VVPC	volume fraction of podocyte cell per glomerulus	%
NVPC	numerical density of podocyte cell per glomerulus	N/glom
GLOM_CAV	number of glomeruli measured for GV_CAV	N
GV_CAV	glomerular volume by Cavaleri method on paraffin sections	$x10^6\ \mu m^3$
GLOM_WG	number of glomeruli measured for GV_WG	N
GV_WG	glomerular volume by Weibel-Gomez method on epon section	$x10^6\ \mu m^3$
P_FPW	foot process width in peripheral glomerular basement membrane	nm
INTACT	percent of intact foot processes on both the peripheral and mesangial glomerular basement membrane	%
P_FEN	percent of endothelial fenestration falling on the peripheral glomerular basement membrane	%
VVINT	cortical interstitial fractional volume	%
VVINT_NS	fraction of cortex that is interstitium	%
VVINT_S	fraction of scar cortex that is interstitium	%
VVSCAR	fraction of total cortex that is scar cortex	%
VVAT	fraction of total cortex that is total atrophic tubules	%
VVAT_NS	fraction of cortex that is atrophic tubules	%
VVATTT_NS	fraction of total tubules that is atrophic tubules	%
VVPT_NS	fraction of cortex that is proximal tubules	%
VVDT_NS	fraction of cortex that is distal tubules	%
VVTT_NS	fraction of cortex that is total tubules (proximal, distal, atrophic)	%
GV	derived glomerular volume variable - using GV_CAV when present and GV_WG when GV_CAV is missing	$x10^6\ \mu m^3$

Supplementary Table 4.11 – description of all kidney morphometry traits

Trait	Glomerular signals	Tubular signals
GBM	365	21
GFR	2	383
GV	105	49
HbA1c	0	1
INTACT	16	0
NVPC	245	1
P_FEN	41	0
P_FPW	1	0
SV	66	0
uACR	778	1969
VPC	64	19
VPCN	2	4
VVAT	1147	942
VVAT_NS	63	6
VVATTT_NS	57	3
VVDT_NS	8	5
VVINT	1	0
VVMES	395	136
VVMGBM	38	22
VVMM	552	490
VVPC	3	0
VVPCN	246	10
VVPT_NS	0	2
VVTT_NS	0	1

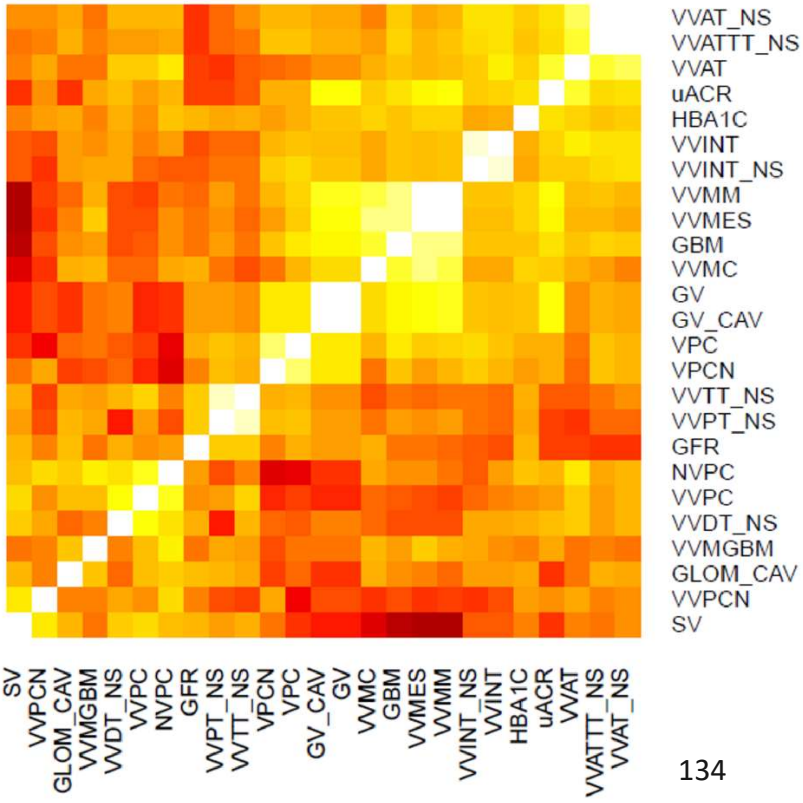
Supplementary Table 4.12 – counts of differentially expressed genes for each clinical/morphometric trait

(A) Biopsy 1 (n=77)

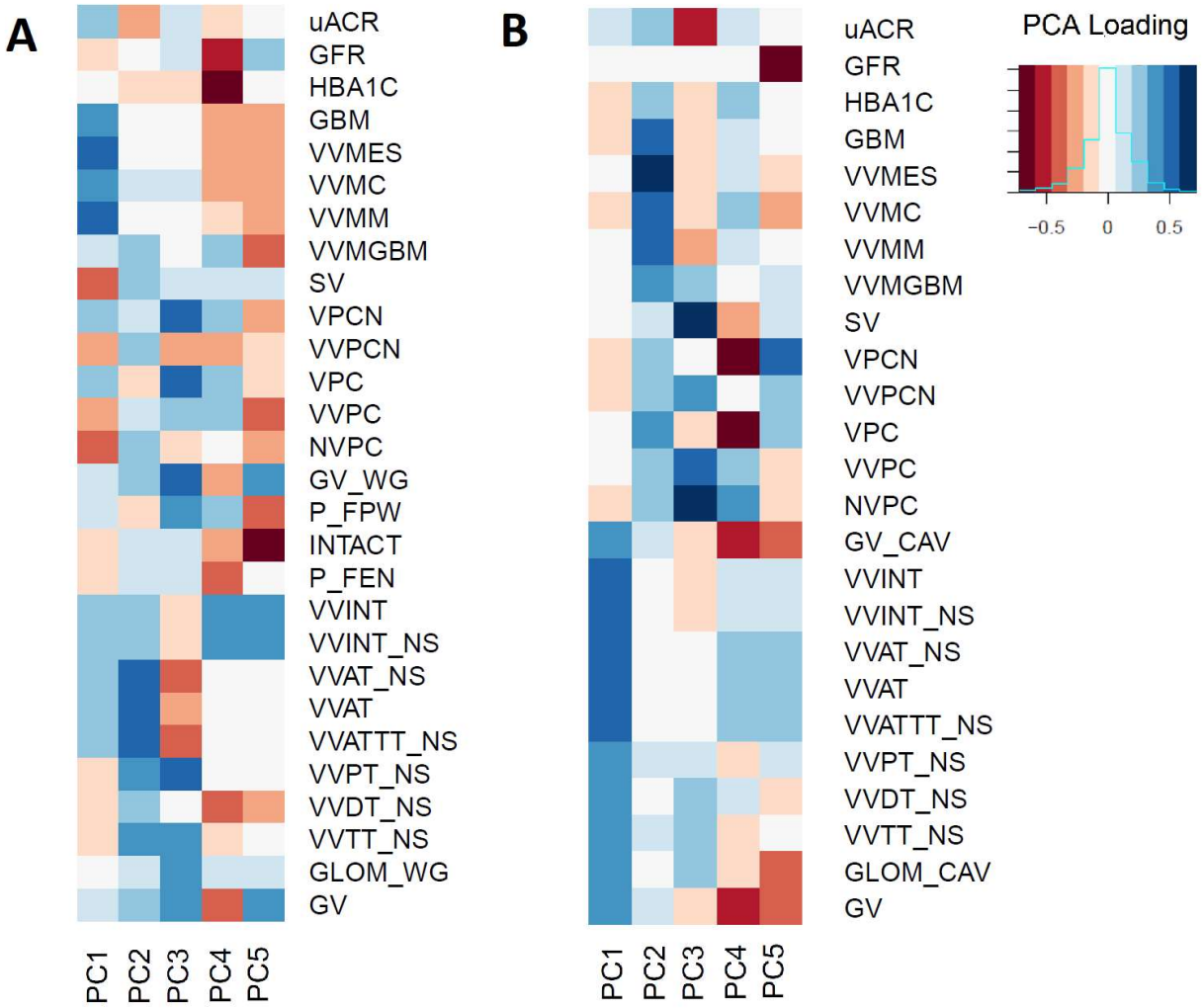


Supplementary Figure 4.2A – Phenotype correlation structure between biopsy 1 clinical and morphometric traits

(B) Biospy 2 (n=60)



Supplementary Figure 4.2B – Phenotype correlation structure between biopsy 2 clinical and morphometric traits



Supplementary Figure 4.3 – PCA loadings for clinical and morphometric traits

(A) shows the PC loadings for biopsy 1 traits and (B) shows the PC loadings for biopsy 2 traits

Chapter 5 Conclusion

5.1 Summary

The study of genetics is an integral part to understanding the biology behind our complex traits and can be approached in a variety of ways. Systems genetics studies across many genomes, transcriptomes, epigenomes, and phenomes provide us opportunities to elucidate the functional mechanisms of trait-associated variants in terms of gene regulation or protein function. In this thesis, we addressed on specific challenges related to systems genetic studies, including meta-imputation of expression across multiple datasets with only summary-level imputation models available, correcting for technical biases towards reference alleles in array-based expression assays, and identifying tissue-specific and population-specific regulatory variants and trait-associated loci in the context of systems genetics with whole genome sequencing, transcriptomics profiles, morphometric traits, and clinical endpoints. With increasing number of sequenced genomes and trait-associated variants identified, it will be increasingly important to interpret each association signal through systems genetics approach that leverages molecular traits as intermediate phenotypes. In this chapter, I will summarize each chapter and describe the current limitations in the methods and results described in each chapter and discuss future directions to expand their scope of research.

In Chapter 2, we developed a method which leverages multiple datasets to accurately impute tissue-specific gene expression levels. Our method, Smartly Weighted Averaging across Multiple Tissues (SWAM) does not train directly from data, but rather combines extant prediction models by assigning weights based on their predictive performance and similarity to the tissue of interest. We demonstrate that when using the same set of resources (GTEx version 6 tissues), SWAM improves prediction accuracy compared to approaches which predict gene

expression by training directly from raw data (PrediXcan, UTMOST). However, the major benefit of using the SWAM meta-imputation framework is the flexibility to combine multiple external resources derived from disjoint sets of individuals. Indeed, prediction accuracy is substantially improved when integrating whole-blood information from DGN samples into the GTEx-only predictions, highlighting the importance of using multiple datasets.

In Chapter 3, we revisit a well-known hybridization bias that arises in microarray studies caused by genetic polymorphisms within target probe sequences. In our work, we leverage the availability of whole genome sequencing data to accurately identify and characterize this bias at both the probe and probeset (gene) level. We adjust gene expression level calculations by removing probes which overlap with study-specific polymorphisms and demonstrate that this approach resolves the negative bias more effectively compared to when using a reference panel to identify probes. We then propose an imputation method in which probes are not removed, but rather intensity levels are imputed based on values of unaffected probes within their probeset. This method was proposed to address the issue of unnecessarily removing probes ultimately reducing accuracy of expression calculation. This approach results in higher concordance between pre- and post- correction for many genes where probes were only mildly affected, while also reducing the negative bias in eQTL analysis.

In Chapter 4, we perform a systems genetic study of Pima Native Americans enrolled in a diabetic nephropathy study. We integrate whole genome sequence data, gene expression and morphometric features derived from two microdissected renal compartments – Glomerular and Tubular – and clinical measurements to provide a landscape of the transcriptomic and genomic profiles of this cohort. Because of the high dimensionality of these datasets, we used various dimension reduction techniques such as PCA and TWAS to reduce the multiple testing burden and increase the chances of signal discovery. Studying this unique population gave us the ability to identify many population-specific and tissue-specific regulatory variants, as well as link various expressed genes with downstream clinical and morphometric traits.

5.2 Meta-imputation of gene expression using summary-level eQTL databases

In our second chapter, we developed a method to leverage multiple datasets to accurately impute gene expression levels. Previously, gene expression imputation models were derived by performing (penalized) regression between genetic markers and measured expression levels, requiring raw individual-level data. Our method, Smartly Weighted Averaging across Multiple Tissues (SWAM) combines these derived prediction models by assigning weights to each model based on their similarity to the tissue of interest. Because SWAM does not train directly from raw data, prediction models derived from disjoint sets of individuals can be combined. The ability to leverage multiple datasets effectively increases the sample size of prediction model building, compared to using only a single resource. We demonstrate that SWAM provides superior imputation accuracy when combining multiple tissues from the same cohort (GTEx) compared to methods that train directly from raw data. In addition, imputation accuracy can be improved further when integrating other external resources, such as adding whole-blood tissue from DGN with the GTEx tissues.

We compared SWAM to other expression imputation methods in the context of TWAS, testing three traits (HDL, LDL and type-2 diabetes). In terms of power, SWAM discovered more trait-associations than other methods that generate imputed expression levels. However, we are aware that there are other methods which do not impute expression, but directly conduct TWAS using summary-level aggregate information from multiple resources, such as MultiXcan and S-MultiXcan. These multi-dataset TWAS methods do outperform SWAM in terms of signal discovery for the traits we tested. However, expression predictions can be useful outside of the context of TWAS such as for mendelian randomization experiments. Overall, among methods that produce expression predictions, we found SWAM to have the highest accuracy when validating with an external resource, and to have the highest power for TWAS discovery. Because multiple resources can be integrated without the need for their raw data, SWAM will be able to take advantage of the increasing number of eQTL resources being generated.

Because many current eQTL resources come from individuals with European ancestry, we validated our results with only the European samples from the GEUVADIS consortium.

However, there could be many population specific variants that regulate expression levels that may only be detected by studying the correct ethnic groups. In addition, identifying these variants could provide the opportunity to identify the true causal variant for other populations by disentangling associations caused by LD structures. When comparing to the African population from GEUVADIS, we found much lower concordance between measured and imputed expression. This trans-ethnic or multi-ethnic aspect of imputation is a challenging problem, particularly so given that we are not training using the full raw data. Our validation of SWAM therefore operates under the assumption that the population of interest matches the training population. One future direction could be to examine the ideal strategy to integrate resources for multi-population or population specific imputation.

With the emerging availability of new gene expression resources, providing the tools for massive integration is very important. In the future I would like to eventually implement a web-based version of SWAM, similar to imputation servers to facilitate this integration. This could simplify and streamline much of the imputation process at a large scale.

Finally, as single-cell RNA-Seq is becoming the pre-eminent technology for gene expression studies, we are able to view transcriptomic profiles at a very high scale and resolution. There have already been many efforts to analyze this high dimensional data by distinguishing and calculating expression levels for different cell types [203–205]. However, technical noise and cell dropouts have also given rise to the need for imputation at the cellular level [206–208]. Since SWAM is already suited to integrate multiple datasets even with different tissue types, our method could be extended to scRNA-Seq, treating different cell types as separate “tissues”. This could potentially enrich imputations and also allow for TWAS conducted for scRNA-Seq.

5.3 Revising array-based expression profiles to empower today’s systems genetics

In the third chapter, we revisited the well-known negative hybridization bias in microarray studies while leveraging the availability of whole genome sequencing. Because microarray

probes are designed to target specific DNA sequences within gene regions, genetic polymorphisms within individuals can weaken the bond between the RNA molecules and probe. This in turn creates a false association between the genotypes and probe intensity, which leads to downstream false positives in eQTL studies, while also hiding true associations. In previous work, the strategy was to identify these probes by using polymorphic sites from reference panels (1000G), but this could incorrectly flag many probes if the population is different from the individuals in the panel. In our work, we use individual-level whole genome sequence data to accurately identify the list of potential biased probes, removing them from the expression calculation. We demonstrate that this approach reduces bias more effectively compared to using 1000G common variants. However, our characterization of flagged probes also revealed that some probes are greatly affected (high technical bias) whereas others are only mildly affected by the SNP-in-probe effect. Because removing probes unnecessarily can potentially add noise to the expression estimates, we devised an imputation method where we use the unaffected probes within a probeset to help estimate the true unbiased intensity for the affected probe. We demonstrate that this imputation method provides similar levels of efficacy in terms of bias correction compared to removing biased probes, while also having higher concordance with original expression levels for genes with only mildly affected probes.

Because of their design, microarrays will always be susceptible to hybridization biases as they require DNA sequences. However, with increasing knowledge of the human genome, modern microarrays can somewhat overcome these weakened hybridization biases by avoiding known polymorphic sites in their probe targets. The work done in this dissertation should be taken with caution as it was done using the HuGene 2.1 ST array, and may not be generalizable to the newer microarray platforms. Furthermore, the study population was also limited to Pima Native Americans, and different populations may be impacted to varying degrees depending on their population-specific variants. One possible future direction would be to characterize the list of probes as well as the extent of bias in said probes for different microarray platforms and different populations. Making publicly available resources such as these could aid future researchers in correcting biases without the need for reference panels.

5.4 Systems genetic study on Pima diabetic nephropathy cohort

In chapter 4, we presented a systems genetics study of the Pima diabetic nephropathy cohort, where we analyzed whole genome sequence data, transcriptomic profiles, clinical phenotypes, and kidney morphometry. Compared the currently published kidney transcriptome resources, this study was unique in the sense that it was one of the few with microdissected renal tissue compartments (as opposed to bulk kidney cortex) and focuses on an underrepresented population. As such, studying this cohort provided the opportunity to both replicate, and discover novel tissue-specific and population-specific insights into biological pathways underlying kidney disease. For example, our eQTL mapping discovered 805 glomerular eGenes and 1,118 tubular eGenes, with roughly 50% replicated in GTEx, and with a plausible 129 novel tissue-specific and 64 novel population-specific genes not previously identified in GTEx. In addition to eQTL mapping, we discovered numerous renal disease-relevant biological pathways for genes significantly associated with clinical and morphometric traits, including cytokine-cytokine receptor interactions, focal adhesion, cancer, and ECM-receptor interactions [180–186]. Our GWAS – despite the small sample size – discovered variants associated with the important VPC trait (volume of podocyte counts) within the *C2CD4B* gene region, which has been previously implicated in diabetes risk [167,168].

Despite the many resources generated from this study, there are limitations and challenges that arose from this deep, complex dataset that could be addressed as future directions. For example, although the study conducted biopsies at two time points, the assaying of expression levels was done in separate batches, thereby confounding the time effect with the batch effect. As such, because it is highly improbable to disentangle these two effects, discovering differentially expressed genes related to disease progression is likely infeasible. One potential future direction could be to use the RNA-Sequencing data (which coincides with the microarray expression from biopsy 2) to assist in separating out the batch and time effects. However, this was not within the scope of the project. Another limitation of this study was that, in order to obtain high quality phenotypes and transcriptomic profiles, the sample size of the cohort was relatively low, particularly in the context of GWAS, which can often have orders of magnitude

higher sample sizes [194]. Moving forward, integration of the transcriptomic facets of this study into larger scale GWA studies could provide further insight into the biological pathways underlying renal disease. Finally, with the advent of single-cell sequencing resources, we may able to deconvolute our gene expression into cell-type-specific resolution [209] to interpret eQTLs in a more precise manner.

5.5 Concluding remarks

The rapid advancement of technology, particularly in computing in the 20th and early 21st century has ushered in an era of revolutionary accessibility to information. With these changes to how quickly we receive and process data, we have been able to delve into various scientific topics with both unprecedented breadth and depth. The field of genetics has been rapidly evolving as technology improves, and with these improvements come new challenges, in both wetlab settings, as well as statistical and computational. The work done in this thesis has addressed some of the challenges in the field of genomics, such as implementing an integrative framework to combine multiple external resources to predict gene expression, correcting for technical biases in older gene expression technologies, and generating resources for emerging population-specific and tissue-specific systems genetics studies. It has been wonderful and fulfilling to contribute but a small part to our already vast expanse of knowledge in the field of genetics. As our understanding of genetics, biology, and science continue to grow, I look forward to continuing to work towards the improvement of public health by helping push the boundaries of our understanding of genetics, biology, and medicine.

References

1. Civelek M, Lusis AJ. Systems genetics approaches to understand complex traits. Nat Rev Genet. 2014 Jan;15(1):34–48.

2. Abbott S, Fairbanks DJ. Experiments on Plant Hybrids by Gregor Mendel. Genetics. 2016 Oct;204(2):407–22.

3. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, Pendlington ZM, Welter D, Burdett T, Hindorff L, Flicek P, Cunningham F, Parkinson H. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017 04;45(D1):D896–901.

4. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, Suveges D, Vrousgou O, Whetzel PL, Amode R, Guillen JA, Riat HS, Trevanion SJ, Hall P, Junkins H, Flicek P, Burdett T, Hindorff LA, Cunningham F, Parkinson H. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019 08;47(D1):D1005–12.

5. Zimmermann MT. The Importance of Biologic Knowledge and Gene Expression Context for Genomic Data Interpretation. Front Genet. 2018 Dec 18;9:670.

6. The Multiple Tissue Human Expression Resource (MuTHER) Consortium, Grundberg E, Small KS, Hedman ÅK, Nica AC, Buil A, Keildson S, Bell JT, Yang T-P, Meduri E, Barrett A, Nisbett J, Sekowska M, Wilk A, Shin S-Y, Glass D, Travers M, Min JL, Ring S, Ho K, Thorleifsson G, Kong A, Thorsteindottir U, Ainali C, Dimas AS, Hassanali N, Ingle C, Knowles D, Krestyaninova M, Lowe CE, Di Meglio P, Montgomery SB, Parts L, Potter S, Surdulescu G, Tsaprouni L, Tsoka S, Bataille V, Durbin R, Nestle FO, O'Rahilly S, Soranzo N, Lindgren CM, Zondervan KT, Ahmadi KR, Schadt EE, Stefansson K, Smith GD, McCarthy MI, Deloukas P, Dermitzakis ET, Spector TD. Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat Genet. 2012 Oct;44(10):1084–9.

7. Nica AC, Parts L, Glass D, Nisbet J, Barrett A, Sekowska M, Travers M, Potter S, Grundberg E, Small K, Hedman ÅK, Bataille V, Tzenova Bell J, Surdulescu G, Dimas AS, Ingle C, Nestle FO, di Meglio P, Min JL, Wilk A, Hammond CJ, Hassanali N, Yang T-P, Montgomery SB, O'Rahilly S, Lindgren CM, Zondervan KT, Soranzo N, Barroso I, Durbin R, Ahmadi K, Deloukas P, McCarthy MI, Dermitzakis ET, Spector TD, The MuTHER Consortium. The Architecture of Gene Regulatory Variation across Multiple Human Tissues: The MuTHER Study. Barsh G, editor. PLoS Genet. 2011 Feb 3;7(2):e1002003.

8. Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU, Prokunina-Olsson L, Ding C-J, Swift AJ, Narisu N, Hu T, Pruim R, Xiao R, Li X-Y, Conneely KN, Riebow NL, Sprau AG, Tong M, White PP, Hetrick KN, Barnhart MW, Bark CW, Goldstein JL, Watkins L, Xiang F, Saramies J, Buchanan TA, Watanabe RM, Valle TT, Kinnunen L, Abecasis GR, Pugh EW, Doheny KF, Bergman RN, Tuomilehto J, Collins FS, Boehnke M. A Genome-Wide Association Study of Type 2 Diabetes in Finns Detects Multiple Susceptibility Variants. Science. 2007 Jun 1;316(5829):1341–5.

9. The LifeLines Cohort Study, The ADIPOGen Consortium, The AGEN-BMI Working Group, The CARDIOGRAMplusC4D Consortium, The CKDGen Consortium, The GLGC, The ICBP, The MAGIC Investigators, The MuTHER Consortium, The MIGen Consortium, The PAGE Consortium, The ReproGen Consortium, The GENIE Consortium, The International Endogene Consortium, Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, Powell C, Vedantam S, Buchkovich ML, Yang J, Croteau-Chonka DC, Esko T, Fall T, Ferreira T, Gustafsson S, Kutalik Z, Luan J, Mägi R, Randall JC, Winkler TW, Wood AR, Workalemahu T, Faul JD, Smith JA, Hua Zhao J, Zhao W, Chen J, Fehrmann R, Hedman ÅK, Karjalainen J, Schmidt EM, Absher D, Amin N, Anderson D, Beekman M, Bolton JL, Bragg-Gresham JL, Buyske S, Demirkan A, Deng G, Ehret GB, Feenstra B, Feitosa MF, Fischer K, Goel A, Gong J, Jackson AU, Kanoni S, Kleber ME, Kristiansson K, Lim U, Lotay V, Mangino M, Mateo Leach I, Medina-Gomez C, Medland SE, Nalls MA, Palmer CD, Pasko D, Pechlivanis S, Peters MJ, Prokopenko I, Shungin D, Stančáková A, Strawbridge RJ, Ju Sung Y, Tanaka T, Teumer A, Trompet S, van der Laan SW, van Setten J, Van Vliet-Ostaptchouk JV, Wang Z, Yengo L, Zhang W, Isaacs A, Albrecht E, Ärnlöv J, Arscott GM, Attwood AP, Bandinelli S, Barrett A, Bas IN, Bellis C, Bennett AJ, Berne C, Blagieva R, Blüher M, Böhringer S, Bonnycastle LL, Böttcher Y, Boyd HA, Bruinenberg M, Caspersen IH, Ida Chen Y-D, Clarke R, Warwick Daw E, de Craen AJM, Delgado G, Dimitriou M, Doney ASF, Eklund N, Estrada K, Eury E, Folkersen L, Fraser RM, Garcia ME, Geller F, Giedraitis V, Gigante B, Go AS, Golay A, Goodall AH, Gordon SD, Gorski M, Grabe H-J, Grallert H, Grammer TB, Gräßler J, Grönberg H, Groves CJ, Gusto G, Haessler J, Hall P, Haller T, Hallmans G, Hartman CA, Hassinen M, Hayward C, Heard-Costa NL, Helmer Q, Hengstenberg C, Holmen O, Hottenga J-J, James AL, Jeff JM, Johansson Å, Jolley J, Juliusdottir T, Kinnunen L, Koenig W, Koskenvuo M, Kratzer W, Laitinen J, Lamina C, Leander K, Lee NR, Lichtner P, Lind L, Lindström J, Sin Lo K, Lobbens S, Lorbeer R, Lu Y, Mach F, Magnusson PKE, Mahajan A, McArdle WL, McLachlan S, Menni C, Merger S, Mihailov E, Milani L, Moayyeri A, Monda KL, Morken MA, Mulas A, Müller G, Müller-Nurasyid M, Musk AW, Nagaraja R, Nöthen MM, Nolte IM, Pilz S, Rayner NW, Renstrom F, Rettig R, Ried JS, Ripke S, Robertson NR, Rose LM, Sanna S, Scharnagl H, Scholtens S, Schumacher FR, Scott WR, Seufferlein T, Shi J, Vernon Smith A, Smolonska J, Stanton AV, Steinthorsdottir V, Stirrups K, Stringham HM, Sundström J, Swertz MA, Swift AJ, Syvänen A-C, Tan S-T, Tayo BO, Thorand B, Thorleifsson G, Tyrer JP, Uh H-W, Vandenput L, Verhulst FC, Vermeulen SH, Verweij N, Vonk JM, Waite LL, Warren HR, Waterworth D, Weedon MN, Wilkens LR, Willenborg C, Wilsgaard T, Wojczynski MK, Wong A, Wright AF, Zhang Q, Brennan EP, Choi M, Dastani Z, Drong AW, Eriksson P, Franco-Cereceda A, Gådin JR, Gharavi AG, Goddard ME,

Handsaker RE, Huang J, Karpe F, Kathiresan S, Keildson S, Kiryluk K, Kubo M, Lee J-Y, Liang L, Lifton RP, Ma B, McCarroll SA, McKnight AJ, Min JL, Moffatt MF, Montgomery GW, Murabito JM, Nicholson G, Nyholt DR, Okada Y, Perry JRB, Dorajoo R, Reinmaa E, Salem RM, Sandholm N, Scott RA, Stolk L, Takahashi A, Tanaka T, van't Hooft FM, Vinkhuyzen AAE, Westra H-J, Zheng W, Zondervan KT, Heath AC, Arveiler D, Bakker SJL, Beilby J, Bergman RN, Blangero J, Bovet P, Campbell H, Caulfield MJ, Cesana G, Chakravarti A, Chasman DI, Chines PS, Collins FS, Crawford DC, Adrienne Cupples L, Cusi D, Danesh J, de Faire U, den Ruijter HM, Dominiczak AF, Erbel R, Erdmann J, Eriksson JG, Farrall M, Felix SB, Ferrannini E, Ferrières J, Ford I, Forouhi NG, Forrester T, Franco OH, Gansevoort RT, Gejman PV, Gieger C, Gottesman O, Gudnason V, Gyllensten U, Hall AS, Harris TB, Hattersley AT, Hicks AA, Hindorff LA, Hingorani AD, Hofman A, Homuth G, Kees Hovingh G, Humphries SE, Hunt SC, Hyppönen E, Illig T, Jacobs KB, Jarvelin M-R, Jöckel K-H, Johansen B, Jousilahti P, Wouter Jukema J, Jula AM, Kaprio J, Kastelein JJP, Keinanen-Kiukaanniemi SM, Kiemeney LA, Knekt P, Kooner JS, Kooperberg C, Kovacs P, Kraja AT, Kumari M, Kuusisto J, Lakka TA, Langenberg C, Le Marchand L, Lehtimäki T, Lyssenko V, Männistö S, Marette A, Matise TC, McKenzie CA, McKnight B, Moll FL, Morris AD, Morris AP, Murray JC, Nelis M, Ohlsson C, Oldehinkel AJ, Ong KK, Madden PAF, Pasterkamp G, Peden JF, Peters A, Postma DS, Pramstaller PP, Price JF, Qi L, Raitakari OT, Rankinen T, Rao DC, Rice TK, Ridker PM, Rioux JD, Ritchie MD, Rudan I, Salomaa V, Samani NJ, Saramies J, Sarzynski MA, Schunkert H, Schwarz PEH, Sever P, Shuldiner AR, Sinisalo J, Stolk RP, Strauch K, Tönjes A, Trégouët D-A, Tremblay A, Tremoli E, Virtamo J, Vohl M-C, Völker U, Waeber G, Willemsen G, Witteman JC, Carola Zillikens M, Adair LS, Amouyel P, Asselbergs FW, Assimes TL, Bochud M, Boehm BO, Boerwinkle E, Bornstein SR, Bottinger EP, Bouchard C, Cauchi S, Chambers JC, Chanock SJ, Cooper RS, de Bakker PIW, Dedoussis G, Ferrucci L, Franks PW, Froguel P, Groop LC, Haiman CA, Hamsten A, Hui J, Hunter DJ, Hveem K, Kaplan RC, Kivimaki M, Kuh D, Laakso M, Liu Y, Martin NG, März W, Melbye M, Metspalu A, Moebus S, Munroe PB, Njølstad I, Oostra BA, Palmer CNA, Pedersen NL, Perola M, Pérusse L, Peters U, Power C, Quertermous T, Rauramaa R, Rivadeneira F, Saaristo TE, Saleheen D, Sattar N, Schadt EE, Schlessinger D, Eline Slagboom P, Snieder H, Spector TD, Thorsteinsdottir U, Stumvoll M, Tuomilehto J, Uitterlinden AG, Uusitupa M, van der Harst P, Walker M, Wallaschofski H, Wareham NJ, Watkins H, Weir DR, Wichmann H-E, Wilson JF, Zanen P, Borecki IB, Deloukas P, Fox CS, Heid IM, O'Connell JR, Strachan DP, Stefansson K, van Duijn CM, Abecasis GR, Franke L, Frayling TM, McCarthy MI, Visscher PM, Scherag A, Willer CJ, Boehnke M, Mohlke KL, Lindgren CM, Beckmann JS, Barroso I, North KE, Ingelsson E, Hirschhorn JN, Loos RJF, Speliotes EK. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015 Feb;518(7538):197–206.

10. You Q, Yang X, Peng Z, Xu L, Wang J. Development and Applications of a High Throughput Genotyping Tool for Polyploid Crops: Single Nucleotide Polymorphism (SNP) Array. Front Plant Sci. 2018 Feb 6;9:104.

11. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science. 1995 Jul 28;269(5223):496–512.

12. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature. 2001 Feb 15;409(6822):860–921.

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

14. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015 Oct 1;526(7571):68–74.

15. Genome Aggregation Database Consortium, Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, Gauthier LD, Brand H, Solomonson M, Watts NA, Rhodes D, Singer-Berk M, England EM, Seaby EG, Kosmicki JA, Walters RK, Tashman K, Farjoun Y, Banks E, Poterba T, Wang A, Seed C, Whiffin N, Chong JX, Samocha KE, Pierce-Hoffman E, Zappala Z, O'Donnell-Luria AH, Minikel EV, Weisburd B, Lek M, Ware JS, Vittal C, Armean IM, Bergelson L, Cibulskis K, Connolly KM, Covarrubias M, Donnelly S, Ferriera S, Gabriel S, Gentry J, Gupta N, Jeandet T, Kaplan D, Llanwarne C, Munshi R, Novod S, Petrillo N, Roazen D, Ruano-Rubio V, Saltzman A, Schleicher M, Soto J, Tibbetts K, Tolonen C, Wade G, Talkowski ME, Neale BM, Daly MJ, MacArthur DG. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020 May;581(7809):434–43.

16. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, Vrieze SI, Chew EY, Levy S, McGue M, Schlessinger D, Stambolian D, Loh P-R, Iacono WG, Swaroop A, Scott LJ, Cucca F, Kronenberg F, Boehnke M, Abecasis GR, Fuchsberger C. Next-generation genotype imputation service and methods. Nat Genet. 2016 Oct;48(10):1284–7.

17. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TFC, McCarroll SA, Visscher PM. Finding the missing heritability of complex diseases. Nature. 2009 Oct 8;461(7265):747–53.

18. Yengo L, Sidorenko J, Kemper KE, Zheng Z, Wood AR, Weedon MN, Frayling TM, Hirschhorn J, Yang J, Visscher PM, the GIANT Consortium. Meta-analysis of genome-wide association studies for height and body mass index in ∼700000 individuals of European ancestry. Human Molecular Genetics. 2018 Oct 15;27(20):3641–9.

19. Wainschtein P, Jain DP, Yengo L, Zheng Z, TOPMed Anthropometry Working Group, Trans-Omics for Precision Medicine Consortium, Cupples LA, Shadyab AH, McKnight B, Shoemaker BM, Mitchell BD, Psaty BM, Kooperberg C, Roden D, Darbar D, Arnett DK, Regan EA, Boerwinkle E, Rotter JI, Allison MA, McDonald M-LN, Chung MK, Smith NL,

Ellinor PT, Vasan RS, Mathias RA, Rich SS, Heckbert SR, Redline S, Guo X, Chen Y-DI, Liu C-T, de Andrade M, Yanek LR, Albert CM, Hernandez RD, McGarvey ST, North KE, Lange LA, Weir BS, Laurie CC, Yang J, Visscher PM. Recovery of trait heritability from whole genome sequence data [Internet]. Genetics; 2019 Mar [cited 2021 May 13]. Available from: http://biorxiv.org/lookup/doi/10.1101/588020

20. Geddes L. Genetic study homes in on height's heritability mystery. Nature. 2019 Apr;568(7753):444–5.

21. Pers TH, Karjalainen JM, Chan Y, Westra H-J, Wood AR, Yang J, Lui JC, Vedantam S, Gustafsson S, Esko T, Frayling T, Speliotes EK, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, Boehnke M, Raychaudhuri S, Fehrmann RSN, Hirschhorn JN, Franke L. Biological interpretation of genome-wide association studies using predicted gene functions. Nat Commun. 2015 Jan 19;6:5890.

22. Cohen J, Pertsemlidis A, Kotowski IK, Graham R, Garcia CK, Hobbs HH. Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. Nat Genet. 2005 Feb;37(2):161–5.

23. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, Helgason A, Walters GB, Gunnarsdottir S, Mouy M, Steinthorsdottir V, Eiriksdottir GH, Bjornsdottir G, Reynisdottir I, Gudbjartsson D, Helgadottir A, Jonasdottir A, Jonasdottir A, Styrkarsdottir U, Gretarsdottir S, Magnusson KP, Stefansson H, Fossdal R, Kristjansson K, Gislason HG, Stefansson T, Leifsson BG, Thorsteinsdottir U, Lamb JR, Gulcher JR, Reitman ML, Kong A, Schadt EE, Stefansson K. Genetics of gene expression and its effect on disease. Nature. 2008 Mar;452(7186):423–8.

24. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. 2010 Apr 1;6(4):e1000888.

25. Nica AC, Montgomery SB, Dimas AS, Stranger BE, Beazley C, Barroso I, Dermitzakis ET. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. PLoS Genet. 2010 Apr 1;6(4):e1000895.

26. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. Nat Rev Genet. 2015 Apr;16(4):197–212.

27. Smith AK, Kilaru V, Kocak M, Almli LM, Mercer KB, Ressler KJ, Tylavsky FA, Conneely KN. Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type. BMC Genomics. 2014;15(1):145.

28. Suzuki M. An integrative analysis sheds light on methylation profiles. Sci Transl Med. 2016 Jan 13;8(321):321ec8-321ec8.

29. Liang D, Elwell AL, Aygün N, Lafferty MJ, Krupa O, Cheek KE, Courtney KP, Yusupova M, Garrett ME, Ashley-Koch A, Crawford GE, Love MI, de la Torre-Ubieta L, Geschwind DH, Stein JL. Cell-type specific effects of genetic variation on chromatin accessibility during human neuronal differentiation [Internet]. Genetics; 2020 Jan [cited 2021 Jan 2]. Available from: http://biorxiv.org/lookup/doi/10.1101/2020.01.13.904862

30. HIPSCI Consortium, Alasoo K, Rodrigues J, Mukhopadhyay S, Knights AJ, Mann AL, Kundu K, Hale C, Dougan G, Gaffney DJ. Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. Nat Genet. 2018 Mar;50(3):424–31.

31. Holdt LM, von Delft A, Nicolaou A, Baumann S, Kostrzewa M, Thiery J, Teupser D. Quantitative Trait Loci Mapping of the Mouse Plasma Proteome (pQTL). Genetics. 2013 Feb;193(2):601–8.

32. Veyrieras J-B, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, Pritchard JK. High-Resolution Mapping of Expression-QTLs Yields Insight into Human Gene Regulation. Gibson G, editor. PLoS Genet. 2008 Oct 10;4(10):e1000214.

33. Gusev A, Lee SH, Trynka G, Finucane H, Vilhjálmsson BJ, Xu H, Zang C, Ripke S, Bulik-Sullivan B, Stahl E, Kähler AK, Hultman CM, Purcell SM, McCarroll SA, Daly M, Pasaniuc B, Sullivan PF, Neale BM, Wray NR, Raychaudhuri S, Price AL, Ripke S, Neale BM, Corvin A, Walters JTR, Farh K-H, Holmans PA, Lee P, Bulik-Sullivan B, Collier DA, Huang H, Pers TH, Agartz I, Agerbo E, Albus M, Alexander M, Amin F, Bacanu SA, Begemann M, Belliveau RA, Bene J, Bergen SE, Bevilacqua E, Bigdeli TB, Black DW, Børglum AD, Bruggeman R, Buccola NG, Buckner RL, Byerley W, Cahn W, Cai G, Campion D, Cantor RM, Carr VJ, Carrera N, Catts SV, Chambert KD, Chan RCK, Chen RYL, Chen EYH, Cheng W, Cheung EFC, Chong SA, Cloninger CR, Cohen D, Cohen N, Cormican P, Craddock N, Crowley JJ, Curtis D, Davidson M, Davis KL, Degenhardt F, Del Favero J, DeLisi LE, Demontis D, Dikeos D, Dinan T, Djurovic S, Donohoe G, Drapeau E, Duan J, Dudbridge F, Durmishi N, Eichhammer P, Eriksson J, Escott-Price V, Essioux L, Fanous AH, Farrell MS, Frank J, Franke L, Freedman R, Freimer NB, Friedl M, Friedman JI, Fromer M, Genovese G, Georgieva L, Gershon ES, Giegling I, Giusti-Rodrguez P, Godard S, Goldstein JI, Golimbet V, Gopal S, Gratten J, Grove J, de Haan L, Hammer C, Hamshere ML, Hansen M, Hansen T, Haroutunian V, Hartmann AM, Henskens FA, Herms S, Hirschhorn JN, Hoffmann P, Hofman A, Hollegaard MV, Hougaard DM, Ikeda M, Joa I, Julià A, Kahn RS, Kalaydjieva L, Karachanak-Yankova S, Karjalainen J, Kavanagh D, Keller MC, Kelly BJ, Kennedy JL, Khrunin A, Kim Y, Klovins J, Knowles JA, Konte B, Kucinskas V, Kucinskiene ZA, Kuzelova-Ptackova H, Kähler AK, Laurent C, Keong JLC, Lee SH, Legge SE, Lerer B, Li M, Li T, Liang K-Y, Lieberman J, Limborska S, Loughland CM, Lubinski J, Lnnqvist J, Macek M, Magnusson PKE, Maher BS, Maier W, Mallet J, Marsal S, Mattheisen M, Mattingsdal M, McCarley RW, McDonald C, McIntosh AM, Meier S, Meijer CJ, Melegh B, Melle I, Mesholam-Gately RI, Metspalu A, Michie PT, Milani L, Milanova V, Mokrab Y, Morris DW, Mors O, Mortensen PB, Murphy KC, Murray RM, Myin-Germeys I, Mller-Myhsok B, Nelis M, Nenadic I, Nertney DA,

Nestadt G, Nicodemus KK, Nikitina-Zake L, Nisenbaum L, Nordin A, O'Callaghan E, O'Dushlaine C, O'Neill FA, Oh S-Y, Olincy A, Olsen L, Van Os J, Pantelis C, Papadimitriou GN, Papiol S, Parkhomenko E, Pato MT, Paunio T, Pejovic-Milovancevic M, Perkins DO, Pietilinen O, Pimm J, Pocklington AJ, Powell J, Price A, Pulver AE, Purcell SM, Quested D, Rasmussen HB, Reichenberg A, Reimers MA, Richards AL, Roffman JL, Roussos P, Ruderfer DM, Salomaa V, Sanders AR, Schall U, Schubert CR, Schulze TG, Schwab SG, Scolnick EM, Scott RJ, Seidman LJ, Shi J, Sigurdsson E, Silagadze T, Silverman JM, Sim K, Slominsky P, Smoller JW, So H-C, Spencer CCA, Stahl EA, Stefansson H, Steinberg S, Stogmann E, Straub RE, Strengman E, Strohmaier J, Stroup TS, Subramaniam M, Suvisaari J, Svrakic DM, Szatkiewicz JP, Sderman E, Thirumalai S, Toncheva D, Tooney PA, Tosato S, Veijola J, Waddington J, Walsh D, Wang D, Wang Q, Webb BT, Weiser M, Wildenauer DB, Williams NM, Williams S, Witt SH, Wolen AR, Wong EHM, Wormley BK, Wu JQ, Xi HS, Zai CC, Zheng X, Zimprich F, Wray NR, Stefansson K, Visscher PM, Adolfsson R, Andreassen OA, Blackwood DHR, Bramon E, Buxbaum JD, Brglum AD, Cichon S, Darvasi A, Domenici E, Ehrenreich H, Esko T, Gejman PV, Gill M, Gurling H, Hultman CM, Iwata N, Jablensky AV, Jönsson EG, Kendler KS, Kirov G, Knight J, Lencz T, Levinson DF, Li QS, Liu J, Malhotra AK, McCarroll SA, McQuillin A, Moran JL, Mortensen PB, Mowry BJ, Nthen MM, Ophoff RA, Owen MJ, Palotie A, Pato CN, Petryshen TL, Posthuma D, Rietschel M, Riley BP, Rujescu D, Sham PC, Sklar P, St. Clair D, Weinberger DR, Wendland JR, Werge T, Daly MJ, Sullivan PF, O'Donovan MC, Ripke S, O'Dushlaine C, Chambert K, Moran JL, Kähler AK, Akterin S, Bergen S, Magnusson PKE, Neale BM, Ruderfer D, Scolnick E, Purcell S, McCarroll S, Sklar P, Hultman CM, Sullivan PF. Partitioning Heritability of Regulatory and Cell-Type-Specific Variants across 11 Common Diseases. The American Journal of Human Genetics. 2014 Nov;95(5):535–52.

34. Jansen RC, Nap JP. Genetical genomics: the added value from segregation. Trends Genet. 2001 Jul;17(7):388–91.

35. Li J, Burmeister M. Genetical genomics: combining genetics with gene expression analysis. Human Molecular Genetics. 2005 Oct 15;14(suppl_2):R163–9.

36. Jacob F, Monod J. Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol. 1961 Jun;3:318–56.

37. Gann A. Jacob and Monod: from operons to EvoDevo. Curr Biol. 2010 Sep 14;20(17):R718-723.

38. Volgin DV. Gene Expression. In: Animal Biotechnology [Internet]. Elsevier; 2014 [cited 2021 Jan 3]. p. 307–25. Available from: https://linkinghub.elsevier.com/retrieve/pii/B9780124160026000171

39. Nica AC, Dermitzakis ET. Expression quantitative trait loci: present and future. Phil Trans R Soc B. 2013 Jun 19;368(1620):20120362.

40. Gilad Y, Rifkin SA, Pritchard JK. Revealing the architecture of gene regulation: the promise of eQTL studies. Trends in Genetics. 2008 Aug;24(8):408–15.

41. Clyde D. Transitioning from association to causation with eQTLs. Nat Rev Genet. 2017 May;18(5):271–271.

42. Göring HHH, Curran JE, Johnson MP, Dyer TD, Charlesworth J, Cole SA, Jowett JBM, Abraham LJ, Rainwater DL, Comuzzie AG, Mahaney MC, Almasy L, MacCluer JW, Kissebah AH, Collier GR, Moses EK, Blangero J. Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. Nat Genet. 2007 Oct;39(10):1208–16.

43. Shan N, Wang Z, Hou L. Identification of trans-eQTLs using mediation analysis with multiple mediators. BMC Bioinformatics. 2019 Mar;20(S3):126.

44. Yao C, Joehanes R, Johnson AD, Huan T, Liu C, Freedman JE, Munson PJ, Hill DE, Vidal M, Levy D. Dynamic Role of trans Regulation of Gene Expression in Relation to Complex Traits. The American Journal of Human Genetics. 2017 Apr;100(4):571–80.

45. Brem RB, Yvert G, Clinton R, Kruglyak L. Genetic dissection of transcriptional regulation in budding yeast. Science. 2002 Apr 26;296(5568):752–5.

46. van Nas A, Ingram-Drake L, Sinsheimer JS, Wang SS, Schadt EE, Drake T, Lusis AJ. Expression quantitative trait loci: replication, tissue- and sex-specificity in mice. Genetics. 2010 Jul;185(3):1059–68.

47. The Geuvadis Consortium, Lappalainen T, Sammeth M, Friedländer MR, ‘t Hoen PAC, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG, Barann M, Wieland T, Greger L, van Iterson M, Almlöf J, Ribeca P, Pulyakhina I, Esser D, Giger T, Tikhonov A, Sultan M, Bertier G, MacArthur DG, Lek M, Lizano E, Buermans HPJ, Padioleau I, Schwarzmayr T, Karlberg O, Ongen H, Kilpinen H, Beltran S, Gut M, Kahlem K, Amstislavskiy V, Stegle O, Pirinen M, Montgomery SB, Donnelly P, McCarthy MI, Flicek P, Strom TM, Lehrach H, Schreiber S, Sudbrak R, Carracedo Á, Antonarakis SE, Häsler R, Syvänen A-C, van Ommen G-J, Brazma A, Meitinger T, Rosenstiel P, Guigó R, Gut IG, Estivill X, Dermitzakis ET. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013 Sep;501(7468):506–11.

48. Moffatt MF, Kabesch M, Liang L, Dixon AL, Strachan D, Heath S, Depner M, von Berg A, Bufe A, Rietschel E, Heinzmann A, Simma B, Frischer T, Willis-Owen SAG, Wong KCC, Illig T, Vogelberg C, Weiland SK, von Mutius E, Abecasis GR, Farrall M, Gut IG, Lathrop GM, Cookson WOC. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature. 2007 Jul 26;448(7152):470–3.

49. Libioulle C, Louis E, Hansoul S, Sandor C, Farnir F, Franchimont D, Vermeire S, Dewit O, de Vos M, Dixon A, Demarche B, Gut I, Heath S, Foglio M, Liang L, Laukens D, Mni M, Zelenika D, Van Gossum A, Rutgeerts P, Belaiche J, Lathrop M, Georges M. Novel Crohn

disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. PLoS Genet. 2007 Apr 20;3(4):e58.

50. Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, Haudenschild CD, Beckman KB, Shi J, Mei R, Urban AE, Montgomery SB, Levinson DF, Koller D. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res. 2014 Jan;24(1):14–24.

51. The GTEx Consortium, Ardlie KG, Deluca DS, Segre AV, Sullivan TJ, Young TR, Gelfand ET, Trowbridge CA, Maller JB, Tukiainen T, Lek M, Ward LD, Kheradpour P, Iriarte B, Meng Y, Palmer CD, Esko T, Winckler W, Hirschhorn JN, Kellis M, MacArthur DG, Getz G, Shabalin AA, Li G, Zhou Y-H, Nobel AB, Rusyn I, Wright FA, Lappalainen T, Ferreira PG, Ongen H, Rivas MA, Battle A, Mostafavi S, Monlong J, Sammeth M, Mele M, Reverter F, Goldmann JM, Koller D, Guigo R, McCarthy MI, Dermitzakis ET, Gamazon ER, Im HK, Konkashbaev A, Nicolae DL, Cox NJ, Flutre T, Wen X, Stephens M, Pritchard JK, Tu Z, Zhang B, Huang T, Long Q, Lin L, Yang J, Zhu J, Liu J, Brown A, Mestichelli B, Tidwell D, Lo E, Salvatore M, Shad S, Thomas JA, Lonsdale JT, Moser MT, Gillard BM, Karasik E, Ramsey K, Choi C, Foster BA, Syron J, Fleming J, Magazine H, Hasz R, Walters GD, Bridge JP, Miklos M, Sullivan S, Barker LK, Traino HM, Mosavel M, Siminoff LA, Valley DR, Rohrer DC, Jewell SD, Branton PA, Sobin LH, Barcus M, Qi L, McLean J, Hariharan P, Um KS, Wu S, Tabor D, Shive C, Smith AM, Buia SA, Undale AH, Robinson KL, Roche N, Valentino KM, Britton A, Burges R, Bradbury D, Hambright KW, Seleski J, Korzeniewski GE, Erickson K, Marcus Y, Tejada J, Taherian M, Lu C, Basile M, Mash DC, Volpi S, Struewing JP, Temple GF, Boyer J, Colantuoni D, Little R, Koester S, Carithers LJ, Moore HM, Guan P, Compton C, Sawyer SJ, Demchok JP, Vaught JB, Rabiner CA, Lockhart NC, Ardlie KG, Getz G, Wright FA, Kellis M, Volpi S, Dermitzakis ET. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science. 2015 May 8;348(6235):648–60.

52. The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020 Sep 11;369(6509):1318–30.

53. Wainberg M, Sinnott-Armstrong N, Mancuso N, Barbeira AN, Knowles DA, Golan D, Ermel R, Ruusalepp A, Quertermous T, Hao K, Björkegren JLM, Im HK, Pasaniuc B, Rivas MA, Kundaje A. Opportunities and challenges for transcriptome-wide association studies. Nat Genet. 2019 Apr;51(4):592–9.

54. Mancuso N, Shi H, Goddard P, Kichaev G, Gusev A, Pasaniuc B. Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. The American Journal of Human Genetics. 2017 Mar;100(3):473–87.

55. GTEx Consortium, Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, Eyler AE, Denny JC, Nicolae DL, Cox NJ, Im HK. A gene-based association method for mapping traits using reference transcriptome data. Nat Genet. 2015 Sep;47(9):1091–8.

56. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BWJH, Jansen R, de Geus EJC, Boomsma DI, Wright FA, Sullivan PF, Nikkola E, Alvarez M, Civelek M, Lusis AJ, Lehtimäki T, Raitoharju E, Kähönen M, Seppälä I, Raitakari OT, Kuusisto J, Laakso M, Price AL, Pajukanta P, Pasaniuc B. Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet. 2016 Mar;48(3):245–52.

57. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007 Jun;447(7145):661–78.

58. Vervier K, Michaelson JJ. SLINGER: large-scale learning for predicting gene expression. Sci Rep. 2016 Dec;6(1):39360.

59. Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-directed, spatially addressable parallel chemical synthesis. Science. 1991 Feb 15;251(4995):767–73.

60. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995 Oct 20;270(5235):467–70.

61. Luo J, Schumacher M, Scherer A, Sanoudou D, Megherbi D, Davison T, Shi T, Tong W, Shi L, Hong H, Zhao C, Elloumi F, Shi W, Thomas R, Lin S, Tillinghast G, Liu G, Zhou Y, Herman D, Li Y, Deng Y, Fang H, Bushel P, Woods M, Zhang J. A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. Pharmacogenomics J. 2010 Aug;10(4):278–91.

62. Blair S, Williams L, Bishop J, Chagovetz A. Microarray temperature optimization using hybridization kinetics. Methods Mol Biol. 2009;529:171–96.

63. Croner RS, Lausen B, Schellerer V, Zeittraeger I, Wein A, Schildberg C, Papadopoulos T, Dimmler A, Hahn EG, Hohenberger W, Brueckl WM. Comparability of microarray data between amplified and non amplified RNA in colorectal carcinoma. J Biomed Biotechnol. 2009;2009:837170.

64. Beekman JM, Boess F, Hildebrand H, Kalkuhl A, Suter L. Gene Expression Analysis of the Hepatotoxicant Methapyrilene in Primary Rat Hepatocytes: An Interlaboratory Study. Environ Health Perspect. 2006 Jan;114(1):92–9.

65. Dobbin KK, Beer DG, Meyerson M, Yeatman TJ, Gerald WL, Jacobson JW, Conley B, Buetow KH, Heiskanen M, Simon RM, Minna JD, Girard L, Misek DE, Taylor JMG, Hanash S, Naoki K, Hayes DN, Ladd-Acosta C, Enkemann SA, Viale A, Giordano TJ. Interlaboratory comparability study of cancer gene expression analysis using oligonucleotide microarrays. Clin Cancer Res. 2005 Jan 15;11(2 Pt 1):565–72.

66. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003 Jan 22;19(2):185–93.

67. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003 Apr;4(2):249–64.

68. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007 Jan 1;8(1):118–27.

69. Okoniewski MJ, Miller CJ. Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. BMC Bioinformatics. 2006 Dec;7(1):276.

70. Casneuf T, Van de Peer Y, Huber W. In situ analysis of cross-hybridisation on microarrays and the inference of expression correlation. BMC Bioinformatics. 2007;8(1):461.

71. Shendure J. The beginning of the end for microarrays? Nat Methods. 2008 Jul;5(7):585–7.

72. Zhao S, Fung-Leung W-P, Bittner A, Ngo K, Liu X. Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. Zhang S-D, editor. PLoS ONE. 2014 Jan 16;9(1):e78644.

73. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF. Complementary DNA sequencing: expressed sequence tags and human genome project. Science. 1991 Jun 21;252(5013):1651–6.

74. Adams MD, Kerlavage AR, Fleischmann RD, Fuldner RA, Bult CJ, Lee NH, Kirkness EF, Weinstock KG, Gocayne JD, White O. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. Nature. 1995 Sep 28;377(6547 Suppl):3–174.

75. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. Science. 1995 Oct 20;270(5235):484–7.

76. Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T, Fukuda S, Sasaki D, Podhajska A, Harbers M, Kawai J, Carninci P, Hayashizaki Y. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci U S A. 2003 Dec 23;100(26):15776–81.

77. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009 Jan;10(1):57–63.

78. Kukurba KR, Montgomery SB. RNA Sequencing and Analysis. Cold Spring Harb Protoc. 2015 Apr 13;2015(11):951–69.

79. Morin RD, Bainbridge M, Fejes A, Hirst M, Krzywinski M, Pugh TJ, McDonald H, Varhol R, Jones SJM, Marra MA. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. BioTechniques. 2008 Jul;45(1):81–94.

80. Cloonan N, Forrest ARR, Kolle G, Gardiner BBA, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, Robertson AJ, Perkins AC, Bruce SJ, Lee CC, Ranade SS, Peckham HE, Manning JM, McKernan KJ, Grimmond SM. Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Methods. 2008 Jul;5(7):613–9.

81. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008 Jun 6;320(5881):1344–9.

82. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008 Jul;5(7):621–8.

83. Li L, Clevers H. Coexistence of quiescent and active adult stem cells in mammals. Science. 2010 Jan 29;327(5965):542–5.

84. Huang S. Non-genetic heterogeneity of cells in development: more than just noise. Development. 2009 Dec;136(23):3853–62.

85. Shalek AK, Satija R, Shuga J, Trombetta JJ, Gennert D, Lu D, Chen P, Gertner RS, Gaublomme JT, Yosef N, Schwartz S, Fowler B, Weaver S, Wang J, Wang X, Ding R, Raychowdhury R, Friedman N, Hacohen N, Park H, May AP, Regev A. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. Nature. 2014 Jun;510(7505):363–9.

86. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, Lao K, Surani MA. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 2009 May;6(5):377–82.

87. Shaffer SM, Dunagin MC, Torborg SR, Torre EA, Emert B, Krepler C, Beqiri M, Sproesser K, Brafford PA, Xiao M, Eggan E, Anastopoulos IN, Vargas-Garcia CA, Singh A, Nathanson KL, Herlyn M, Raj A. Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. Nature. 2017 Jun 15;546(7658):431–5.

88. Wang Y, Navin NE. Advances and Applications of Single-Cell Sequencing Technologies. Molecular Cell. 2015 May;58(4):598–609.

89. Chen G, Ning B, Shi T. Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. Front Genet. 2019 Apr 5;10:317.

90. Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium, Shen X, Howard DM, Adams MJ, Hill WD, Clarke T-K, Deary IJ, Whalley HC, McIntosh AM. A

phenome-wide association and Mendelian Randomisation study of polygenic risk for depression in UK Biobank. Nat Commun. 2020 Dec;11(1):2301.

91. Zhou W, Zhao Z, Nielsen JB, Fritsche LG, LeFaive J, Gagliano Taliun SA, Bi W, Gabrielsen ME, Daly MJ, Neale BM, Hveem K, Abecasis GR, Willer CJ, Lee S. Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. Nat Genet. 2020 Jun;52(6):634–9.

92. Westra H-J, Franke L. From genome to function by studying eQTLs. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease. 2014 Oct;1842(10):1896–902.

93. Alzheimer's Disease Genetics Consortium, Hu Y, Li M, Lu Q, Weng H, Wang J, Zekavat SM, Yu Z, Li B, Gu J, Muchnik S, Shi Y, Kunkle BW, Mukherjee S, Natarajan P, Naj A, Kuzma A, Zhao Y, Crane PK, Lu H, Zhao H. A statistical framework for cross-tissue transcriptome-wide association analysis. Nat Genet. 2019 Mar;51(3):568–76.

94. Barbeira AN, Pividori M, Zheng J, Wheeler HE, Nicolae DL, Im HK. Integrating predicted transcriptome from multiple tissues improves association detection. Plagnol V, editor. PLoS Genet. 2019 Jan 22;15(1):e1007889.

95. Schleinitz D, Krause K, Wohland T, Gebhardt C, Linder N, Stumvoll M, Blüher M, Bechmann I, Kovacs P, Gericke M, Tönjes A. Identification of distinct transcriptome signatures of human adipose tissue from fifteen depots. Eur J Hum Genet. 2020 Dec;28(12):1714–25.

96. Shen K, Zeppillo T, Limon A. Regional transcriptome analysis of AMPA and GABAA receptor subunit expression generates E/I signatures of the human brain. Sci Rep. 2020 Dec;10(1):11352.

97. Mancuso CA, Canfield JL, Singla D, Krishnan A. A flexible, interpretable, and accurate approach for imputing the expression of unmeasured genes. Nucleic Acids Research. 2020 Dec 2;48(21):e125–e125.

98. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002 Jan 1;30(1):207–10.

99. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 2013 Jan;41(Database issue):D991-995.

100. Alberts R, Terpstra P, Li Y, Breitling R, Nap J-P, Jansen RC. Sequence Polymorphisms Cause Many False cis eQTLs. Storey J, editor. PLoS ONE. 2007 Jul 18;2(7):e622.

101. Ciobanu DC, Lu L, Mozhui K, Wang X, Jagalur M, Morris JA, Taylor WL, Dietz K, Simon P, Williams RW. Detection, Validation, and Downstream Analysis of Allelic Variation in Gene Expression. Genetics. 2010 Jan;184(1):119–28.

102. Ramasamy A, Trabzuni D, Gibbs JR, Dillman A, Hernandez DG, Arepalli S, Walker R, Smith C, Ilori GP, Shabalin AA, Li Y, Singleton AB, Cookson MR, NABEC, Hardy J, UKBEC, Ryten M, Weale ME. Resolving the polymorphism-in-probe problem is critical for correct interpretation of expression QTL studies. Nucleic Acids Res. 2013 Apr;41(7):e88.

103. Quigley D. Equalizer reduces SNP bias in Affymetrix microarrays. BMC Bioinformatics. 2015 Jul 30;16:238.

104. Dannemann M, Lachmann M, Lorenc A. 'maskBAD'--a package to detect and remove Affymetrix probes with binding affinity differences. BMC Bioinformatics. 2012 Apr 16;13:56.

105. Võsa U, Claringbould A, Westra H-J, Bonder MJ, Deelen P, Zeng B, Kirsten H, Saha A, Kreuzhuber R, Kasela S, Pervjakova N, Alvaes I, Fave M-J, Agbessi M, Christiansen M, Jansen R, Seppälä I, Tong L, Teumer A, Schramm K, Hemani G, Verlouw J, Yaghootkar H, Sönmez R, Brown A, Kukushkina V, Kalnapenkis A, Rüeger S, Porcu E, Kronberg-Guzman J, Kettunen J, Powell J, Lee B, Zhang F, Arindrarto W, Beutner F, BIOS Consortium, Brugge H, i2QTL Consortium, Dmitreva J, Elansary M, Fairfax BP, Georges M, Heijmans BT, Kähönen M, Kim Y, Knight JC, Kovacs P, Krohn K, Li S, Loeffler M, Marigorta UM, Mei H, Momozawa Y, Müller-Nurasyid M, Nauck M, Nivard M, Penninx B, Pritchard J, Raitakari O, Rotzchke O, Slagboom EP, Stehouwer CDA, Stumvoll M, Sullivan P, Hoen PAC 't, Thiery J, Tönjes A, van Dongen J, van Iterson M, Veldink J, Völker U, Wijmenga C, Swertz M, Andiappan A, Montgomery GW, Ripatti S, Perola M, Kutalik Z, Dermitzakis E, Bergmann S, Frayling T, van Meurs J, Prokisch H, Ahsan H, Pierce B, Lehtimäki T, Boomsma D, Psaty BM, Gharib SA, Awadalla P, Milani L, Ouwehand W, Downes K, Stegle O, Battle A, Yang J, Visscher PM, Scholz M, Gibson G, Esko T, Franke L. Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis [Internet]. Genomics; 2018 Oct [cited 2021 Jan 7]. Available from: http://biorxiv.org/lookup/doi/10.1101/447367

106. Gillies CE, Putler R, Menon R, Otto E, Yasutake K, Nair V, Hoover P, Lieb D, Li S, Eddy S, Fermin D, McNulty MT, Hacohen N, Kiryluk K, Kretzler M, Wen X, Sampson MG, Sedor J, Dell K, Schachere M, Lemley K, Whitted L, Srivastava T, Haney C, Sethna C, Grammatikopoulos K, Appel G, Toledo M, Greenbaum L, Wang C, Lee B, Adler S, Nast C, LaPage J, Athavale A, Neu A, Boynton S, Fervenza F, Hogan M, Lieske JC, Chernitskiy V, Kaskel F, Kumar N, Flynn P, Kopp J, Castro-Rubio E, Blake J, Trachtman H, Zhdanova O, Modersitzki F, Vento S, Lafayette R, Mehta K, Gadegbeku C, Johnstone D, Cattran D, Hladunewich M, Reich H, Ling P, Romano M, Fornoni A, Barisoni L, Bidot C, Kretzler M, Gipson D, Williams A, Pitter R, Nachman P, Gibson K, Grubbs S, Froment A, Holzman L, Meyers K, Kallem K, Cerecino F, Sambandam K, Brown E, Johnson N, Jefferson A, Hingorani S, Tuttle K, Curtin L, Dismuke S, Cooper A, Freedman B, Lin JJ, Gray S, Kretzler M, Barisoni L, Gadegbeku C, Gillespie B, Gipson D, Holzman L, Mariani L, Sampson MG,

Song P, Troost J, Zee J, Herreshoff E, Kincaid C, Lienczewski C, Mainieri T, Williams A, Abbott K, Roy C, Urv T, Brooks J. An eQTL Landscape of Kidney Tissue in Human Nephrotic Syndrome. The American Journal of Human Genetics. 2018 Aug;103(2):232–44.

107. Kumar V, Abbas AK, Fausto N, Robbins SL, Cotran RS, editors. Robbins and Cotran pathologic basis of disease. 7th ed. Philadelphia: Elsevier Saunders; 2005. 1525 p.

108. Ko Y-A, Yi H, Qiu C, Huang S, Park J, Ledo N, Köttgen A, Li H, Rader DJ, Pack MA, Brown CD, Susztak K. Genetic-Variation-Driven Gene-Expression Changes Highlight Genes with Important Functions for Kidney Disease. The American Journal of Human Genetics. 2017 Jun;100(6):940–53.

109. Porrini E, Ruggenenti P, Luis-Lima S, Carrara F, Jiménez A, de Vries APJ, Torres A, Gaspari F, Remuzzi G. Estimated GFR: time for a critical appraisal. Nat Rev Nephrol. 2019 Mar;15(3):177–90.

110. Sandholm N, Salem RM, McKnight AJ, Brennan EP, Forsblom C, Isakova T, McKay GJ, Williams WW, Sadlier DM, Mäkinen V-P, Swan EJ, Palmer C, Boright AP, Ahlqvist E, Deshmukh HA, Keller BJ, Huang H, Ahola AJ, Fagerholm E, Gordin D, Harjutsalo V, He B, Heikkilä O, Hietala K, Kytö J, Lahermo P, Lehto M, Lithovius R, Österholm A-M, Parkkonen M, Pitkäniemi J, Rosengård-Bärlund M, Saraheimo M, Sarti C, Söderlund J, Soro-Paavonen A, Syreeni A, Thorn LM, Tikkanen H, Tolonen N, Tryggvason K, Tuomilehto J, Wadén J, Gill GV, Prior S, Guiducci C, Mirel DB, Taylor A, Hosseini SM, DCCT/EDIC Research Group, Parving H-H, Rossing P, Tarnow L, Ladenvall C, Alhenc-Gelas F, Lefebvre P, Rigalleau V, Roussel R, Tregouet D-A, Maestroni A, Maestroni S, Falhammar H, Gu T, Möllsten A, Cimponeriu D, Ioana M, Mota M, Mota E, Serafinceanu C, Stavarachi M, Hanson RL, Nelson RG, Kretzler M, Colhoun HM, Panduru NM, Gu HF, Brismar K, Zerbini G, Hadjadj S, Marre M, Groop L, Lajer M, Bull SB, Waggott D, Paterson AD, Savage DA, Bain SC, Martin F, Hirschhorn JN, Godson C, Florez JC, Groop P-H, Maxwell AP. New Susceptibility Loci Associated with Kidney Disease in Type 1 Diabetes. Böger CA, editor. PLoS Genet. 2012 Sep 20;8(9):e1002921.

111. Muller YL, Piaggi P, Hanson RL, Kobes S, Bhutta S, Abdussamad M, Leak-Johnson T, Kretzler M, Huang K, Weil EJ, Nelson RG, Knowler WC, Bogardus C, Baier LJ. A cis-eQTL in PFKFB2 is associated with diabetic nephropathy, adiposity and insulin secretion in American Indians. Human Molecular Genetics. 2015 May 15;24(10):2985–96.

112. Kerlin BA, Blatt NB, Fuh B, Zhao S, Lehman A, Blanchong C, Mahan JD, Smoyer WE. Epidemiology and Risk Factors for Thromboembolic Complications of Childhood Nephrotic Syndrome: A Midwest Pediatric Nephrology Consortium (MWPNC) Study. The Journal of Pediatrics. 2009 Jul;155(1):105-110.e1.

113. Mikhaylova AV, Thornton TA. Accuracy of Gene Expression Prediction From Genotype Data With PrediXcan Varies Across and Within Continental Populations. Front Genet. 2019 Apr 3;10:261.

114. Asimit JL, Hatzikotoulas K, McCarthy M, Morris AP, Zeggini E. Trans-ethnic study design approaches for fine-mapping. Eur J Hum Genet. 2016 Sep;24(9):1330–6.

115. Danjou F, Zoledziewska M, Sidore C, Steri M, Busonero F, Maschio A, Mulas A, Perseu L, Barella S, Porcu E, Pistis G, Pitzalis M, Pala M, Menzel S, Metrustry S, Spector TD, Leoni L, Angius A, Uda M, Moi P, Thein SL, Galanello R, Abecasis GR, Schlessinger D, Sanna S, Cucca F. Genome-wide association analyses based on whole-genome sequencing in Sardinia provide insights into regulation of hemoglobin levels. Nat Genet. 2015 Nov;47(11):1264–71.

116. Sidore C, Busonero F, Maschio A, Porcu E, Naitza S, Zoledziewska M, Mulas A, Pistis G, Steri M, Danjou F, Kwong A, Ortega del Vecchyo VD, Chiang CWK, Bragg-Gresham J, Pitzalis M, Nagaraja R, Tarrier B, Brennan C, Uzzau S, Fuchsberger C, Atzeni R, Reinier F, Berutti R, Huang J, Timpson NJ, Toniolo D, Gasparini P, Malerba G, Dedoussis G, Zeggini E, Soranzo N, Jones C, Lyons R, Angius A, Kang HM, Novembre J, Sanna S, Schlessinger D, Cucca F, Abecasis GR. Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. Nat Genet. 2015 Nov;47(11):1272–81.

117. FinnGen Project, Locke AE, Steinberg KM, Chiang CWK, Service SK, Havulinna AS, Stell L, Pirinen M, Abel HJ, Chiang CC, Fulton RS, Jackson AU, Kang CJ, Kanchi KL, Koboldt DC, Larson DE, Nelson J, Nicholas TJ, Pietilä A, Ramensky V, Ray D, Scott LJ, Stringham HM, Vangipurapu J, Welch R, Yajnik P, Yin X, Eriksson JG, Ala-Korpela M, Järvelin M-R, Männikkö M, Laivuori H, Dutcher SK, Stitziel NO, Wilson RK, Hall IM, Sabatti C, Palotie A, Salomaa V, Laakso M, Ripatti S, Boehnke M, Freimer NB. Exome sequencing of Finnish isolates enhances rare-variant association power. Nature. 2019 Aug;572(7769):323–8.

118. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. Nat Rev Genet. 2019 Aug;20(8):467–84.

119. Pingault J-B, O'Reilly PF, Schoeler T, Ploubidis GB, Rijsdijk F, Dudbridge F. Using genetic data to strengthen causal inference in observational research. Nat Rev Genet. 2018 Sep;19(9):566–80.

120. Zhang W, Voloudakis G, Rajagopal VM, Readhead B, Dudley JT, Schadt EE, Björkegren JLM, Kim Y, Fullard JF, Hoffman GE, Roussos P. Integrative transcriptome imputation reveals tissue-specific and shared biological mechanisms mediating susceptibility to complex traits. Nat Commun. 2019 Dec;10(1):3834.

121. Barfield R, Feng H, Gusev A, Wu L, Zheng W, Pasaniuc B, Kraft P. Transcriptome-wide association studies accounting for colocalization using Egger regression. Genet Epidemiol. 2018 Jul;42(5):418–33.

122. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. International Journal of Epidemiology. 2015 Apr 1;44(2):512–25.

123. GTEx Consortium, Gamazon ER, Segrè AV, van de Bunt M, Wen X, Xi HS, Hormozdiari F, Ongen H, Konkashbaev A, Derks EM, Aguet F, Quan J, Nicolae DL, Eskin E, Kellis M, Getz G, McCarthy MI, Dermitzakis ET, Cox NJ, Ardlie KG. Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. Nat Genet. 2018 Jul;50(7):956–67.

124. Uhlen M, Zhang C, Lee S, Sjöstedt E, Fagerberg L, Bidkhori G, Benfeitas R, Arif M, Liu Z, Edfors F, Sanli K, von Feilitzen K, Oksvold P, Lundberg E, Hober S, Nilsson P, Mattsson J, Schwenk JM, Brunnström H, Glimelius B, Sjöblom T, Edqvist P-H, Djureinovic D, Micke P, Lindskog C, Mardinoglu A, Ponten F. A pathology atlas of the human cancer transcriptome. Science. 2017 Aug 18;357(6352):eaan2507.

125. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, Cortes A, Welsh S, Young A, Effingham M, McVean G, Leslie S, Allen N, Donnelly P, Marchini J. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018 Oct;562(7726):203–9.

126. The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. BMJ. 2018 May 2;k1952.

127. Collins FS, Varmus H. A New Initiative on Precision Medicine. N Engl J Med. 2015 Feb 26;372(9):793–5.

128. Sieberts SK, Perumal TM, Carrasquillo MM, Allen M, Reddy JS, Hoffman GE, Dang KK, Calley J, Ebert PJ, Eddy J, Wang X, Greenwood AK, Mostafavi S, CommonMind Consortium (CMC), The AMP-AD Consortium, Omberg L, Peters MA, Logsdon BA, De Jager PL, Ertekin-Taner N, Mangravite LM. Large eQTL meta-analysis reveals differing patterns between cerebral cortical and cerebellar brain regions. Sci Data. 2020 Oct 12;7(1):340.

129. Zeng B, Lloyd-Jones LR, Montgomery GW, Metspalu A, Esko T, Franke L, Vosa U, Claringbould A, Brigham KL, Quyyumi AA, Idaghdour Y, Yang J, Visscher PM, Powell JE, Gibson G. Comprehensive Multiple eQTL Detection and Its Application to GWAS Interpretation. Genetics. 2019 Jul;212(3):905–18.

130. Tran HTN, Ang KS, Chevrier M, Zhang X, Lee NYS, Goh M, Chen J. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biol. 2020 Dec;21(1):12.

131. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010 Oct;11(10):733–9.

132. Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genomics and Bioinformatics. 2020 Sep 1;2(3):lqaa078.

133. GTEx Consortium. Genetic effects on gene expression across human tissues. Nature. 2017 Oct 12;550(7675):204–13.

134. The GTEx Consortium, Zhang Y, Quick C, Yu K, Barbeira A, Luca F, Pique-Regi R, Kyung Im H, Wen X. PTWAS: investigating tissue-relevant causal molecular mechanisms of complex traits using probabilistic TWAS analysis. Genome Biol. 2020 Dec;21(1):232.

135. Bhattacharya A, García-Closas M, Olshan AF, Perou CM, Troester MA, Love MI. A framework for transcriptome-wide association studies in breast cancer in diverse study populations. Genome Biol. 2020 Dec;21(1):42.

136. Zhou D, Jiang Y, Zhong X, Cox NJ, Liu C, Gamazon ER. A unified framework for joint-tissue transcriptome-wide association and Mendelian randomization analysis. Nat Genet. 2020 Nov;52(11):1239–46.

137. GTEx Consortium, Barbeira AN, Dickinson SP, Bonazzola R, Zheng J, Wheeler HE, Torres JM, Torstenson ES, Shah KP, Garcia T, Edwards TL, Stahl EA, Huckins LM, Nicolae DL, Cox NJ, Im HK. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat Commun. 2018 Dec;9(1):1825.

138. Fuior EV, Gafencu AV. Apolipoprotein C1: Its Pleiotropic Effects in Lipid Metabolism and Beyond. IJMS. 2019 Nov 26;20(23):5939.

139. Okoro PC, Schubert R, Guo X, Johnson WC, Rotter JI, Hoeschele I, Liu Y, Im HK, Luke A, Dugas LR, Wheeler HE. Transcriptome prediction performance across machine learning models and diverse ancestries. Human Genetics and Genomics Advances. 2021 Apr;2(2):100019.

140. Cai M, Xiao J, Zhang S, Wan X, Zhao H, Chen G, Yang C. A unified framework for cross-population trait prediction by leveraging the genetic correlation of polygenic traits. The American Journal of Human Genetics. 2021 Apr;108(4):632–55.

141. Pividori M, Rajagopal PS, Barbeira A, Liang Y, Melia O, Bastarache L, Park Y, Consortium Gte, Wen X, Im HK. PhenomeXcan: Mapping the genome to the phenome through the transcriptome. Sci Adv. 2020 Sep;6(37):eaba2083.

142. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010;33(1):1–22.

143. Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, Loveland JE, Ruef BJ, Hart E, Suner M-M, Landrum MJ, Aken B, Ayling S, Baertsch R, Fernandez-Banet J, Cherry JL, Curwen V, DiCuccio M, Kellis M, Lee J, Lin MF, Schuster M, Shkeda A, Amid C, Brown G, Dukhanina O, Frankish A, Hart J, Maidak BL, Mudge J,

Murphy MR, Murphy T, Rajan J, Rajput B, Riddick LD, Snow C, Steward C, Webb D, Weber JA, Wilming L, Wu W, Birney E, Haussler D, Hubbard T, Ostell J, Durbin R, Lipman D. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. Genome Research. 2009 Jul 1;19(7):1316–23.

144. Global Lipids Genetics Consortium. Discovery and refinement of loci associated with lipid levels. Nat Genet. 2013 Nov;45(11):1274–83.

145. the DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat Genet. 2012 Sep;44(9):981–90.

146. Kerimov N, Hayhurst JD, Peikova K, Manning JR, Walter P, Kolberg L, Samoviča M, Sakthivel MP, Kuzmin I, Trevanion SJ, Burdett T, Jupp S, Parkinson H, Papatheodorou I, Yates A, Zerbino DR, Alasoo K. eQTL Catalogue: a compendium of uniformly processed human gene expression and splicing QTLs [Internet]. Genomics; 2020 Jan [cited 2021 Mar 23]. Available from: http://biorxiv.org/lookup/doi/10.1101/2020.01.29.924266

147. Joehanes R, Zhang X, Huan T, Yao C, Ying S, Nguyen QT, Demirkale CY, Feolo ML, Sharopova NR, Sturcke A, Schäffer AA, Heard-Costa N, Chen H, Liu P, Wang R, Woodhouse KA, Tanriverdi K, Freedman JE, Raghavachari N, Dupuis J, Johnson AD, O'Donnell CJ, Levy D, Munson PJ. Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies. Genome Biol. 2017 Dec;18(1):16.

148. Dai M. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Research. 2005 Nov 27;33(20):e175–e175.

149. Sandberg R, Larsson O. Improved precision and accuracy for microarrays using updated probe set definitions. BMC Bioinformatics. 2007 Dec;8(1):48.

150. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14(4):R36.

151. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010 May;28(5):511–5.

152. Jun G, Wing MK, Abecasis GR, Kang HM. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. Genome Res. 2015 Jun;25(6):918–25.

153. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY,

Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011 May;43(5):491–8.

154. Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, Chen H-C, Agarwala R, McLaren WM, Ritchie GRS, Albracht D, Kremitzki M, Rock S, Kotkiewicz H, Kremitzki C, Wollam A, Trani L, Fulton L, Fulton R, Matthews L, Whitehead S, Chow W, Torrance J, Dunn M, Harden G, Threadgold G, Wood J, Collins J, Heath P, Griffiths G, Pelan S, Grafham D, Eichler EE, Weinstock G, Mardis ER, Wilson RK, Howe K, Flicek P, Hubbard T. Modernizing reference genome assemblies. PLoS Biol. 2011 Jul;9(7):e1001091.

155. Carvalho BS, Irizarry RA. A framework for oligonucleotide microarray preprocessing. Bioinformatics. 2010 Oct 1;26(19):2363–7.

156. Irizarry RA. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Research. 2003 Feb 15;31(4):15e–15.

157. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S, Freimer NB, Sabatti C, Eskin E. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010 Apr;42(4):348–54.

158. Canela-Xandri O, Rawlik K, Tenesa A. An atlas of genetic associations in UK Biobank. Nat Genet. 2018 Nov;50(11):1593–9.

159. NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, Taliun SAG, Corvelo A, Gogarten SM, Kang HM, Pitsillides AN, LeFaive J, Lee S, Tian X, Browning BL, Das S, Emde A-K, Clarke WE, Loesch DP, Shetty AC, Blackwell TW, Smith AV, Wong Q, Liu X, Conomos MP, Bobo DM, Aguet F, Albert C, Alonso A, Ardlie KG, Arking DE, Aslibekyan S, Auer PL, Barnard J, Barr RG, Barwick L, Becker LC, Beer RL, Benjamin EJ, Bielak LF, Blangero J, Boehnke M, Bowden DW, Brody JA, Burchard EG, Cade BE, Casella JF, Chalazan B, Chasman DI, Chen Y-DI, Cho MH, Choi SH, Chung MK, Clish CB, Correa A, Curran JE, Custer B, Darbar D, Daya M, de Andrade M, DeMeo DL, Dutcher SK, Ellinor PT, Emery LS, Eng C, Fatkin D, Fingerlin T, Forer L, Fornage M, Franceschini N, Fuchsberger C, Fullerton SM, Germer S, Gladwin MT, Gottlieb DJ, Guo X, Hall ME, He J, Heard-Costa NL, Heckbert SR, Irvin MR, Johnsen JM, Johnson AD, Kaplan R, Kardia SLR, Kelly T, Kelly S, Kenny EE, Kiel DP, Klemmer R, Konkle BA, Kooperberg C, Köttgen A, Lange LA, Lasky-Su J, Levy D, Lin X, Lin K-H, Liu C, Loos RJF, Garman L, Gerszten R, Lubitz SA, Lunetta KL, Mak ACY, Manichaikul A, Manning AK, Mathias RA, McManus DD, McGarvey ST, Meigs JB, Meyers DA, Mikulla JL, Minear MA, Mitchell BD, Mohanty S, Montasser ME, Montgomery C, Morrison AC, Murabito JM, Natale A, Natarajan P, Nelson SC, North KE, O'Connell JR, Palmer ND, Pankratz N, Peloso GM, Peyser PA, Pleiness J, Post WS, Psaty BM, Rao DC, Redline S, Reiner AP, Roden D, Rotter JI, Ruczinski I, Sarnowski C, Schoenherr S, Schwartz DA, Seo J-S, Seshadri S, Sheehan VA, Sheu WH, Shoemaker MB, Smith NL, Smith JA, Sotoodehnia N, Stilp AM, Tang W, Taylor KD, Telen M, Thornton TA, Tracy RP, Van Den Berg DJ, Vasan RS, Viaud-

Martinez KA, Vrieze S, Weeks DE, Weir BS, Weiss ST, Weng L-C, Willer CJ, Zhang Y, Zhao X, Arnett DK, Ashley-Koch AE, Barnes KC, Boerwinkle E, Gabriel S, Gibbs R, Rice KM, Rich SS, Silverman EK, Qasba P, Gan W, Papanicolaou GJ, Nickerson DA, Browning SR, Zody MC, Zöllner S, Wilson JG, Cupples LA, Laurie CC, Jaquish CE, Hernandez RD, O'Connor TD, Abecasis GR. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature. 2021 Feb 11;590(7845):290–9.

160. Sofer T, Kurniansyah N, Aguet F, Ardlie K, Durda P, Nickerson DA, Smith JD, Liu Y, Gharib SA, Redline S, Rich SS, Rotter JI, Taylor KD. Benchmarking Association Analyses of Continuous Exposures with RNA-seq in Observational Studies [Internet]. Bioinformatics; 2021 Feb [cited 2021 Apr 8]. Available from: http://biorxiv.org/lookup/doi/10.1101/2021.02.12.430989

161. Folkersen L, Gustafsson S, Wang Q, Hansen DH, Hedman ÅK, Schork A, Page K, Zhernakova DV, Wu Y, Peters J, Ericsson N, Bergen SE, Boutin T, Bretherick AD, Enroth S, Kalnapenkis A, Gådin JR, Suur B, Chen Y, Matic L, Gale JD, Lee J, Zhang W, Quazi A, Ala-Korpela M, Choi SH, Claringbould A, Danesh J, Davey-Smith G, de Masi F, Elmståhl S, Engström G, Fauman E, Fernandez C, Franke L, Franks P, Giedraitis V, Haley C, Hamsten A, Ingason A, Johansson Å, Joshi PK, Lind L, Lindgren CM, Lubitz S, Palmer T, Macdonald-Dunlop E, Magnusson M, Melander O, Michaelsson K, Morris AP, Mägi R, Nagle M, Nilsson PM, Nilsson J, Orho-Melander M, Polasek O, Prins B, Pålsson E, Qi T, Sjögren M, Sundström J, Surendran P, Võsa U, Werge T, Wernersson R, Westra H-J, Yang J, Zhernakova A, Ärnlöv J, Fu J, Smith G, Esko T, Hayward C, Gyllensten U, Landen M, Siegbahn A, Wilson JF, Wallentin L, Butterworth AS, Holmes MV, Ingelsson E, Mälarstig A. Genomic evaluation of circulating proteins for drug target characterisation and precision medicine [Internet]. Genetics; 2020 Apr [cited 2021 Apr 8]. Available from: http://biorxiv.org/lookup/doi/10.1101/2020.04.03.023804

162. Folkersen L, Gustafsson S, Wang Q, Hansen DH, Hedman ÅK, Schork A, Page K, Zhernakova DV, Wu Y, Peters J, Eriksson N, Bergen SE, Boutin TS, Bretherick AD, Enroth S, Kalnapenkis A, Gådin JR, Suur BE, Chen Y, Matic L, Gale JD, Lee J, Zhang W, Quazi A, Ala-Korpela M, Choi SH, Claringbould A, Danesh J, Davey Smith G, de Masi F, Elmståhl S, Engström G, Fauman E, Fernandez C, Franke L, Franks PW, Giedraitis V, Haley C, Hamsten A, Ingason A, Johansson Å, Joshi PK, Lind L, Lindgren CM, Lubitz S, Palmer T, Macdonald-Dunlop E, Magnusson M, Melander O, Michaelsson K, Morris AP, Mägi R, Nagle MW, Nilsson PM, Nilsson J, Orho-Melander M, Polasek O, Prins B, Pålsson E, Qi T, Sjögren M, Sundström J, Surendran P, Võsa U, Werge T, Wernersson R, Westra H-J, Yang J, Zhernakova A, Ärnlöv J, Fu J, Smith JG, Esko T, Hayward C, Gyllensten U, Landen M, Siegbahn A, Wilson JF, Wallentin L, Butterworth AS, Holmes MV, Ingelsson E, Mälarstig A. Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. Nat Metab. 2020 Oct;2(10):1135–48.

163. West KM, Blacksher E, Burke W. Genomics, Health Disparities, and Missed Opportunities for the Nation's Research Agenda. JAMA. 2017 May 9;317(18):1831.

164. Need AC, Goldstein DB. Next generation disparities in human genomics: concerns and remedies. Trends in Genetics. 2009 Nov;25(11):489–94.

165. Petrovski S, Goldstein DB. Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. Genome Biol. 2016 Dec;17(1):157.

166. Hindorff LA, Bonham VL, Brody LC, Ginoza MEC, Hutter CM, Manolio TA, Green ED. Prioritizing diversity in human genomics research. Nat Rev Genet. 2018 Mar;19(3):175–85.

167. Yamauchi T, Hara K, Maeda S, Yasuda K, Takahashi A, Horikoshi M, Nakamura M, Fujita H, Grarup N, Cauchi S, Ng DPK, Ma RCW, Tsunoda T, Kubo M, Watada H, Maegawa H, Okada-Iwabu M, Iwabu M, Shojima N, Shin HD, Andersen G, Witte DR, Jørgensen T, Lauritzen T, Sandbæk A, Hansen T, Ohshige T, Omori S, Saito I, Kaku K, Hirose H, So W-Y, Beury D, Chan JCN, Park KS, Tai ES, Ito C, Tanaka Y, Kashiwagi A, Kawamori R, Kasuga M, Froguel P, Pedersen O, Kamatani N, Nakamura Y, Kadowaki T. A genome-wide association study in the Japanese population identifies susceptibility loci for type 2 diabetes at UBE2E2 and C2CD4A-C2CD4B. Nat Genet. 2010 Oct;42(10):864–8.

168. Strawbridge RJ, Dupuis J, Prokopenko I, Barker A, Ahlqvist E, Rybin D, Petrie JR, Travers ME, Bouatia-Naji N, Dimas AS, Nica A, Wheeler E, Chen H, Voight BF, Taneera J, Kanoni S, Peden JF, Turrini F, Gustafsson S, Zabena C, Almgren P, Barker DJP, Barnes D, Dennison EM, Eriksson JG, Eriksson P, Eury E, Folkersen L, Fox CS, Frayling TM, Goel A, Gu HF, Horikoshi M, Isomaa B, Jackson AU, Jameson KA, Kajantie E, Kerr-Conte J, Kuulasmaa T, Kuusisto J, Loos RJF, Luan J, Makrilakis K, Manning AK, Martinez-Larrad MT, Narisu N, Nastase Mannila M, Ohrvik J, Osmond C, Pascoe L, Payne F, Sayer AA, Sennblad B, Silveira A, Stancakova A, Stirrups K, Swift AJ, Syvanen A-C, Tuomi T, van ’t Hooft FM, Walker M, Weedon MN, Xie W, Zethelius B, the DIAGRAM Consortium, the GIANT Consortium, the MuTHER Consortium, the CARDIoGRAM Consortium, the C4D Consortium, Ongen H, Malarstig A, Hopewell JC, Saleheen D, Chambers J, Parish S, Danesh J, Kooner J, Ostenson C-G, Lind L, Cooper CC, Serrano-Rios M, Ferrannini E, Forsen TJ, Clarke R, Franzosi MG, Seedorf U, Watkins H, Froguel P, Johnson P, Deloukas P, Collins FS, Laakso M, Dermitzakis ET, Boehnke M, McCarthy MI, Wareham NJ, Groop L, Pattou F, Gloyn AL, Dedoussis GV, Lyssenko V, Meigs JB, Barroso I, Watanabe RM, Ingelsson E, Langenberg C, Hamsten A, Florez JC. Genome-Wide Association Identifies Nine Common Variants Associated With Fasting Proinsulin Levels and Provides New Insights Into the Pathophysiology of Type 2 Diabetes. Diabetes. 2011 Oct 1;60(10):2624–34.

169. Nair V, Komorowsky CV, Weil EJ, Yee B, Hodgin J, Harder JL, Godfrey B, Ju W, Boustany-Kari CM, Schwarz M, Lemley KV, Nelson PJ, Nelson RG, Kretzler M. A molecular morphometric approach to diabetic kidney disease can link structure to function and outcome. Kidney International. 2018 Feb;93(2):439–49.

170. Kang HM, Ye C, Eskin E. Accurate Discovery of Expression Quantitative Trait Loci Under Confounding From Spurious and Genuine Regulatory Hotspots. Genetics. 2008 Dec;180(4):1909–25.

171. Quick C, Guan L, Li Z, Li X, Dey R, Liu Y, Scott L, Lin X. A versatile toolkit for molecular QTL mapping and meta-analysis at scale [Internet]. Genetics; 2020 Dec [cited 2021 Apr 14]. Available from: http://biorxiv.org/lookup/doi/10.1101/2020.12.18.423490

172. Ding J, Gudjonsson JE, Liang L, Stuart PE, Li Y, Chen W, Weichenthal M, Ellinghaus E, Franke A, Cookson W, Nair RP, Elder JT, Abecasis GR. Gene expression in skin and lymphoblastoid cells: Refined statistical method reveals extensive overlap in cis-eQTL signals. Am J Hum Genet. 2010 Dec 10;87(6):779–89.

173. Wen X, Pique-Regi R, Luca F. Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. Li B, editor. PLoS Genet. 2017 Mar 9;13(3):e1006646.

174. Wen X, Lee Y, Luca F, Pique-Regi R. Efficient Integrative Multi-SNP Association Analysis via Deterministic Approximation of Posteriors. The American Journal of Human Genetics. 2016 Jun;98(6):1114–29.

175. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. JSTOR. 1995;57(1):289–300.

176. Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research. 2000 Jan 1;28(1):27–30.

177. Kanehisa M. Toward understanding the origin and evolution of cellular organisms. Protein Science. 2019 Nov;28(11):1947–51.

178. Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: integrating viruses and cellular organisms. Nucleic Acids Research. 2021 Jan 8;49(D1):D545–51.

179. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, McDermott MG, Monteiro CD, Gundersen GW, Ma'ayan A. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016 Jul 8;44(W1):W90–7.

180. Donate-Correa J, Ferri CM, Sánchez-Quintana F, Pérez-Castro A, González-Luis A, Martín-Núñez E, Mora-Fernández C, Navarro-González JF. Inflammatory Cytokines in Diabetic Kidney Disease: Pathophysiologic and Therapeutic Implications. Front Med. 2021 Jan 22;7:628289.

181. Magno A, Herat L, Carnagarin R, Schlaich M, Matthews V. Current Knowledge of IL-6 Cytokine Family Members in Acute and Chronic Kidney Disease. Biomedicines. 2019 Mar 13;7(1):19.

182. Navarro-González JF, Mora-Fernández C. The Role of Inflammatory Cytokines in Diabetic Nephropathy. JASN. 2008 Mar;19(3):433–42.

183. Lennon R, Randles MJ, Humphries MJ. The Importance of Podocyte Adhesion for a Healthy Glomerulus. Front Endocrinol [Internet]. 2014 Oct 14 [cited 2021 Apr 14];5. Available from: http://journal.frontiersin.org/article/10.3389/fendo.2014.00160/abstract

184. Lausecker F, Tian X, Inoue K, Wang Z, Pedigo CE, Hassan H, Liu C, Zimmer M, Jinno S, Huckle AL, Hamidi H, Ross RS, Zent R, Ballestrem C, Lennon R, Ishibe S. Vinculin is required to maintain glomerular barrier integrity. Kidney International. 2018 Mar;93(3):643–55.

185. Majo S, Courtois S, Souleyreau W, Bikfalvi A, Auguste P. Impact of Extracellular Matrix Components to Renal Cell Carcinoma Behavior. Front Oncol. 2020 Apr 28;10:625.

186. Simon EE, McDonald JA. Extracellular matrix receptors in the kidney cortex. American Journal of Physiology-Renal Physiology. 1990 Nov 1;259(5):F783–92.

187. Caplin B, Wang Z, Slaviero A, Tomlinson J, Dowsett L, Delahaye M, Salama A, The International Consortium for Blood Pressure Genome-Wide Association Studies, Wheeler DC, Leiper J. Alanine-Glyoxylate Aminotransferase-2 Metabolizes Endogenous Methylarginines, Regulates NO, and Controls Blood Pressure. Arterioscler Thromb Vasc Biol. 2012 Dec;32(12):2892–900.

188. Parving H-H. Microalbuminuria in essential hypertension and diabetes mellitus: Journal of Hypertension. 1996 Sep;14(Supplement 2):S89–94.

189. Palatini P. +Microalbuminuria in hypertension. Current Science Inc. 2003 May;5(3):208–14.

190. Takase H, Sugiura T, Ohte N, Dohi Y. Urinary Albumin as a Marker of Future Blood Pressure and Hypertension in the General Population. Medicine. 2015 Feb;94(6):e511.

191. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 10 Years of GWAS Discovery: Biology, Function, and Translation. The American Journal of Human Genetics. 2017 Jul;101(1):5–22.

192. Comeron JM, Kreitman M, De La Vega FM. On the power to detect SNP/phenotype association in candidate quantitative trait loci genomic regions: a simulation study. Pac Symp Biocomput. 2003;478–89.

193. Nishino J, Ochi H, Kochi Y, Tsunoda T, Matsui S. Sample Size for Successful Genome-Wide Association Study of Major Depressive Disorder. Front Genet. 2018 Jun 28;9:227.

194. Mills MC, Rahal C. A scientometric review of genome-wide association studies. Commun Biol. 2019 Dec;2(1):9.

195. Köttgen A, Pattaro C, Böger CA, Fuchsberger C, Olden M, Glazer NL, Parsa A, Gao X, Yang Q, Smith AV, O'Connell JR, Li M, Schmidt H, Tanaka T, Isaacs A, Ketkar S, Hwang S-J, Johnson AD, Dehghan A, Teumer A, Paré G, Atkinson EJ, Zeller T, Lohman K, Cornelis MC, Probst-Hensch NM, Kronenberg F, Tönjes A, Hayward C, Aspelund T, Eiriksdottir G, Launer LJ, Harris TB, Rampersaud E, Mitchell BD, Arking DE, Boerwinkle E, Struchalin M, Cavalieri M, Singleton A, Giallauria F, Metter J, de Boer IH, Haritunians T, Lumley T, Siscovick D, Psaty BM, Zillikens MC, Oostra BA, Feitosa M, Province M, de Andrade M, Turner ST, Schillert A, Ziegler A, Wild PS, Schnabel RB, Wilde S, Munzel TF, Leak TS, Illig T, Klopp N, Meisinger C, Wichmann H-E, Koenig W, Zgaga L, Zemunik T, Kolcic I, Minelli C, Hu FB, Johansson Å, Igl W, Zaboli G, Wild SH, Wright AF, Campbell H, Ellinghaus D, Schreiber S, Aulchenko YS, Felix JF, Rivadeneira F, Uitterlinden AG, Hofman A, Imboden M, Nitsch D, Brandstätter A, Kollerits B, Kedenko L, Mägi R, Stumvoll M, Kovacs P, Boban M, Campbell S, Endlich K, Völzke H, Kroemer HK, Nauck M, Völker U, Polasek O, Vitart V, Badola S, Parker AN, Ridker PM, Kardia SLR, Blankenberg S, Liu Y, Curhan GC, Franke A, Rochat T, Paulweber B, Prokopenko I, Wang W, Gudnason V, Shuldiner AR, Coresh J, Schmidt R, Ferrucci L, Shlipak MG, van Duijn CM, Borecki I, Krämer BK, Rudan I, Gyllensten U, Wilson JF, Witteman JC, Pramstaller PP, Rettig R, Hastie N, Chasman DI, Kao WH, Heid IM, Fox CS. New loci associated with kidney function and chronic kidney disease. Nat Genet. 2010 May;42(5):376–84.

196. Zeggini E, Ioannidis JP. Meta-analysis in genome-wide association studies. Pharmacogenomics. 2009 Feb;10(2):191–201.

197. Qiu C, Huang S, Park J, Park Y, Ko Y-A, Seasock MJ, Bryer JS, Xu X-X, Song W-C, Palmer M, Hill J, Guarnieri P, Hawkins J, Boustany-Kari CM, Pullen SS, Brown CD, Susztak K. Renal compartment–specific genetic variation analyses identify new pathways in chronic kidney disease. Nat Med. 2018 Nov;24(11):1721–31.

198. Ju W, Greene CS, Eichinger F, Nair V, Hodgin JB, Bitzer M, Lee Y, Zhu Q, Kehata M, Li M, Jiang S, Rastaldi MP, Cohen CD, Troyanskaya OG, Kretzler M. Defining cell-type specificity at the transcriptional level in human disease. Genome Res. 2013 Nov;23(11):1862–73.

199. Trapnell C. Defining cell types and states with single-cell genomics. Genome Res. 2015 Oct;25(10):1491–8.

200. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, Kling DE, Gauthier LD, Levy-Moonshine A, Roazen D, Shakir K, Thibault J, Chandran S, Whelan C, Lek M, Gabriel S, Daly MJ, Neale B, MacArthur DG, Banks E. Scaling accurate genetic variant discovery to tens of thousands of samples [Internet]. Genomics; 2017 Nov [cited 2021 Apr 14]. Available from: http://biorxiv.org/lookup/doi/10.1101/201178

201. Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat Protoc. 2012 Mar;7(3):500–7.

202. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E. Efficient Control of Population Structure in Model Organism Association Mapping. Genetics. 2008 Mar;178(3):1709–23.

203. Alquicira-Hernandez J, Sathe A, Ji HP, Nguyen Q, Powell JE. scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. Genome Biol. 2019 Dec;20(1):264.

204. Kobak D, Berens P. The art of using t-SNE for single-cell transcriptomics. Nat Commun. 2019 Dec;10(1):5416.

205. Poulin J-F, Tasic B, Hjerling-Leffler J, Trimarchi JM, Awatramani R. Disentangling neural cell diversity using single-cell transcriptomics. Nat Neurosci. 2016 Sep;19(9):1131–41.

206. Xu J, Cai L, Liao B, Zhu W, Yang J. CMF-Impute: an accurate imputation tool for single-cell RNA-seq data. Mathelier A, editor. Bioinformatics. 2020 May 1;36(10):3139–47.

207. Li WV, Li JJ. An accurate and robust imputation method scImpute for single-cell RNA-seq data. Nat Commun. 2018 Dec;9(1):997.

208. Xu Y, Zhang Z, You L, Liu J, Fan Z, Zhou X. scIGANs: single-cell RNA-seq imputation using generative adversarial networks. Nucleic Acids Research. 2020 Sep 4;48(15):e85–e85.

209. Wang X, Park J, Susztak K, Zhang NR, Li M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nat Commun. 2019 Dec;10(1):380.