CHROMATIN ACCESSIBILITY CONTEXT IDENTIFIES REGULATORY MECHANISMS FOR CARDIOMETABOLIC TRAITS

Hannah Janine Perrin

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Genetics Department in the School of Medicine.

Chapel Hill
2022

Approved by:

Karen L. Mohlke

Samir Kelada

Christopher Mack

Folami Ideraabdulla

Jason Stein

Michael Love

ABSTRACT

Hannah Janine Perrin: Chromatin accessibility context identifies regulatory mechanisms for cardiometabolic traits
Under the direction of Karen L. Mohlke

Cardiovascular diseases (CVD) and associated cardiovascular and metabolic (cardiometabolic) traits pose a significant global health burden. Identifying molecular mechanisms for cardiometabolic traits would improve diagnosis and treatment of disease. Genome-wide association studies (GWAS) have identified thousands of loci associated with cardiometabolic traits. However, the mechanisms of most remain unclear, especially at the large number of noncoding loci. One mechanism of noncoding loci is to regulate gene expression in cell- and context-dependent manners. Regulatory elements can be identified through chromatin accessibility. Therefore, chromatin accessibility in disease-relevant cell types and contexts can be integrated with gene expression and GWAS data to identify regulatory elements that affect gene expression to contribute to cardiometabolic traits. I profiled chromatin accessibility in adipose and liver tissue and in adipocytes exposed to disease-relevant contexts of differentiation, excess free fatty acids, hypoxia, and inflammation. I identified context-dependent regulatory elements that change after exposure to disease-relevant contexts in adipocytes and between sexes in liver tissue. I integrated context-dependent regulatory elements with multiple genomic datasets such as eQTL, Hi-C, and context-dependent gene expression to link elements to candidate genes. Additionally, I integrated context-dependent regulatory elements to GWAS to link elements to traits. Functional testing of candidate regulatory elements identified context- and allele-dependent transcriptional activity. While they require future functional testing, the work in this dissertation identifies hundreds of candidate regulatory mechanisms for noncoding GWAS loci. Furthermore, these chromatin accessibility profiles provide a useful resource for future work on identifying regulatory mechanisms of GWAS loci that may improve diagnosis and treatment of disease.

ACKNOWLEDGEMENTS

I am exceedingly grateful for all the people who supported me throughout my Ph.D. First, I would like to thank my mentor, Karen Mohlke. Thank you for all of your guidance and support. Thank you for helping me grow as a scientist, improve my writing and presentation skills, and supporting professional development opportunities outside of lab. Your enthusiasm for science, even (or especially) when experiments don't go as planned, made your lab an exciting workplace.

I would also like to thank numerous faculty and administrators at UNC who supported me throughout my graduate research. Thank you to my committee: Samir Kelada, Chris Mack, Mike Love, Jason Stein, and Folami Ideraabdullah. Through many meetings (official and unofficial) and emails you have all provided frequent encouragement and assistance throughout my graduate research. Thank you to John Cornett and Cara Marlow, for helping me navigate graduate school paperwork, answering frequent questions, and for always providing encouragement.

I would particularly like to thank Kevin Currin, with whom I worked most closely during my graduate research. Your contributions to experimental design, data analysis, and interpretation were invaluable to every project we worked on together. Your frequent, thought-provoking questions improved every experiment and helped me grow as a scientist. Your patience in teaching computational skills helped me straddle the often-difficult divide between wet and dry lab science. I can't thank you enough.

I would like to thank all Mohlke lab members, past and present, who helped me in countless ways throughout my graduate research. I would especially like to thank Maren Cannon, who was my student mentor during my rotation in Karen's lab. You were an excellent mentor, patiently answering countless questions. And you've continued to provide fantastic guidance on my career goals, thank you for all the time you've taken to help me. Thank you, Rani and Gautam for performing functional experiments for Chapter 2. Thank you, Rani for all the help with ATAC-seq library preparation for Chapter 4 and helping in countless ways to make lab work easier. Thank you, Kenneth Ng for assistance in generating ATAC-seq libraries for Chapter 2. Thank you, James Davis for helping with functional experiment design. Thank you, Cassie, Alaine, Raulie, Sarah, Tori, Emma, Apoorva, Shelley, and Sophie for teaching me about your projects, which helped me learn background for my own project and put my project into a larger perspective.

I would also like to thank collaborators who answered so many questions and provided help throughout my graduate research. Thank you, Alexias Safi, Mike Erdos, Arushi Varshney, and Nandini Manickam for assistance in optimizing nuclei isolation and ATAC-seq protocols for our samples. Thank you, Kristina Garske for providing and answering questions about adipocyte Hi-C data. Thank you, Martin Wabitsch for providing SGBS cells and answering questions regarding data publication. Thank you, Lianxin Hu and Qing Zhang for use of their hypoxia chamber. Thank you, Markku Laakso for providing METSIM adipose samples. Thank you, Amy Etheridge for assistance in obtaining and tracking liver samples. I would also like to thank the UNC High-Throughout Sequencing Facility and Advanced Analytics cores for assistance in large sequencing projects throughout my graduate research.

I would like to thank my mentors at GeneDx. Amanda Hussey, Lori Crumbliss, Isabelle Olivos-Glander, Patty McAndrew, and many more. My experiences at GeneDx instilled my love for genetics that pushed me onto graduate school. I appreciate all of the learning opportunities I experienced at GeneDx in both leadership and science.

Thank you to all my friends who provided support and made graduate school fun. From Friendsgivings, game nights, Spartan races, rock climbing, and more you all helped me find balance. Thank you to the Mohlke lab book club for excellent discussions and introducing me to a wide variety of topics.

Thank you to my family who always supported me. Thank you, Zola for being a fantastic and adorable companion. Thank you, Aunt Suzanne for encouraging me to continue with graduate school. Thank you, Claire for all your support and confidence boosting. Thank you, Alan for your support and being weird with me. Thank you, mom, dad, and Ryan for your love and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AFR	African ancestry
AI	Allelic imbalance
ATAC-seq	Assay for transposase-accessible chromatin followed by sequencing
BMI	Body mass index
BP	Base pair
CVD	Cardiovascular disease
DNase-seq	DNase hypersensitivity
ENCODE	Encyclopedia of DNA elements
eQTL	Expression quantitative trait loci
EUR	European ancestry
FAIRE-seq	Formaldehyde-assisted isolation of regulatory elements
FC	Fold change
FDR	False discovery rate
GWAS	Genome-wide association study
HDL-C	High density lipoprotein cholesterol
KB	Kilobase
LD	Linkage disequilibrium
LDL-C	Low density lipoprotein cholesterol
LFC	Log$_2$ fold change
MACS2	Model-based Analysis of ChIP-seq version 2
METSIM	Metabolic syndrome in men

PCA	Principal component analysis
NAFLD	Non-alcoholic fatty liver disease
QTL	Quantitative trait loci
RNA-seq	Ribonucleic acid sequencing
SGBS	Simpson-Golabi-Behmel syndrome
T2D	Type 2 diabetes
TF	Transcription factor
TSS	Transcription start site
WHR	Waist-hip ratio

CHAPTER 1: INTRODUCTION

Overview of cardiometabolic traits

Cardiovascular diseases (CVD), metabolic diseases such as Type 2 Diabetes (T2D), and associated cardiovascular and metabolic (cardiometabolic) traits are a significant global health burden. CVD remains the leading cause of death worldwide, at a rate of 18.6 million deaths per year[1,2]. In the United States, CVD is estimated to affect 126.9 million adults and cost $363.4 billion per year in healthcare[2]. Metabolic diseases are also a leading cause of mortality[3] and increase the risk of CVD[1,2,4]. T2D is the 8th leading cause of death in the United States in 2020, at a rate of 24.8 deaths per 100,000[3]. Associated prevalent cardiometabolic traits are additional risk factors for both CVD and T2D[2] and include obesity and high cholesterol[5] which, respectively, affect 96.5 million and 28 million adults in the United States[2]. Central obesity, or increased distribution of fat around the abdomen, increases risk of CVD, T2D, and related cardiometabolic traits more than obesity alone and is measured by traits such as waist-hip ratio (WHR)[5,6]. Obesity can lead to dyslipidemia, or an imbalance of lipids such as cholesterol and triglycerides[2,5]. High low-density lipoprotein cholesterol (LDL-C) increases risk of CVD through accumulation of plaques in the bloodstream[7]. Low high-density lipoprotein cholesterol (HDL-C) increases risk of CVD because HDL-C has beneficial properties including clearing LDL-C plaques to be metabolized in the liver and roles in endothelial cell repair[7]. Increased understanding of mechanisms of CVD and metabolic disease would improve diagnosis and treatment, therefore improving health outcomes.

Biological processes of cardiometabolic traits

Cardiometabolic traits are complex and involve processes in multiple tissues including adipose, liver, skeletal muscle, and pancreas[8]. One example of a cardiometabolic trait involving multiple tissues is insulin resistance. Insulin resistance is a complex process that is caused by traits such as central obesity[5] and causes additional cardiometabolic traits including T2D[8]. Insulin resistance occurs when adipose, liver, and muscle become less responsive to insulin which reduces their ability to uptake glucose from the blood. Beta cells in the pancreas compensate by producing more insulin but, without interventions such as lifestyle changes or pharmacological treatment, insulin resistance eventually leads to elevated blood glucose and T2D[8]. While many tissues are involved in cardiometabolic traits, in this dissertation I focus on adipose and liver.

Adipose and cardiometabolic traits

Adipose tissue affects cardiometabolic traits such as body fat distribution, blood cholesterol levels, and insulin resistance through lipid storage and hormone secretion processes[9,10]. Adipose tissue is heterogeneous and contains multiple cell types, including adipocytes, preadipocytes, immune cells, and vascular cells[11]. Adipocytes are a cell type within adipose tissue that store excess lipids through hyperplasia, during which preadipocytes differentiate into adipocytes to increase storage, or hypertrophy, during which adipocytes expand to increase storage[12]. Hypertrophy of adipocytes is associated with hypoxia caused by the increased size and inflammation caused by tissue fibrosis and necrosis[12]. Dysfunctional, hypertrophic adipocytes have limited ability to store excess nutrients resulting in elevated blood levels of glucose and lipids[12]. Excess lipids in the blood can lead to insulin resistance and storage of lipids in visceral fat deposits and organs such as liver and skeletal muscle[12].

Adipose and cell models for component adipose cell types are useful to study adipose biology. Obesity is a leading cause of CVD and metabolic disease, however the body fat distribution of the excess adipose can predict disease risk better than total body fat[5,10]. An increased WHR, a measure of body fat distribution and central obesity, is associated with an increased risk of disease[10]. Body mass index (BMI) is an easily measured trait that is often used to capture estimates of total body fat[5]. Visceral adipose, which is accumulation of fat around internal organs and associated with central obesity, occurs during processes such as insulin resistance[12,13], has also been shown to increase risk of CVD more than subcutaneous adipose[10]. However, enlarged adipocytes from both visceral and subcutaneous adipose have been associated with inflammation and metabolic dysfunction[14]. Compared to other cardiometabolic trait-relevant tissues, subcutaneous adipose can be easy to collect through a needle biopsy or minor outpatient surgical biopsy. In Chapter 2 of this dissertation, I study samples of subcutaneous abdominal adipose from individuals in the METabolic Syndrome In Men (METSIM) study[15]. The METSIM study consists of ~10,000 Finnish men with dense genotype, gene expression, and cardiometabolic trait phenotype data[15]. The ability to integrate and identify associations between datasets makes the METSIM study a useful resource for additional studies.

Cell models of component cell types of adipose such as adipocytes are useful for studies of adipose biology. Adipocyte cell models can be exposed to cardiometabolic trait relevant environments such as excess lipids in the form of free fatty acids, inflammation, and hypoxia to study effects in a controlled environment. Mouse 3T3-L1 adipocytes are a common cell model[16–19], however, it is also useful to study human cell models because species-specific differences have been identified[20]. In Chapters 2 and 3 of this dissertation I use Simpson-Golabi-Behmel Syndrome (SGBS) cells, a well-characterized human, diploid preadipocyte cell model that can be differentiated into mature adipocytes[21,22].

Liver and cardiometabolic traits

Liver tissue is important to cardiometabolic traits through regulation of lipids, glucose, and cholesterol[23]. Dysfunctional adipose leads to excess lipids accumulating in the blood which are taken up by the liver[12]. Dysfunctional liver tissue can contribute to cardiometabolic traits such as insulin resistance through decreased storage of glucose and increased accumulation of lipids[23]. Accumulation of lipids in liver through insulin resistance is a risk factor cardiometabolic traits such as non-alcoholic fatty liver disease (NAFLD)[23]. Like many cardiometabolic traits[24–26], NAFLD demonstrates sex differences in prevalence with a higher risk for men than pre-menopausal women[25]. Sex differences can also contribute to differences in drug metabolism that could inform treatment options[27]. Sex differences in cardiometabolic disease can be caused by genetic factors from the sex chromosomes, epigenetics such as chromatin accessibility, gene regulation, environmental factors, and endogenous factors such as hormones[27]. Studying sex differences in disease relevant tissues could aid identification of sex-specific mechanisms of disease. Compared to adipose tissue, liver tissue collection is more invasive. In Chapter 4 of this dissertation, I study liver samples from male and female deceased organ donors without known disease[28].

Multiple tissues and processes contribute to cardiometabolic traits and many of the molecular mechanisms of these complex processes remain poorly understood. Studying cardiometabolic trait-relevant tissues and cell types in relevant contexts will improve understanding of the molecular mechanisms of disease.

Identifying candidate mechanisms at cardiometabolic trait loci

Cardiometabolic traits develop from both genetic and environmental causes that can interact. Environmental factors that contribute to cardiometabolic traits includes nutrition, physical activity, and smoking[2]. Lifestyle interventions such as diet and increased physical activity have been shown to improve cardiometabolic traits and decrease risk of disease[2,29–31]. However, there is strong evidence for genetic factors contributing to cardiometabolic traits.

Genetics of cardiometabolic traits

Genetic factors can contribute to risk of developing cardiometabolic disease. Using methods such as family and twin studies, heritability estimates range from 30-70% for T2D[32,33], 40-60% for coronary artery disease[34–36], 36-61% for WHR[37], and 22-91% for LDL-C[38,39]. Genetic factors can also interact with other risk factors like sex, age, and environment to increase risk in an additive manner[40,41]. These interactions can explain some of the large variations in heritability estimates[40]. For example, gene-age interactions have been identified for weight, where heritability estimates are low for infants (5-9%) and increased to 74-87% by age 19[42]. Gene-sex interactions have also been identified for traits such as BMI and triglycerides[40]. Heritability estimates show that genetic variation can contribute to cardiometabolic traits.

Genome-wide association studies (GWAS) identify associations between genetic variants and traits. GWAS have identified thousands of loci associated with cardiometabolic traits[43,44] including over 300 for T2D[45–47], over 100 for measure of body fat distribution such as WHR[48,49] with seven loci demonstrating sexual dimorphism[49], and over 400 for blood lipid traits[50,51] with 64 demonstrating sexual dimorphism[51]. Although some GWAS loci are in protein coding regions of the genome and can be directly linked to genes with predicted functional effects, as many as 90% of identified GWAS loci are located in noncoding regions of the genome[52,53]. It is difficult to identify mechanisms of noncoding GWAS loci because there are often multiple candidate variants, genes, cell types, and relevant cellular contexts[53–55]. Noncoding regions can represent regulatory elements that alter gene expression in cell type and context-dependent manners[53].

Context-dependent gene expression

A single change in cellular context can trigger gene expression changes in a large number of genes in order to adapt an organism to the environment. Dynamic gene regulation in response to cellular context is regulated through transcription factors, proteins that bind to regulatory elements to alter transcription[56]. Transcription factor activity can be regulated by altering the activity or abundance of the transcription factor or altering chromatin to make binding more effective[56,57]. Transcription factor activity also can be altered through post translational modifications, such as phosphorylation, which can switch a transcription factor between active and inactive states[56]. Chromatin remodeling that changes accessibility can expose transcription factor binding sites, allowing an active transcription factor to bind to a DNA regulatory element and affect gene expression[56,58].

Transcription factors bind to specific DNA sequences, called transcription factor binding motifs[56]. DNA variants within a region of chromatin accessibility can change binding affinity for a transcription factor, therefore changing the response to a cellular context. Key transcription factors that play a role in adipocyte differentiation include *PPARγ* and *C/EBPα*[59]. A transcription factor responsible for many sex-dependent differences in liver gene expression is *STAT5b*[60]. Identifying context-dependent chromatin accessibility can reveal regulatory elements that alter gene expression, and variants within these regulatory elements could alter transcription factor binding to produce genetic differences in response to a cellular context.

Identifying regulatory elements

While noncoding GWAS loci remain difficult to understand, many are predicted to have regulatory mechanisms that alter gene expression to affect a trait[1]. Gene regulation has been studied using gene expression quantitative trait loci (eQTL) studies. eQTL studies identify variants associated with changes in gene expression and have been performed in many tissues including adipose and liver[61,62]. Noncoding GWAS loci are enriched for colocalized eQTL associations[63], suggesting a regulatory mechanism at these loci. Some GWAS loci colocalize with eQTL in trait-relevant tissues including adipose and liver[61–67], while other GWAS loci colocalize with eQTL found in specific contexts, such as stimulated, but not naïve, immune cells[68]. Despite eQTL localization identifying potential candidate genes, mechanisms of GWAS loci can remain unclear due to a large number of candidate variants at many loci[53]. Therefore, regulatory element and gene expression profiles in cardiometabolic trait-relevant contexts can be used to identify molecular mechanisms at noncoding GWAS loci.

Regulatory elements such as enhancers, silencers, or promoters can be detected by a variety of epigenomic assays including histone modifications and chromatin accessibility profiling[58]. Chromatin accessibility is a known feature of active regulatory elements[58] and can be profiled using sequencing methods such as the Assay for Transposase Accessible Chromatin (ATAC-seq)[69]. ATAC-seq requires less material and time compared to other methods of chromatin accessibility profiling such as DNase hypersensitivity (DNase-seq) and formaldehyde-assisted isolation of regulatory elements (FAIRE-seq)[69]. ATAC-seq is typically performed with 50,000 cells, compared to millions required for other protocols[69]. ATAC-seq can also be performed in a day using a Tn5 transposase enzyme that targets accessible regions to cut and ligate sequencing adaptors in a single step while other protocols often require multiple days[69]. Chromatin immunoprecipitation (ChIP-seq) is another method that can identify genome-wide changes contributing to gene regulation by detecting protein-DNA interactions[70]. ChIP-seq can identify sites where transcription factor proteins directly bind to DNA, however, each transcription factor must be assayed individually and requires a high-quality antibody[70]. ChIP-seq data has been generated for specific transcription factors in adipocytes[71–74] and liver[74,75]. Compared to ChIP-seq, ATAC-seq can capture changes from all transcription factors at once, however ATAC-seq does not identify which transcription factor is acting at a regulatory element[69]. Additional computational methods such as identification of transcription factor binding motifs and experimental methods such as electrophoretic mobility shift assays can identify specific transcription factors involved in a regulatory element identified by chromatin accessibility[53].

Large consortiums such as the Encyclopedia of DNA Elements (ENCODE)[76] and Roadmap Epigenomics Project[77] have identified regulatory maps for many cell types and contexts. ENCODE profiled histone modifications, transcription factor binding, and chromatin accessibility in hundreds of cell and tissue types[76]. Roadmap Epigenomics Project profiled histone modifications, DNA methylation, and chromatin accessibility in 111 cell and tissue types[77]. These profiles are valuable resources for studying regulation, however some cell types and many cell contexts that are relevant to disease remain under-annotated. Annotation of chromatin accessibility, a known feature of active elements[58], in additional disease-relevant contexts can be integrated with other genomic datasets such as eQTL and GWAS to identify regulatory elements that alter gene expression to affect cardiometabolic traits.

Identifying candidate genes for regulatory elements

Linking regulatory elements to genes in disease-relevant contexts can identify targets for drugs that could increase or decrease their function. Regulatory elements can be linked to genes through integration with other genomic data such as eQTL, chromosome conformation capture, and context-dependent gene expression[78]. Each type of genomic data has advantages and disadvantages; therefore, it is useful to use multiple methods. Analyses such as eQTL identify variants associated with differences in gene expression between individuals or contexts[68,79]. Identifying active regulatory elements through chromatin accessibility profiles in cardiometabolic trait-relevant tissues, cell types, and contexts that overlap eQTL associated variants can identify candidate variants and link them to candidate genes[28,53]. However, eQTL can be underpowered to detect associations due to small sample size or missing relevant cell types or contexts[78]. Chromosome conformation capture techniques such as promoter capture Hi-C identify regions of the genome in close contact with each other, including active regulatory elements and promoters[80]. However, chromatin conformation capture can identify large regions that interact resulting in difficulty identifying the active regulatory element[80]. Identifying active regulatory elements through chromatin accessibility profiles in cardiometabolic trait-relevant tissues, cell types, and contexts that overlap chromosome conformation capture regions can link a smaller candidate regulatory element to a candidate gene[28]. Gene regulation can vary by cell type and cell context[53,68], therefore paired differential chromatin accessibility and gene expression profiles in the same conditions can be used as an additional line of evidence to link regulatory elements to differential gene expression. However, there are not well-established methods to link differential chromatin accessibility to differential gene expression even in paired data other than proximity which is indirect.

Each genomic dataset can be used to link regulatory elements to genes. Due to the advantages and disadvantages of different data types, it is useful to use multiple methods. Linking a regulatory element to the same gene through multiple methods increases confidence in the prediction. However, a regulatory element linked to a gene by only one method can merit further investigation. Integrating chromatin accessibility with gene expression and GWAS data can identify candidate regulatory elements and variants that alter gene expression and contribute to a cardiometabolic trait, however functional testing is necessary.

Functional testing of regulatory variants

Variants in regulatory elements that can be linked to a gene and disease trait can be experimentally manipulated to validate predicted genetic mechanisms of disease[53]. There are many approaches to functionally test a candidate regulatory mechanism[53,81]. The primary goal of functional validation such as transcriptional reporter assays, electrophoretic mobility shift assays, or allelic imbalance testing is to identify allelic differences in transcriptional activity. Transcriptional reporter assays such as a luciferase assay test variant alleles in a candidate regulatory region for differences in transcriptional activity of a reporter gene such as luciferase[53]. Allelic imbalance can be tested for with sequencing data such as ATAC-seq chromatin accessibility profiles. An allelic imbalance test can be used at heterozygous sites and tests for disproportionate representation of one allele compared to the other[53].

Identification of variants that contribute to disease could improve prediction of individuals at risk of disease and determine which individuals may respond better to a specific drug treatment[53]. For example, a variant associated with myocardial infarction was found to create a *C/EBP* transcription factor binding site in a liver regulatory element that affected expression of *SORT1*, a gene that alters low-density lipoprotein cholesterol[82]. Identification of this functional variant could be used to identify individuals at risk of myocardial infarction and individuals who may respond better to treatments for high low-density lipoprotein cholesterol specific to the *SORT1* pathway[82].

Aims and overview

In this dissertation, I contribute to defining the impact of genetic variation and cellular context on chromatin accessibility and cardiometabolic traits. I hypothesize based on previous research that variants in context-dependent regions of chromatin accessibility affect gene regulation to contribute to disease traits. In Chapter 2, I identify chromatin accessibility and gene expression that change with adipocyte differentiation. I link context-dependent chromatin accessibility to candidate genes using three approaches. I link context-dependent chromatin accessibility to variants associated with cardiometabolic traits. I also identify a consensus map of chromatin accessibility in 11 adipose tissue samples. In Chapter 3, I describe investigations into adipocyte chromatin accessibility in other cardiometabolic trait-relevant contexts such as free fatty acids, hypoxia, and inflammation. In Chapter 4, I identify sex-biased chromatin accessibility regions that change between males and females in human liver tissue. I link sex-biased chromatin accessibility to variants associated with differential expression in liver and to disease traits. In Chapter 5, I summarize my results, reflect on my research, discuss limitations, and consider future directions. In this dissertation, I identify hundreds of candidate variants in disease-relevant contexts that could help define mechanisms responsible for variation in cardiometabolic traits.

CHAPTER 2: CHROMATIN ACCESSIBILITY AND GENE EXPRESSION DURING ADIPOCYTE DIFFERENTIATION IDENTIFY CONTEXT-DEPENDENT EFFECTS AT CARDIOMETABOLIC GWAS LOCI[1],[2]

Introduction

Genome-wide association studies (GWAS) have identified thousands of loci associated with cardiometabolic traits, yet most mechanisms remain unclear due to unknown functional variants, genes, cell types, and relevant contexts, especially at the large number of noncoding loci[54]. Noncoding loci can regulate gene expression in cell type and context-dependent manners[53]. Some GWAS loci colocalize with gene expression quantitative trait loci (eQTL) in trait-relevant tissues[61,62,64–67], although other GWAS loci colocalize with eQTL found only in one context, such as stimulated, but not naïve, immune cells[68]. Therefore, mapping transcriptional regulatory elements and gene expression in disease-relevant contexts can be used to characterize molecular mechanisms of GWAS loci. Enhancers and other regulatory elements can be detected by identifying regions of chromatin accessibility[58] using sequencing methods such as the Assay for Transposase Accessible Chromatin (ATAC-seq)[69]. Chromatin accessibility in cardiometabolic-relevant cell types and contexts can be integrated with GWAS and eQTL data to

[1] The work in this chapter has been previously published and adapted for this dissertation chapter[83]. The citation is: Perrin HJ, Currin KW, Vadlamudi S, Pandey GK, Ng KK, Wabitsch M, Laakso M, Love MI, Mohlke KL. Chromatin accessibility and gene expression during adipocyte differentiation identify context-dependent effects at cardiometabolic GWAS loci. PLoS Genet. 2021 Oct 26;17(10):e1009865. doi: 10.1371/journal.pgen.1009865. PMID: 34699533; PMCID: PMC8570510.

[2] Hannah Perrin performed chromatin accessibility assays, analyzed data, generated figures, and wrote and edited the manuscript and the response to reviewers.

identify regulatory elements and variants that alter gene expression to affect cardiometabolic traits.

Adipose tissue influences cardiometabolic traits such as body fat distribution, insulin sensitivity, blood cholesterol levels, and inflammation through its roles in lipid storage and hormone secretion[9,10]. Hundreds of GWAS loci for cardiometabolic traits are colocalized with eQTL in adipose tissue[61,64,65], and variants at GWAS loci for some cardiometabolic traits, such as waist-to-hip ratio adjusted for body mass index (BMI) and high-density lipoprotein (HDL) cholesterol, are overrepresented in transcriptional regulatory elements in adipose tissue[84,85]. At a subset of colocalized GWAS-eQTL signals, adipose tissue gene expression may mediate the effect of the genetic variant on GWAS traits[61]. Adipose is a heterogeneous tissue that contains multiple cell types, including adipocytes, preadipocytes, immune cells, and vascular cells[11]. Adipose tissue stores lipids through either hyperplasia, during which preadipocytes differentiate into mature adipocytes to store excess energy, or hypertrophy, during which existing adipocytes expand to store excess energy[12]. Thus, identifying variants with regulatory effects at specific stages of adipocyte differentiation may uncover additional mechanisms at GWAS loci for cardiometabolic traits.

Genetic and environmental variation between individuals can contribute to differences in chromatin accessibility[76]. Chromatin accessibility maps generated from multiple individuals can capture accessible regions that reflect genetic effects and diverse environmental contexts. Existing human adipose tissue chromatin accessibility maps are comprised of data from one to six individuals and differ by tissue donor characteristics (e.g. BMI, age, sex), adipose depot, tissue extraction site, and storage conditions[76,84,86]. Given the cell type heterogeneity of tissue samples, it is also useful to characterize the component cell types in controlled environments.

Chromatin accessibility during adipogenesis has been studied in models such as mouse 3T3-L1 cells[87], however additional studies in human models could improve interpretation of human non-coding genetic variants. Simpson-Golabi-Behmel Syndrome (SGBS) cells are a well-characterized diploid preadipocyte cell model that can be differentiated into mature adipocytes and is useful for studying adipocyte differentiation[21,22].

In this study, we identified differences in chromatin accessibility and gene expression between preadipocytes, immature adipocytes, and mature adipocytes in SGBS cells. In addition, we generated a consensus map of subcutaneous adipose tissue chromatin accessibility using 11 samples obtained from METabolic Syndrome in Men (METSIM) participants[15]. We used three methods to link differentially accessible regulatory elements to candidate genes and identified variants at cardiometabolic GWAS loci that resided in elements more accessible in preadipocytes or adipocytes. Finally, we identified variants at the *SCD* and *EYA2* loci that showed context-dependent and/or allelic effects on transcriptional activity, detecting potential mechanisms by which specific variants alter gene expression to affect cardiometabolic traits.

Results

Changes in chromatin accessibility across adipocyte differentiation timepoints identify context-dependent regulatory elements

We profiled chromatin accessibility during adipocyte differentiation with ATAC-seq in SGBS cells[69,88]. We analyzed a final set of ten replicates of preadipocytes (D0), ten replicates of immature adipocytes differentiated for four days (D4), and five replicates of mature adipocytes differentiated for fourteen days (D14) (Figure 2.1A and Table 2-1). Our libraries had ~33-156 million filtered reads each, and showed high quality, with an average transcription start site

(TSS) enrichment of 6.8, and an average fraction of reads in peaks (FRiP) of 48.5%. To test for differences in chromatin accessibility between timepoints, we generated a set of 147,587 accessible chromatin regions (ATAC-seq peaks) at any time point (Table 2-2) by merging the top 100,000 consensus peaks for each time point (ranked by median peak p-value across replicates, see Methods). Principal component analysis (PCA) showed that replicates clustered by differentiation timepoint, with preadipocytes and adipocytes separated by the first principal component, which explained 74% of the variance.

To predict regulatory elements involved in adipocyte differentiation, we identified differentially accessible peaks, hereafter called 'context-dependent peaks', between each pairwise comparison of the three timepoints ($\log_2$ fold change (LFC)>1; false discovery rate (FDR)<5%; Table 2-3). Based on the 10,000 context-dependent peaks with the most significant difference in any comparison, a heatmap showed that replicates clustered by timepoint (Figure 2.1B). Most (86%) of the changes in chromatin accessibility between D0 and D14 were observed by D4, and only 233 peaks were specifically more accessible in mature adipocytes (D14>D0 and D14>D4), suggesting that chromatin accessibility changes early after the initiation of differentiation and remains largely stable between D4 and D14. To characterize the major differences, we identified context-dependent peaks more accessible in preadipocytes in both comparisons (D0>D4 and D0>D14; 18,244 peaks), hereafter called 'preadipocyte-dependent peaks', and context-dependent peaks more accessible in immature and mature adipocytes (D4>D0 and D14>D0; 15,919 peaks), hereafter called 'adipocyte-dependent peaks'. In analyses described below, we used the preadipocyte-dependent and adipocyte-dependent peaks for enrichment analyses and general comparisons between preadipocytes and adipocytes, and we used context-dependent peaks from all pairwise comparisons to identify regulatory elements linked to genes and GWAS loci.

We evaluated the relevance of context-dependent peaks for biological processes and transcription factors known to be involved in adipocyte differentiation. Preadipocyte-dependent peaks were enriched (P<1x10-10) near genes associated with roles in several cell cycle processes, including positive regulation of DNA replication. Mature adipocyte-dependent peaks were enriched near genes with roles in cardiovascular development. Adipocyte-dependent peaks were enriched near genes with roles in several metabolic processes, including response to insulin, regulation of fatty acid oxidation, and intracellular lipid transport. In addition, preadipocyte-dependent peaks were enriched (P<1x10-5) for transcription factor motifs for TEAD and GATA, which inhibit adipocyte differentiation[89,90], while adipocyte-dependent peaks were enriched for motifs of transcription factors that promote adipogenesis, such as CEBP, PPAR, and LXR[12,91] and transcription factors involved in glucose metabolism such as GRE[92]. Thus, adipocyte- and preadipocyte-dependent peaks are found near genes and contain transcription factor motifs relevant to their cell contexts, increasing confidence that these peaks capture relevant biology. Although genomic proximity between regulatory elements and genes is a strong predictor of a regulatory relationship[93], regulatory elements may not always affect the nearest genes.

To compare these SGBS peaks to adipose tissue peaks, we expanded our previous set of adipose tissue ATAC-seq profiles[84] from 3 to 17 samples that fulfilled sequencing quality thresholds (Methods, Table 2-4 and Figure 2.6). In the 17 tissue samples, we identified 79,598 consensus adipose tissue peaks present in three or more samples. After removing 6 outlier samples identified using PCA, overlap with adipose regulatory elements, and other factors, we also identified 51,855 consensus adipose tissue peaks using 11 adipose tissue samples. The 11-sample peak set had a higher percentage of peaks within the Roadmap Epigenomics Project[77] adipose nuclei enhancers and promoters (45% enhancer, 39% promoter) compared to the 17-sample peak set (34% enhancer, 28% promoter), and a similar percentage compared to our previous 3-sample peak set[84] (49% enhancer, 39% promoter) (Figure 2.1D). We proceeded with the 11-sample consensus adipose peak set for further analyses because it provides higher consistency with Roadmap adipose enhancers and promoters relative to the 17-sample set and may capture more genetic and environmental variation in chromatin accessibility than the 3-sample set.

To determine if context-dependent SGBS peaks marked previously annotated adipose regulatory elements, we compared the SGBS peaks to Roadmap Epigenomics Project adipose nuclei chromatin states[77] and to the 11-sample adipose tissue peaks. A higher percentage of adipocyte-dependent peaks were found within Roadmap adipose nuclei enhancers and promoters (60% enhancer, 3.9% promoter) compared to preadipocyte-dependent peaks (12% enhancer, 0.73% promoter) (Figure 2.1E). Similarly, 36% of adipocyte-dependent peaks overlapped (shared at least 1 base) adipose tissue peaks, while only 1.8% of preadipocyte-dependent peaks overlapped adipose tissue peaks, consistent with adipose tissue containing more adipocytes than preadipocytes[11,94]. Peaks found in SGBS and adipose tissue may have more relevance to adipose biology than peaks found in SGBS cells alone. These results show that our adipocyte-dependent and consensus adipose tissue peaks demonstrate strong similarity with existing adipocyte promoters and enhancers.

Changes in gene expression across adipocyte differentiation

We generated RNA-seq data from six replicates of SGBS preadipocytes (D0), six replicates of immature adipocytes differentiated for four days (D4), and four replicates of mature adipocytes differentiated for fourteen days (D14) (Figure 2.1A). We generated ~36-56 million filtered reads overlapping transcripts per replicate (Table 2-5) and identified 18,299 expressed genes (median normalized count >=1 across libraries). PCA showed that replicates clustered by differentiation timepoint, with preadipocytes and adipocytes separated by the first principal component, which explained 54% of the variance.

To identify changes in gene expression during adipocyte differentiation, we identified genes differentially expressed between each pairwise comparison of the three timepoints (LFC>1; FDR<5%; Table 2-3). A heatmap of these 'context-dependent genes' showed that replicates clustered by timepoint (Figure 2.1C). In addition, we identified context-dependent genes that were observed in multiple timepoint comparisons. In contrast to context-dependent chromatin accessibility, for which 86% of changes between D0 and D14 were observed already by D4, only 1,282 of 2,107 (61%) context-dependent genes between D0 and D14 were observed already by D4. Although further analysis is needed, this result is consistent with previous studies that identified changes in chromatin accessibility that occurred earlier during adipocyte differentiation and remained more stable than changes in gene expression[87,95].

We tested context-dependent genes for enrichment of biological processes known to be involved in adipocyte differentiation. Genes expressed more strongly in preadipocytes than adipocytes were enriched (P<1x10-10) for several cell cycle processes including cell cycle regulation and nuclear division. Genes expressed more strongly in adipocytes than preadipocytes showed enrichment (P<1x10-10) for several differentiation and metabolic processes such as response to insulin, glucose homeostasis, fatty acid metabolic processes, and lipid biosynthetic processes. We also identified context-dependent transcription factors whose binding motifs were enriched in context-dependent peaks, including preadipocyte-dependent GATA family members that had motifs enriched in preadipocyte-dependent peaks, and the adipocyte-dependent gene PPARG whose motifs were enriched in adipocyte-dependent peaks. Adipocyte-dependent genes also included known adipocyte-dependent genes such as ADIPOQ[96]. These results indicate that the context-dependent genes have functions relevant to the corresponding cell types.

Three approaches to link genes to context-dependent peaks

Linking context-dependent peaks to genes remains challenging because most peaks are located in non-coding regions with multiple genes nearby. Approaches to predict genes affected by a peak have varied sensitivity and specificity[53], thus we used three approaches to identify additional genes and to gain confidence in genes identified by more than one method. The three approaches used to link context-dependent peaks to genes were: overlap with adipocyte promoter capture Hi-C[97,98], overlap with adipose eQTL variants[61], and context-dependent expression of genes linked by either of the first two approaches (Figure 2.2A-C).

In the first approach, we identified context-dependent peaks that overlapped adipocyte promoter capture Hi-C regions[97,98] (overlap>=1 base pair, Figure 2.2A). We identified 14,080 peaks linked to 9,080 genes (28,696 peak-gene pairs). We investigated the extent to which increasing the overlap threshold between peaks and Hi-C fragments would change our results. Of the 14,594 peak-Hi-C fragment overlapping pairs (some peaks overlap more than one Hi-C fragment and vice versa), 12,380 (85%) have over 50% of peak bases within the Hi-C fragment and 10,329 (71%) have the entire peak within the Hi-C fragment, suggesting that we would obtain similar results using more strict overlap thresholds. Of the 14,080 peaks, 3,436 were preadipocyte-dependent and 4,873 were adipocyte-dependent (5,771 were context-dependent but not preadipocyte- or adipocyte-dependent, hereafter called ‘other context-dependent peaks’). We identified more links for adipocyte peaks than for preadipocyte peaks, consistent with our use of Hi-C data only from mature adipocytes, not preadipocytes. Most distances from peaks to gene TSS linked by Hi-C (85%) were within 500 kb, and 97% were within 1.2 Mb (Figure 2.2D).

In the second approach, we identified context-dependent peaks that overlapped adipose eQTL signals[61], defining each signal as all variants in high linkage disequilibrium with a lead eQTL variant (r2>0.8, Figure 2.2B). Of 3,002 peaks linked to 2,369 genes (3,794 peak-gene pairs), 805 linked from preadipocyte-dependent peaks and 996 from adipocyte-dependent peaks (1,201 linked from other context-dependent peaks). The larger number of links from adipocyte peaks than preadipocyte peaks is consistent with use of eQTL from adipose tissue, which contains more adipocytes than preadipocytes[11,94]. We identified 4,549 adipose eQTL variants within the context-dependent peaks; these variants could be part of the mechanisms regulating expression level of the corresponding genes. Most distances from peaks to gene TSS linked by eQTL (87%) were within 200 kb, and all were within 1 Mb, the distance threshold used for the eQTL study (Figure 2.2D).

In the third approach, we identified context-dependent peaks linked to a gene by Hi-C or eQTL overlap for which the gene also showed context-dependent expression between any timepoint comparison (Figure 2.2C). Of the 14,080 peaks identified by Hi-C, 4,462 peaks also linked to a context-dependent gene (1,000 linked from preadipocyte-dependent peaks and 1,681 linked from adipocyte-dependent peaks, 1,781 linked from other context-dependent peaks). Of the 3,002 peaks identified by eQTL, 720 contained a context-dependent gene (134 linked from preadipocyte-dependent peaks, 298 linked from adipocyte-dependent peaks, 288 linked from context-dependent but not preadipocyte- or adipocyte-dependent peaks).

Each approach to link regulatory peaks to genes can add an additional level of evidence to support the predicted gene target. We next identified peaks linked to the same gene through more than one approach. Of 16,076 total peaks linked to a gene through at least one of the three approaches, 78 peaks were linked to the same gene through all three approaches and 5,145 peaks were linked to the same gene through two or more approaches (Figure 2.2E). Of the 78 peaks linked to 59 genes through all three approaches, interesting candidate regulatory elements include four peaks linked to CDKN2B, whose gene product has known roles in cell cycle control and whose regulation has been linked to coronary artery disease[99,100]. Of the 5,145 peaks linked to 1,670 genes through at least two approaches, 1,143 linked from preadipocyte-dependent peaks and 1,945 linked from adipocyte-dependent peaks (2,057 linked from other context-dependent peaks). Although peaks linked by all three approaches have the most supporting evidence, to prevent overlooking interesting candidates we considered peaks linked by two or more methods when evaluating candidates for functional evaluation.

Trait heritability enrichment within context-dependent peaks

We used stratified LD score regression[101] to compare heritability enrichment for selected cardiometabolic traits in preadipocyte-dependent peaks, adipocyte-dependent peaks, and bulk adipose tissue peaks. Given that preadipocyte-dependent and adipocyte-dependent peaks cover a small portion of the genome (~0.45%), we also ran stratified LD score regression on the top 100,000 consensus peaks (ranked by median peak p-value across replicates) in each SGBS differentiation day. For comparison, we also ran stratified LD score regression using the adipose tissue peaks.

Different traits were enriched in adipocyte-dependent and preadipocyte-dependent peaks. For waist-hip ratio adjusted for BMI (WHRadjBMI), we observed significant enrichment for adipocyte-dependent peaks (z-score=4.7, P<1.2x10-6) and adipose tissue peaks (z-score=5.2, P<1.0x10-7) but not for preadipocyte-dependent peaks (z-score=-1.1, P<0.86) (Figure 2.3A). Results were consistent for the top 100,000 consensus peaks in each SGBS differentiation day; the modest enrichment in D0 peaks could be partly due to peaks shared between timepoints. We also observed nominal enrichment for HDL heritability in adipocyte-dependent and adipose tissue peaks. In contrast, we observed significant enrichment for coronary artery disease (CAD) heritability in SGBS D0 (z-score=2.8, P<2.9x10-3) and adipose tissue (z-score=3.3, P<5.4x10-4), with weaker and still nominally significant enrichment in D4 (z-score=2.4, P<9.0x10-3) and D14 (z-score=2.3, P<0.01); the lack of enrichment in preadipocyte-dependent and adipocyte-dependent peaks may be due to their low genomic coverage. All peak sets showed less heritability enrichment relative to baseline for rheumatoid arthritis, a negative control, except for adipocyte-dependent peaks, which showed nominal enrichment (z-score=1.8, P<0.04), suggesting that adipocytes may have moderate relevance for this trait. We did not observe enrichment of BMI heritability in any peak set, consistent with our previous finding that BMI GWAS loci were not enriched in adipose tissue or SGBS peaks[84] and with findings from other studies that BMI loci are enriched in central nervous system cell types and pathways[101,102]. A complementary approach using all traits in the GWAS catalog[44] grouped by Experimental Factor Ontology terms showed similar results (Figure 2.3B). The most enriched terms for adipocytes included waist-hip ratio, cholesterol, inflammatory traits, and birthweight, whereas the most enriched terms for preadipocytes included atrial fibrillation and inflammatory traits. We also observed enrichments for traits with less apparent, but established connections to cardiometabolic traits, including forced expiratory volume, a measure of lung function that has been shown to be lower in individuals with metabolic syndrome and high body fat percentage[103,104], and intraocular pressure, which has been shown to be higher in individuals with metabolic syndrome and markers of obesity[105,106]. Taken together, we found that peaks in adipocytes contribute more to heritability of WHRadjBMI, whereas preadipocytes may contribute more to heritability of CAD, though to a lesser degree.

Fine-mapping of GWAS variants using context-dependent peaks and allelic imbalance

To identify genetic variants that may have context-dependent effects on disease-relevant traits, we identified distinct signals from the GWAS catalog[44] (see Methods) for which a proxy variant (LD r2>0.8) is located within a context-dependent peak. Of 4,954 context-dependent peaks that overlapped GWAS signals, 1,448 were preadipocyte-dependent and 1,461 were adipocyte-dependent.

At some GWAS loci, these context-dependent peaks can be linked to genes. We observed 4,284 peak-gene pairs that overlapped GWAS variants, and 799 of these pairs, representing 659 unique peaks, were supported by two or more approaches (Figure 2.3C). Of these 659 peaks, 265 were adipocyte-dependent, 143 were preadipocyte-dependent, and 251 were other context-dependent peaks. Of these 659 peaks, 191 (29%) overlapped adipose tissue peaks, which generally had weaker signals than the SGBS peaks. At one locus, we identified two peaks more accessible in adipocytes that overlap adipose eQTL variants for ADIPOQ (peak96641: rs76071583; peak96640: rs143257534), which also showed adipocyte-dependent expression.

These peaks also overlap adipose consensus peaks and GWAS variants associated with adiponectin levels[107], including rs76071583, previously shown to exhibit allelic differences in binding of the transcription factor CEPB-α and transcriptional activity in adipocytes[108]. CEBPA has higher expression in adipocytes than preadipocytes (D4>D0 LFC=8.6, D14>D0 LFC=9.2), consistent with the context-dependent regulatory effect.

To identify GWAS variants that may alter chromatin accessibility at different stages of differentiation, we also identified variants exhibiting allelic imbalance (AI) in ATAC-seq reads across SGBS technical replicates. Because SGBS cells originate from one individual, we could only test for AI at heterozygous variants in one individual. We identified 574, 996, and 489 variants showing significant AI (FDR<5%) on D0, D4, and D14, respectively, and 582 AI variants in 454 context-dependent peaks, including 90 peaks that harbored more than one AI variant. Of the 454 context-dependent peaks, 64 were linked to a target gene by two approaches, 55 contained GWAS variants that exhibited AI, and 13 linked to both a target gene and GWAS variant. At an example with both types of data, a variant (rs11039149) that showed significant AI in days 4 and 14 was found within a peak more accessible in D4 compared to D0 (peak23801) and is an eQTL variant for the adipocyte-dependent gene NR1H3. The more accessible allele rs11039149-G is associated with lower NR1H3 expression. rs11039149 is a GWAS variant for HDL cholesterol[50] and proinsulin[109]. NR1H3 has previously been shown to be involved in lipid transport[110], and one or more of these variants could alter NR1H3 expression and affect associated metabolic traits. Combining ATAC-seq AI, context-dependent peaks, and target genes helps connect variants to regulatory elements and genes and can identify variants with context-dependent effects on gene regulation.

Functional evaluation of candidate regulatory elements reveals context- and allele-dependent mechanisms

Of the 659 context-dependent peaks that we linked to target genes and GWAS signals, we tested two for allele-dependent effects on transcriptional activity using reporter gene assays in SGBS preadipocytes and 12-day differentiated adipocytes. At a first GWAS locus for palmitoleic acid[111], we identified an adipocyte-dependent peak (Figure 2.4A, peak19405; D4>D0: LFC=3.8; D14>D0: LFC=2.8) that we linked to the gene *SCD*, encoding Stearoyl-CoA Desaturase, through two approaches, overlap of the peak with an adipose eQTL variant (rs603424, P=1.6x10-9) associated with *SCD*[61] and adipocyte-dependent expression of *SCD* (D4>D0: LFC=6.3; D14>D0: LFC=8.2) (Figure 2.4A). *SCD* codes for an enzyme involved in fatty acid synthesis[112]. Peak19405 also overlaps a consensus adipose tissue peak and contains rs603424, the G allele of which is associated with higher *SCD* expression[61] and higher palmitoleic acid[111]. We tested a 592-bp region spanning the majority of peak19405 for allele-dependent functional effects. In adipocytes, the construct containing the rs603424-G allele demonstrated significantly increased transcriptional activity compared to the construct containing the rs603424-A allele (forward P=0.003, reverse P=0.0001; Figure 2.4B), consistent with the direction of effect observed in the adipose eQTL. Together, these results suggest that in adipocytes but not preadipocytes, rs603424-G increases transcriptional activity of *SCD* to increase palmitoleic acid levels.

At a second GWAS locus for type 2 diabetes[113], we identified two candidate regulatory elements and tested both for allele-dependent effects on transcriptional activity. One candidate is an adipocyte-dependent peak (Figure 2.5A, peak81750, containing rs55966194, D4>D0: LFC=4.2 and D14>D0: LFC=3.1) that we linked to *EYA2*, encoding Eyes Absent Transcriptional

Coactivator and Phosphatase 2, through colocalization with an adipose eQTL (rs55966194, P=6.0x10-10)[61] and adipocyte-dependent expression of the linked gene (D4>D0: LFC=1.7; D14>D0: LFC=1.4) (Figure 2.5A). *EYA2* codes for a protein that has been linked to adipocyte lipolysis[114]. Also, at this locus, a second candidate regulatory element is an adipose-specific peak not detected in SGBS and which contains variant rs59791349, which is a proxy variant for an adipose eQTL for *EYA2*[61] and GWAS locus for type 2 diabetes[113]. The C alleles for both rs55966194 and rs59791349 are associated with higher *EYA2* expression and increased risk of type 2 diabetes. We tested regions spanning the majority of each peak for allele-dependent transcriptional activity. The 419-bp region for adipocyte-dependent peak81750 containing the rs55966194-C allele demonstrated modest allelic differences only in the reverse orientation (P=0.06, Figure 2.5B), whereas the 288-bp region for the adipose peak containing rs59791349-C demonstrated significantly higher transcriptional activity than the rs59791349-T allele in both orientations and both cell types (adipocytes forward P=0.0029, adipocytes reverse P=0.0058; preadipocytes forward P=0.0008, preadipocytes reverse P=0.0015; Figure 2.5C). The allelic differences in transcriptional activity were consistent with the direction of effect of the adipose eQTL. These results suggest that in both preadipocytes and adipocytes, rs59791349-C increases transcriptional activity of *EYA2* to increase risk of diabetes. Altogether, the experiments at these two loci demonstrate that context-dependent peaks can, but do not always, predict allele-dependent transcriptional activity, as other mechanisms may be involved. These results also suggest the value of using both cell type-specific and tissue-derived regulatory elements to identify functional regulatory variants.

Discussion

In this study, we generated chromatin accessibility and gene expression profiles for preadipocytes, immature adipocytes, and mature adipocytes and identified context-dependent peaks during adipocyte differentiation as candidate regulatory elements. We linked these regulatory elements to candidate genes using three approaches and identified context-dependent regulatory elements at GWAS loci. Our consensus subcutaneous adipose tissue peak map based on profiles from 11 individuals provided a resource to expand on existing human adipose peak maps[86,115,116] and to prioritize among peaks from the SGBS cell model. Finally, we identified 659 context-dependent regulatory elements at GWAS loci that were linked to genes and showed through functional tests that elements can exhibit context-dependent allelic differences in transcriptional activity, identifying plausible disease mechanisms.

Chromatin accessibility profiles differ between samples for biological and technical reasons. Biological reasons can include cell type and cell context. A technical source of variation between our profiles could be due to heterogenous sequencing protocols with a mix of paired-end, single-end, and variable read lengths. We addressed the heterogenous sequencing protocols in our analyses as described in Methods, but it could contribute to differences between libraries. Reassuringly, our SGBS ATAC-seq libraries cluster by day despite differences in sequencing parameters. Additionally, while SGBS cells are a useful human adipocyte model, some aspects of the chromatin accessibility profile could be due to the cells growing in culture or the overgrowth syndrome disease state that allows the cells to grow without being transformed. To address these limitations, we identified SGBS peaks that overlapped adipose tissue peaks, including the peak we tested at the *SCD* locus. Although it remains challenging to compare between species, we observed enrichment of motifs for well-known adipogenesis transcription factors CEBP, PPAR, and RXR within adipocyte-dependent peaks, consistent with a study of changes during adipogenesis in a 3T3-L1 mouse line[87].

Our differential analyses of peaks and gene expression profiles between timepoints suggest that most peak changes occur between D0 and D4 and remain stable between D4 and D14, while a larger proportion of gene expression changes occur between D4 and D14. The observation that peak changes occur early and remain largely stable is consistent with a previous study that found a majority of chromatin accessibility changes in a 3T3-L1 mouse-derived adipocyte cell line occurred between two and four hours after the initiation of differentiation[87]. The observation that gene expression may continue to change throughout later stages of differentiation is consistent with a study that showed gene expression changing between 7-day intervals up to day 21 in human adipose-derived stromal cells[95]. After initial analyses suggested that few context-dependent peaks arose between D4 and D14, we investigated chromatin accessibility at an earlier timepoint of immature adipocytes differentiated for two days (D2). Preliminary analysis of D2 also showed that no peaks were differential between D2 and D4, so we did not generate further D2 data. Similarities between D4 and D14 also led us to focus on the subsets of context-dependent peaks that were specific to preadipocytes (D0>D4 and D0>D14) or adipocytes (D4>D0 and D14>D0), rather than the limited number that were specific to mature adipocytes (D14>D0 and D14>D4).

We used two approaches to link context-dependent peaks to genes: overlap with existing adipocyte promoter capture Hi-C regions and with known adipose eQTL variants, and we determined which of these linked genes also showed expression differences between differentiation timepoints. Promoter capture Hi-C has the advantage of identifying direct connections between regulatory elements and genes, even over large distances. However, physical proximity does not necessarily imply a regulatory relationship, and the data we used was for adipocytes, not preadipocytes, and therefore could have detected connections for the D4 and D14 timepoints better than for D0. While most promoter capture Hi-C fragments have high resolution (median ~3 kb in the analyzed dataset), the location of restriction sites in the genome limits resolution for some fragments (~10% of fragments had >10 kb resolution). Our second approach based on overlap with adipose eQTL variants has the advantage that the identified variants are associated with differences in gene expression. Two disadvantages of the eQTL approach are that eQTL studies may be underpowered, so not all associations are discovered, and that the adipose tissue used in the eQTL study is comprised of multiple cell types, not only adipocytes. Although adipose tissue is heterogenous, it is known to contain more adipocytes than preadipocytes[11,94]. Therefore, the eQTL method also could have detected connections for the D4 and D14 timepoints better than for D0. Peaks linked to genes by eQTL tended to be closer to the TSS of the linked gene compared to Hi-C, partially due to the shorter distance window used in the eQTL data than the Hi-C data. To incorporate differential gene expression into the identification of peak-to-gene links, we initially considered using proximity between context-dependent peaks and context-dependent genes. However, proximity is indirect and requires selecting an arbitrary threshold for maximum distance between peak and gene. Thus, we used context-dependent gene expression as additional supporting evidence for links made by other methods. Although indirect, context-dependent genes have the advantage of being observed in the same cell model and at the same timepoints, and can help determine if a regulatory element has a positive or negative effect on gene expression. Due to the advantages and disadvantages of the different approaches, the largely different peak-to-gene links detected were not surprising. Using multiple approaches to link regulatory elements to candidate genes can overcome the limitations of each approach, and genes identified by multiple methods can increase confidence, although genes linked by even a single method merit further investigation.

We used AI in SGBS ATAC-seq reads to provide suggestive evidence that GWAS variants may alter chromatin accessibility at different stages of adipocyte differentiation. Although we only tested AI at heterozygous variants from one individual, which limits heterozygous sites available for testing, we identified 55 peaks containing GWAS variants that exhibited AI, 13 of which were linked to genes. ATAC-seq in additional cell lines with diverse genotypes would improve the ability to detect AI. Previous studies have mapped AI and chromatin accessibility QTL in different contexts[68,117,118], which allowed for testing of more variants and identification of more robust context-dependent genetic effects on gene regulation. Our results demonstrate that AI in ATAC-seq reads from one individual can be used to predict regulatory variants, although identifying AI in larger sample sizes would lead to more comprehensive and robust results and more genetic variants.

We followed up context-dependent regulatory elements at two GWAS loci by testing variants for effects on context-dependent transcriptional reporter gene activity. Due to the bias towards adipocytes of our methods to link peaks to genes, we focused on regulatory elements more accessible in adipocytes. At *SCD*, we observed consistent evidence of context- and allele-dependent transcriptional activity among technical replicates. The regulatory element that was more accessible in adipocytes contained an allele associated with increased adipose tissue expression of *SCD*[61] and increased palmitoleic acid[111]. *SCD* codes for an enzyme involved in fatty acid synthesis[112], therefore increased *SCD* expression is a likely mechanism to increase palmitoleic acid levels. In reporter assays, the element showed higher transcriptional activity in adipocytes than preadipocytes, and the allele associated with higher adipose *SCD* expression showed higher transcriptional activity, only in adipocytes. These data suggest that the regulatory element we identified increases *SCD* expression to increase palmitoleic acid levels in adipocytes.

At the second locus we examined, *EYA2*, the results are more complex. We identified two candidate regulatory elements, one that was adipocyte-dependent and one that was present in the consensus adipose tissue map. Both regulatory elements contained variants associated with adipose tissue expression of *EYA2*[61] and type 2 diabetes[113]. Both regulatory elements demonstrated higher expression of the reporter gene in adipocytes than preadipocytes, consistent with the context in which one element was more accessible and with the large proportion of adipocytes in adipose tissue[11,94]. However, only the consensus adipose element demonstrated clear allele-dependent transcriptional activity. This result demonstrates that, while identifying loci with context-dependent peaks linked to genes and traits still is useful for identifying candidates, it does not mean the identified variant is responsible. However, the variant within the adipocyte-dependent peak at this locus may still exhibit allelic effects on regulatory activity that are not detectible in *in vitro* transcriptional reporter assays. For the *EYA2* locus, our adipose consensus map guided us to investigate an additional candidate regulatory element that demonstrated an allele-dependent effect on transcriptional activity. *EYA2* codes for a transcriptional coactivator that has been linked to many developmental processes and adipocyte lipolysis, consistent with a role in adipocyte biology and metabolic traits[114,119]. Our reporter assays demonstrate allelic differences in transcriptional activity for elements at two loci, however, additional experiments are needed to validate specific regulatory elements within these peaks in the context of chromatin accessibility and the effect on regulation on the predicted gene.

This study extends our previous study that reported ATAC-seq peaks in SGBS cells and adipose tissue from three individuals[84]. Genetic variation contributes to differences in peaks, so we profiled adipose tissue in additional individuals to capture peaks that could have been missed in fewer samples due to genetic variants or environmental/physiological differences between individuals. In general, ATAC-seq data from frozen adipose tissue demonstrated lower quality than our SGBS preadipocytes and other frozen tissues[88,120,121], despite our efforts to optimize library preparation with different buffers, detergents, and ratios of transposase to nuclei. Freezing has been shown to affect ATAC-seq library quality and comparisons of ATAC-seq profiles in samples using various freezing methods suggest cryopreserved tissue demonstrated higher quality than flash-frozen tissue[120]. High lipid content could also have affected adipose tissue profile quality, as adipose tissue has a high ratio of adipocyte cells[11,94] and lipid content somewhat affected ATAC-seq in cultured SGBS cells, as fewer D14 adipocyte samples met QC thresholds compared to D0 and D4 cells, despite being cultured and processed in parallel. The consensus map of adipose peaks based on the 11 samples of at least moderate quality showed similar overlap with adipocyte nuclei promoters and enhancers as our previous map based on three samples, but the inclusion of additional samples should make the 11-sample consensus map more robust.

Overall, we demonstrated that context-dependent chromatin accessibility identifies context-dependent regulatory elements that can aid understanding of mechanisms behind cardiometabolic traits. By identifying adipocyte differentiation context-dependent regulatory elements and linking them to genes and GWAS traits, we filtered from 58,387 context-dependent regulatory elements to 659 elements with a candidate mechanism. Additional study of these regulatory elements could lead to a better understanding of the role of adipocytes and adipocyte differentiation in cardiometabolic disease traits as well as other relevant traits we identified through enrichment analyses such as lung function. This could also be applied to other adipocyte contexts to identify additional context-dependent mechanisms.

Methods

Ethics statement:

The Ethics Committee of the University of Eastern Finland in Kuopio and the Kuopio University Hospital approved the METSIM study and it was carried out in accordance with the Helsinki Declaration. Formal written consent was obtained from METSIM participants.

Cell culture:

SGBS cells[21] were generously provided by Dr. Martin Wabitsch (University of Ulm) and cultured as previously described[122]. Briefly, we cultured SGBS preadipocytes in serum-containing basal medium (DMEM:F12 + 33uM biotin + 17uM pantothenate) with 10% FBS until confluent, then rinsed in phosphate-buffered-saline (PBS) and differentiated for four days in medium supplemented with 0.01 mg/mL transferrin, 20 nM insulin, 200 nM cortisol, 0.4 nM triiodothyronine, 50 nM dexamethasone, 500 uM IBMX, and 2 uM rosiglitazone. After four days, we maintained differentiated SGBS cells in basal medium supplemented with 0.01 mg/mL transferrin, 20 nM insulin, 200 nM cortisol, 0.4 nM triiodothyronine until harvested. HEK293T cells (ATCC, Manassas, VA) were grown in DMEM supplemented with 10% FBS.

Adipose tissue:

Human subcutaneous abdominal adipose tissue biopsies were obtained from METabolic Syndrome in Men (METSIM)[15] participants as previously described[65]. Adipose tissue was obtained through either a needle or surgical biopsy and flash frozen and stored at -80oC until use.

ATAC-seq library preparation:

We profiled chromatin accessibility in SGBS cells at D0, D4, and D14 of adipocyte differentiation following the omni-ATAC-seq protocol[88] using unique, dual-barcoded indices. We isolated nuclei and used a cell countess to aliquot 50,000 nuclei per library. After initial optimization of Tn5:nuclei ratios, we proceeded with 5 uL of Tn5 per library, some early libraries were prepared with 2.5 uL of Tn5 as indicated. For adipose tissue samples we used the original or omni-ATAC-seq protocol[69,88] as indicated. We cleaned the transposase reaction and final library with Zymo DNA Clean and Concentrator (D4029). We visualized and quantified libraries using a TapeStation, and sequenced with paired-end or single-end reads on a Highseq or Novaseq as indicated (S1 and S7 Tables).

ATAC-seq read alignment and peak calling:

For METSIM samples, ATAC-seq read lengths ranged from 50-150 bp, depending on sequencing center, so all libraries were trimmed to a uniform length of 50 bp before processing. Three METSIM ATAC-seq libraries were single-end and were processed with a single-end version of the following pipeline. All other libraries were paired-end. SGBS ATAC-seq reads were not length-trimmed before processing, although some libraries had 50bp reads and others had 150bp reads. We trimmed sequencing adapters and low quality base calls from the 3' ends of reads using cutadapt[123] with parameters -q 20 –minimum-length 36. We aligned trimmed reads to the hg19 human genome[124] using bowtie2[125] with parameters –minins 36 –maxins 1000 –no-mixed –no-discordant –no-unal and selected nuclear chromosomal alignments with mapq>20 using samtools[125]. We removed alignments overlapping high-signal regions (Duke excluded and ENCODE/DAC exclusion list regions)[126] using BEDTools pairToBed[127] with the parameter -type notospan. We removed duplicate alignments using Picard MarkDuplicates (https://github.com/broadinstitute/picard) and generated ATAC-seq quality metrics using ataqv.[128] Ataqv is only designed for paired-end reads, so we used a customized approach to calculate TSS enrichment for the single-end METSIM libraries. To calculate TSS enrichment, we generated 2,001-bp windows containing the TSS and 1 kb flanking regions on either end for the set of 5,307 RefSeq housekeeping TSSs used by ataqv for TSS enrichment. We then calculated the number of ATAC-seq reads overlapping each base within these 2,001 bp windows for each METSIM sample using BEDTools coverage with the -d option and made a matrix of coverage for these windows using python. Finally, we summed the coverage across each TSS window within the same sample and calculated TSS enrichment by dividing the summed coverage at the TSS by the mean summed coverage of the 100 bases at the leftmost and rightmost ends of the windows using R.

Prior to peak calling, we trimmed alignments so their 5' ends corresponded to the Tn5 binding site (+4 for + strand alignments and -5 for – strand alignments)[69] and smoothed signal by extending alignments 100 bp on either side of the Tn5 binding sites using BEDTools slop[127]. We called peaks (FDR<5%) with MACS2[129] with parameters -q 0.05 –nomodel –bdg and generated ATAC-seq signal bigwig files from MACS2 bedGraph files using the bedGraphToBigWig tool from ucsctools[130]. For SGBS libraries, we proceeded with analyses on a final set of libraries that met our signal-to-noise quality thresholds with a fraction of reads in peaks (FRiP) greater than 20% and a transcription start site enrichment greater than 5[76]. For METSIM libraries, we selected libraries that had TSS enrichment >= 4 calculated from our customized script that works on single-end and paired-end reads. TSS enrichment values produced by our script are generally higher than those calculated by ataqv, and TSS enrichment of 4 from our script corresponds roughly to TSS enrichment of 3 from ataqv.

For each analyzed day of SGBS differentiation, we generated a set of consensus ATAC-seq peaks using the following method. First, we merged peak genomic coordinates across replicates for a given day using BEDTools merge[127]. Second, we defined consensus peaks as merged peaks that overlapped individual replicate peaks in greater than 50% of replicates (at least 3 out of 5 replicates for D14 and 6 out of 10 replicates for D0 and D4).

Identification of differentially accessible peaks:

We generated a set of merged peaks to test for differential chromatin accessibility by merging the top 100,000 consensus peaks in each day (ranked by median peak p-value across replicates). We quantified the accessibility of these merged peaks in each library using featureCounts[131]. We computed the GC percent of each peak using BEDTools nuc[127] and generated within-library GC bias normalization factors using full quantile normalization with EDASeq[132]. We then used EDASeq GC bias normalization factors within DESeq2[133] and used DESeq2 size factors to control for differences in sequencing depth between libraries. We tested for differential chromatin accessibility using DESeq2[133] and classified peaks with FDR<5% and log fold change (LFC)>1 as significantly differential.

Enrichment of transcription factor motifs in differential peaks:

We tested for enrichment of 319 transcription factor (TF) motifs in adipocyte or preadipocyte-dependent peaks using the findMotifsGenome tool from HOMER[134] with the -size 200 option. We used peaks that were not differential in any pairwise day comparison (FDR>50%, absolute value of LFC<1) as background in the enrichment analyses. We classified motifs with a p-value less than the Bonferroni-corrected threshold of 1.6x10-4 (0.05/319 motifs) as significant.

Gene ontology enrichment of genes near differential peaks:

We tested if genes near adipocyte and preadipocyte-dependent peaks were enriched for specific biological processes using the Genomic Regions Enrichment of Annotations Tool (GREAT) web tool (http://great.stanford.edu/public/html/)[135] with the GO Biological Process ontology[135,136]. We ran GREAT version 4.0.4 with the default parameters of basal plus extension, proximal 5 kb upstream to 1 kb downstream, distal 1000 kb (1 Mb), and a whole genome background. We classified ontology terms with Minimum Region-based Fold Enrichment>=2 and FDR<5% as significantly enriched.

Identification of adipose tissue consensus peaks

We constructed an initial set of adipose tissue consensus peaks using the 17 METSIM libraries with TSS enrichment>=4 (our customized TSS enrichment script). To construct consensus peaks, we took the union of peaks across all 17 samples and selected union peaks that overlapped (shared at least one base) with a peak in 3 samples. To identify outlier samples, we computed PCA of ATAC-seq read counts within consensus peaks and performed hierarchical clustering of the top 10 PCs (Fig 2.6). We identified 6 outlier samples: four samples were generated with the omni-ATAC-seq protocol[88] (whereas all other samples were generated using the original ATAC-seq protocol[69]), one sample had a much higher percentage of mitochondrial reads compared to other samples, and one sample had substantially fewer peaks compared to other samples. Adipose tissue peaks from the 11-sample peak set showed stronger overlap with Roadmap Epigenomics adipose nuclei enhancers (Figs 2.1D and 2.6D) and stronger enrichment for all tested traits except BMI (Figs 2.3 and 2.6E) compared to the 17-sample set. Therefore, we removed these 6 samples and generated consensus peaks with 11 samples, using the same approach as for 17 samples.

RNA-seq library preparation, read alignment, and identification of differentially expressed genes:

We isolated total RNA from SGBS cells at D0, D4, and D14 of differentiation using the Total RNA Purification Kit (product #17200) from Norgen Biotek (Ontario, Canada). Novogene (Beijing, China) generated poly-A RNA libraries and performed paired-end RNA sequencing (RNA-seq, read length 150 bp) using a NovaSeq 6000 (Illumina, California, USA). We trimmed sequencing adapters and low quality base calls from the 3' ends of RNA-seq reads using cutadapt[123] with parameters -q 20 –minimum-length 36. We aligned reads to the hg19 human genome[124] using STAR[137] with parameters --sjdbOverhang 149 --twopassMode Basic --quantMode TranscriptomeSAM --outFilterMultimapNmax 20 --alignSJoverhangMin 8 --alignSJDBoverhangMin 1 --outFilterMismatchNmax 999 --outFilterMismatchNoverReadLmax 0.04 --alignIntronMin 20 --alignIntronMax 1000000 --alignMatesGapMax 1000000. We quantified expression of genes from GENCODE v29 lift37[138] and corrected for GC bias using salmon44 with parameters –seqBias –gcBias –gencode. We generated RNA-seq quality metrics using the CollectRnaSeqMetrics tool from Picard (https://github.com/broadinstitute/picard). We used PCA to determine which replicates clustered. Within timepoint clusters, we observed additional clustering by batch that we corrected for in downstream analysis.

To identify differentially expressed genes, we imported salmon transcript quantifications and collapsed to the gene level using tximport[139]. We retained 18,299 genes with median DESeq2-normalized count >= 1 across all libraries. We tested for differential gene expression using DESeq2[133] and classified genes with FDR<5% and LFC>1 as significantly different across pairs of timepoints.

Gene ontology enrichment of differential genes:

We tested if differentially expressed genes were enriched for specific biological processes using the PANTHER statistical overrepresentation test[140] with the GO-Slim Biological Process ontology[136,141]. We ran PANTHER using Fisher's exact test for calculating enrichment and used all 18,299 genes examined in the differential expression analysis as background for the enrichment tests. We classified ontology terms with fold enrichment>=2 and FDR<5% as significantly enriched.

Identification of genes linked to context-dependent peaks:

Hi-C: We identified context-dependent peaks that intersect (overlap>=1 base pair) with the "other-end" fragments of "bait-other" Hi-C loops and either end of "bait-bait" loops from previously published adipocyte promoter capture Hi-C data[97,98] using BedTools[127]. We linked peaks to genes that were on the opposite end of the Hi-C "bait-bait" loops. We categorized Hi-C interaction types as “bait-bait” if the “other-end” fragment also covered a bait fragment and “bait-other” if the “other-end” fragment did not cover a bait fragment.

eQTL: We identified context-dependent peaks that overlapped eQTL proxy variants ($r2>0.8$ with the eQTL lead, 1000G phase 3 EUR LD calculated using PLINK v1.9[142]) using previously published primary and conditional eQTL mapped in METSIM adipose tissue[61,65] using BedTools[127]. We identified the best eQTL proxy within the peak as the variant with the strongest LD with the lead variant at the signal. If a peak contained proxy variants from both primary and conditional signals with equally strong LD, we selected the primary signal proxy as the best proxy. We also listed all eQTL variants that intersected a peak.

Differential Expression: To provide additional evidence for peak-gene links identified by Hi-C or eQTL, we identified if the linked gene was also differentially expressed (FDR<5% and LFC>1) between any timepoint comparisons. We investigated linking context-dependent peaks to differentially expressed genes based on proximity between the peak and gene TSS, but proximity is indirect and based on the even distribution of peaks from TSS as distance increased, any threshold would have been arbitrary so we concluded that proximity alone was not strong evidence to link a peak and gene (S9 Fig).

SGBS genotyping and imputation:

We genotyped two SGBS DNA samples with 335 samples from a separate study using the Infinium Multi-Ethnic Global array (Illumina, San Diego, CA, USA), which contains over 1.7 million variants. The additional 335 samples were used to calculate genotyping call rates, but all subsequent analyses were performed using only SGBS genotypes. We removed variants with call rate <95%, performed multiple quality checks with the checkVCF.py tool (https://genome.sph.umich.edu/wiki/CheckVCF.py), and oriented alleles relative to the hg19 reference genome[124] using PLINK.[142] We restricted to variants that had the same genotype call in both SGBS samples for downstream analyses. We phased autosomal variants using Eagle v2.4[143] and imputed missing variants using Minimac3[144] with the 1000 Genomes (1000G) phase 3 reference panel[145]. The imputation r2 statistic used to assess imputation quality is not meaningful when imputation is performed on a single sample. Therefore, we retained variants with genotype probability (GP) > 0.9. In our batch of SGBS cells, a subset of cells showed loss of heterozygosity on regions of chromosomes 7 and 10 (chr7:1-31,000,000 and chr10:131,000,000-135,534,747); variants overlapping these regions were removed prior to downstream analyses.

ATAC-seq allelic imbalance:

To identify heterozygous variants exhibiting allelic imbalance (AI) in SGBS ATAC-seq reads, we first removed reads exhibiting allelic mapping bias and duplicated reads using WASP[146]. We counted reads aligning to each allele of biallelic heterozygous single nucleotide variants using ASEReadCounter[147] with the option –min-base-quality 30 and removed variants that had aligned bases other than the two genotyped alleles. For each SGBS differentiation day, we selected a set of variants to test for AI that had at least 20 total reads combined across both alleles and at least 3 reads on each allele in greater than 50% of replicates for the given day (3 replicates for D14 and 6 replicates for D0 and D4). We tested for AI separately by day using DESeq2[133] with the design formula ~0+sample+allele, where 'sample' represents an individual ATAC-seq replicate. Using DESeq2, we tested if the ratio of alternate allele counts to reference allele counts was greater than $\log_2(55/45)$ using a Wald test, estimated dispersions of allelic ratios using maximum likelihood, and adjusted for multiple testing using the BH procedure. We used an LFC threshold of $\log_2(55/45)$ rather than $\log_2(50/50)$, to preferentially select variants showing strong AI, especially given high variability in allelic ratios. We considered variants with FDR<5% to show significant AI.

Overlap of GWAS signals with context-dependent peaks:

We downloaded the NHGRI-EBI GWAS catalog[44] on January 17, 2020 and lifted variant positions from hg38 to hg19 using pyliftover (https://github.com/konstantint/pyliftover), a python implementation of the UCSC liftOver tool[148]; We rescued a subset of variants that did not successfully lift over using variant rsIDs to convert between hg38 and hg19 coordinates. We restricted to significant associations (p<5x10-8) for single nucleotide variants (haplotype associations and variant-variant interactions were removed) that were biallelic in the dbSNP[149] build 151 common variant set. To generate a set of LD-distinct association signals, we performed LD-clumping using swiss (https://github.com/statgen/swiss) in a trait-agnostic manner[61]; the most significant p-value per variant was selected, regardless of trait, and variants within strong LD (r2>0.8, 1000G phase 3 EURs) and within 1 Mb of another variant with a more significant p-value (not necessarily for the same trait) were removed. However, we retained all variants and associated traits at each signal for reference in supplemental tables.

To map GWAS catalog trait terms to standardized ontology terms, we downloaded the GWAS to Experimental Factor Ontology (EFO) mappings file from the GWAS catalog on May 13, 2021 and extracted the EFO term corresponding to each trait. We identified GWAS signals that had at least one proxy variant (LD r2>0.8 with the signal lead variant, 1000G phase 3 EURs, calculated with PLINK v1.9[142]) found within context-dependent peaks using BEDTools[127]. For each specific EFO term, we counted the number of signals containing that EFO term, including all variant-trait associations at a signal, not just the strongest association; we only counted each term once per signal. We performed this counting procedure for both the entire LD-clumped GWAS catalog and the subset of the catalog that overlapped the ATAC-seq peak set of interest. Because our goal in using EFO terms is to reduce the complexity of the GWAS catalog traits, we removed any GWAS traits that mapped to 5 or more EFO terms for our analyses that count EFO terms, which only removed <1% of GWAS traits. However, we retained all GWAS traits and EFO terms in S15 Table for reference. To normalize for the overall frequency of an EFO term in the clumped catalog, we divided the number of ATAC-seq counts by the number of total counts for each EFO term and multiplied by 100 to express as a percentage. When ranking by normalized ATAC-seq count to get the top 10 EFO terms for preadipocyte-dependent and adipocyte-dependent peaks, we restricted to terms that had total count >=100.

Enrichment of heritability in ATAC-seq peaks:

We used stratified LD score regression as implemented in LDSC v1.0.1[101] to test if ATAC-seq peaks were enriched for heritability of 9 GWAS traits: 8 cardiometabolic traits BMI[150], HDL cholesterol[151], LDL cholesterol[151], triglycerides[151], total cholesterol[151], coronary artery disease[152], WHRadjBMI[150], T2D[45], and rheumatoid arthritis[153] as a negative control. We tested for heritability enrichment separately in 7 different ATAC-seq peak sets: preadipocyte-dependent peaks, adipocyte-dependent peaks, the top 100,000 consensus peaks for SGBS D0, D4, and D14, and consensus peaks mapped in 17 adipose tissue samples and 11 adipose tissue samples. Using LDSC, we calculated LD scores for ATAC-seq peaks using HapMap3 SNPs[154] and LD calculated from 1000G phase 3 EURs[145]. We computed partitioned heritability separately for each ATAC-seq peak set using LDSC correcting for the baseline v1.2 model, which consists of 52 genic and functional annotations[101]. We used the regression coefficient z-score reported by LDSC to assess the importance of each ATAC-seq peak set for each trait relative to the baseline model, where a positive z-score means that SNP heritability is increased in a given ATAC-seq peak set relative to the baseline model and a negative z-score means that heritability is decreased in the peak set relative to the baseline[155]. We calculated p-values by testing if the coefficient z-score was greater than 0, assuming a standard normal distribution. We classify results with a p-value threshold of 0.05 as nominally significant and 0.0056 (0.05/9 traits) as significant. We compare the relative importance of each ATAC-seq peak set to heritability for a given trait by comparing coefficient z-scores.

Prioritization of candidate regulatory elements for functional testing:

We identified context-dependent peaks linked to a candidate gene by two or more of our three methods to predict target genes. We identified a further subset of these context-dependent peaks that overlapped a cardiometabolic GWAS signal and an adipose peak. We used further lines of evidence to prioritize these candidate regulatory elements for functional testing including: location of variants closer to the summit of a peak as opposed to the shoulder and literature review of linked gene's relevance to adipose biology.

Transcriptional reporter luciferase assays:

SGBS preadipocytes and adipocytes were maintained and transcriptional reporter luciferase assays were performed as previously described[122] with the following changes. Primers were designed to amplify the entire chromatin accessibility region containing variants of interest. Amplified regions containing variant reference and alternate alleles were cloned individually into the XbaI-SbfI restriction sites of the pLS-mP-Luc lentiviral luciferase vectors (a gift from Nadav Ahituv, Addgene plasmid # 106253) or pGL4.23 firefly luciferase reporter vector (Promega) in ‘forward’ and ‘reverse’ orientations (named with respect to the genome reference). The variants were cloned upstream of the minimal promoter and verified by Sanger DNA Sequencing. For lentivirus production, HEK293T cells were grown to 70-80% confluency in 100 mm plates and co-transfected with 9.5 μg of a pLS-MP-Luc construct, 8 μg of packaging plasmid (psPAX2, a gift from Didier Trono, Addgene plasmid # 12260), and 2.5 μg of an envelope plasmid (pMD2.G, a gift from Didier Trono, Addgene plasmid # 12259) using Lipofectamine 2000 transfection reagent (Invitrogen). Media was replaced with fresh growth media 18 hours after transfection. Viral supernatant was harvested 48 and 72 hours after transfection and concentrated using 4X Lenti-X concentrator (Clontech). Lentiviral titer was measured using Lenti-X qRT-PCR Titration Kit (Takara Bio), and functional titers were represented as transduction units. For data normalization, empty pLS-MP-Luc and Renilla luciferase vector pLS-SV40-mp-Rluc viruses (a gift from Nadav Ahituv, Addgene plasmid # 106292) were prepared and quantified in a similar manner.

For preadipocytes, 25,000 SGBS cells were plated the day before transduction, and 35,000 SGBS cells were plated and differentiated for adipocytes into 24 well plates and spin-infected with appropriate titer of construct and Renilla virus in the presence of 10 ug/ml polybrene media. For viral based transcriptional luciferase assays, two independent construct viruses were used for each allele in each orientation and were transduced in tetraplicate wells. After 8 hrs of transduction, media was replaced with fresh growth media, and luciferase and Renilla activity was measured 48 - 72 hours post transduction using Dual Luciferase Reporter Assay System (Promega). For plasmid based transcriptional luciferase assays, we used primers to amplify the regions of interest and we cloned the constructs containing the variants into pGL4.23 firefly luciferase reporter vector (Promega). Five independent clones for each allele in each orientation were cotransfected with Renilla luciferase vector in triplicate wells using Lipofectamine 3000 (Lifetechnologies). Luciferase and Renilla activity were measured after 28hrs of transfection.

For both viral- and plasmid-based assays, luciferase activity of experimental clones was normalized to Renilla luciferase as well as empty vector activity to control for differences in transfection efficiency. All transcriptional reporter assays were repeated on different days. Data are reported as fold change in activity relative to an empty vector. We used a Student's t-test to compare luciferase activity between alleles and between contexts.

Acknowledgements

Thank you to everyone who contributed to the work in this chapter, including Maren Canon and Alexias Safi for guidance on optimizing the ATAC-seq protocol, Kenneth Ng for assistance with ATAC-seq library preparation, Martin Wabitsch for providing SGBS cells, Markku Laakso for providing METSIM adipose samples, Swarooparani Vadlamudi and Gautam Pandey for performing functional assays, Kristina Garske for adipocyte Hi-C data, Roadmap Epigenomics Consortium for data, Kevin Currin for collaborating on study design, data analysis, and interpretation, Michael Love for advising on data analyses, and Karen Mohlke for advising on study design, analyses, and interpretation.

Figures

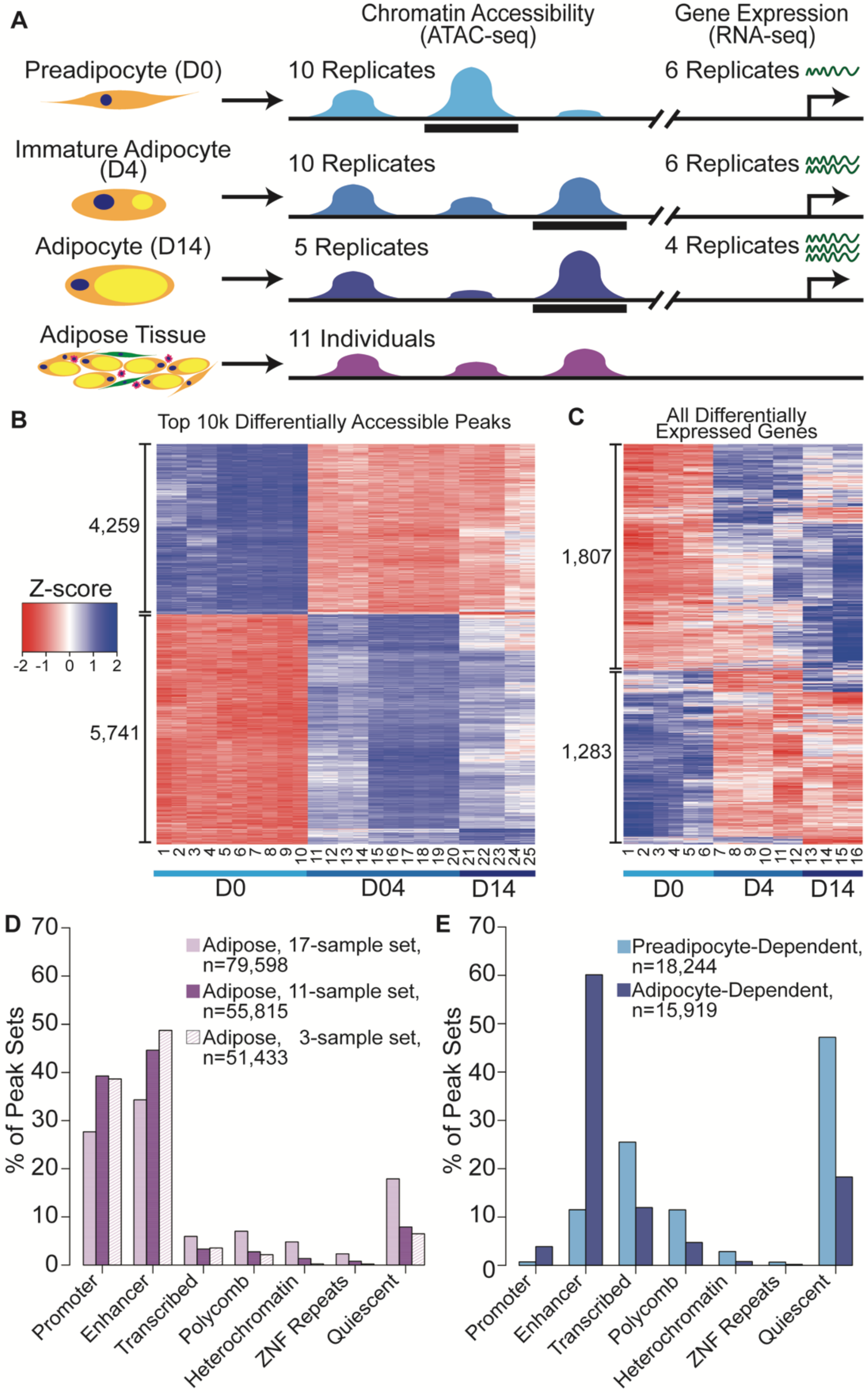
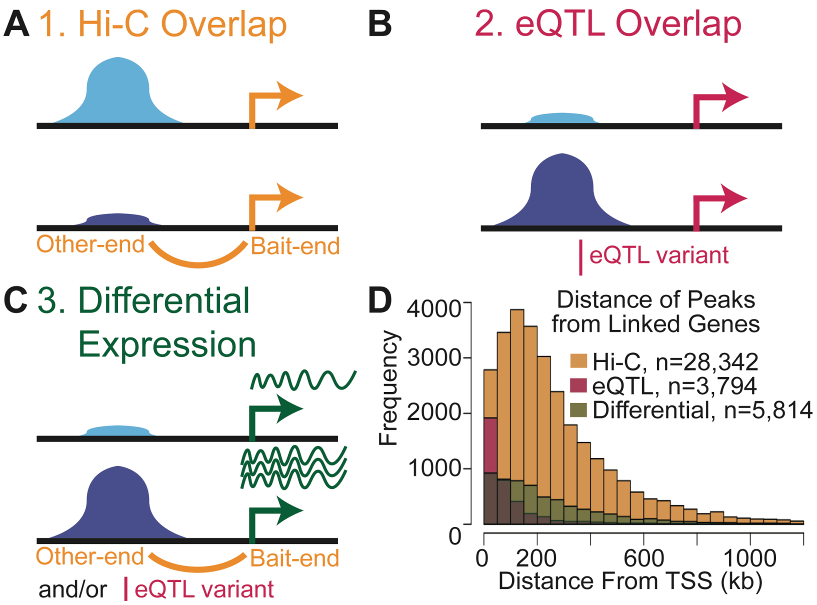

A
Chromatin Accessibility
(ATAC-seq)
Gene Expression
(RNA-seq)
Preadipocyte (D0)
10 Replicates
6 Replicates
Immature Adipocyte
(D4)
10 Replicates
6 Replicates
Adipocyte (D14)
5 Replicates
4 Replicates
Adipose Tissue
11 Individuals
B
Top 10k Differentially Accessible Peaks
Z-score
-2 -1 0 1 2
4,259
5,741
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
D0
D04
D14
C
All Differentially
Expressed Genes
1,807
1,283
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
D0
D4
D14
D
% of Peak Sets
0 10 20 30 40 50 60 70
Adipose, 17-sample set, n=79,598
Adipose, 11-sample set, n=55,815
Adipose, 3-sample set, n=51,433
Promoter
Enhancer
Transcribed
Polycomb
Heterochromatin
ZNF Repeats
Quiescent
E
% of Peak Sets
0 10 20 30 40 50 60 70
Preadipocyte-Dependent, n=18,244
Adipocyte-Dependent, n=15,919
Promoter
Enhancer
Transcribed
Polycomb
Heterochromatin
ZNF Repeats
Quiescent

Figure 2.1: Genome-wide profiles of chromatin accessibility and gene expression at three timepoints of adipocyte differentiation.

(A) Schematic of experimental design. SGBS cells were harvested as preadipocytes (D0), immature adipocytes (D4), and adipocytes (D14). Chromatin accessibility (blue) and gene expression (green) profiles were generated on replicates from each timepoint. Context-dependent peaks are shown as black bars. Chromatin accessibility profiles also were generated from subcutaneous adipose tissue (purple) of 17 individuals and an optimized consensus CA map was developed from a subset of 11 individuals. (B) Heatmap of the top 10,000 context-dependent peaks (from S4 Table) colored by z-score. (C) Heatmap of expression level of all 3,090 context-dependent genes (from S9 Table) colored by z-score. Library numbers correspond to quality metrics in S8 Table. (D-E) Values in S9 Table. (D) Adipose peak overlap with chromatin states of Roadmap Epigenomics Project adipose nuclei for three sets of adipose consensus peaks. (E) Preadipocyte- and adipocyte-dependent peak overlap with chromatin states of Roadmap adipose nuclei.



E

Method	Peaks	Genes	Peak-Gene Pairs
1	14,080	9,080	28,996
2	3,002	2,369	3,794
3	5,077	1,552	5,897
Any 2+	5,145	1,670	6,050

Figure 2.2: Linking context-dependent chromatin accessibility to candidate genes.

(A-C) Schematic of three approaches to link peaks to genes. Day 0 (light blue) and day 14 (dark blue) context-dependent peaks are represented. (A) Context-dependent peaks that overlap elements connected to gene promoters using adipocyte promoter capture Hi-C (orange). (B) Context-dependent peaks that overlap adipose gene eQTL variants (r2>.8 with lead, red). (C) Context-dependent peaks linked to a gene through Hi-C or eQTL for which the linked gene was also differentially expressed between any timepoints (green). (D) Histogram of distances from edges of peaks to the transcription start site of a linked gene within 1.2 Mb. Values in S12 Table. (E) Numbers of context-dependent peaks linked to genes by each method and by two or more methods. Values summarize full results in S12 Table.

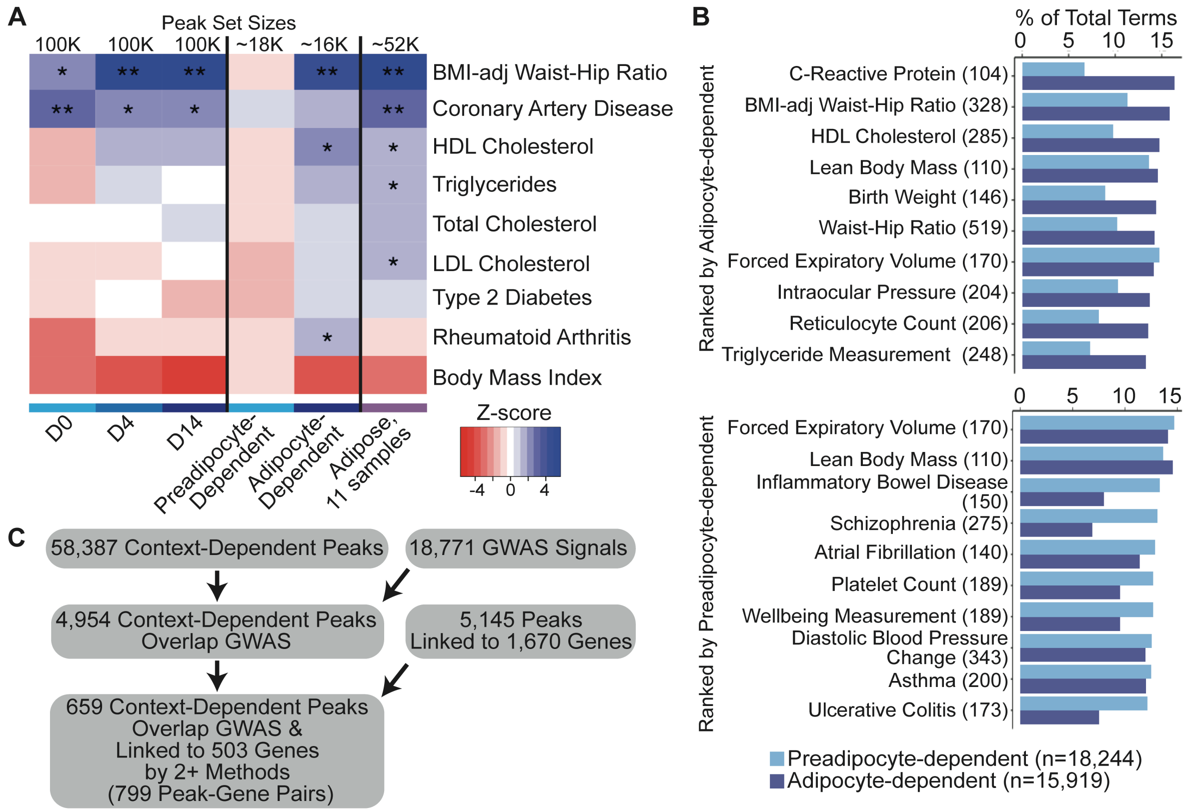


Figure 2.3: Linking peaks to GWAS signals.

(A) Heatmap of cardiometabolic trait GWAS locus enrichment; rheumatoid arthritis was selected for comparison. Peak sets include 100,000 peaks from individual days, preadipocyte- and adipocyte-dependent peaks derived from pairs of timepoints, and adipose tissue peaks. Values in S13 Table. **, P < 0.0056; *, P < 0.05 (B) Barplots of normalized counts of specific experimental factor ontology (EFO) terms for GWAS signals with a variant in a context-dependent peak. Barplots show the top ten EFO terms ranked by normalized count for either preadipocyte-dependent peaks, or adipocyte-dependent peaks. Total number of signals for each term used in the overlap is noted in parentheses in the axis label. Total number of signals for each term overlapping a context-dependent peak is noted to the right of the "All Context-dependent" bar. Values in S14 Table. (C) Flowchart identifying context-dependent peaks overlapping GWAS signals and linked to genes through 2 or more methods.

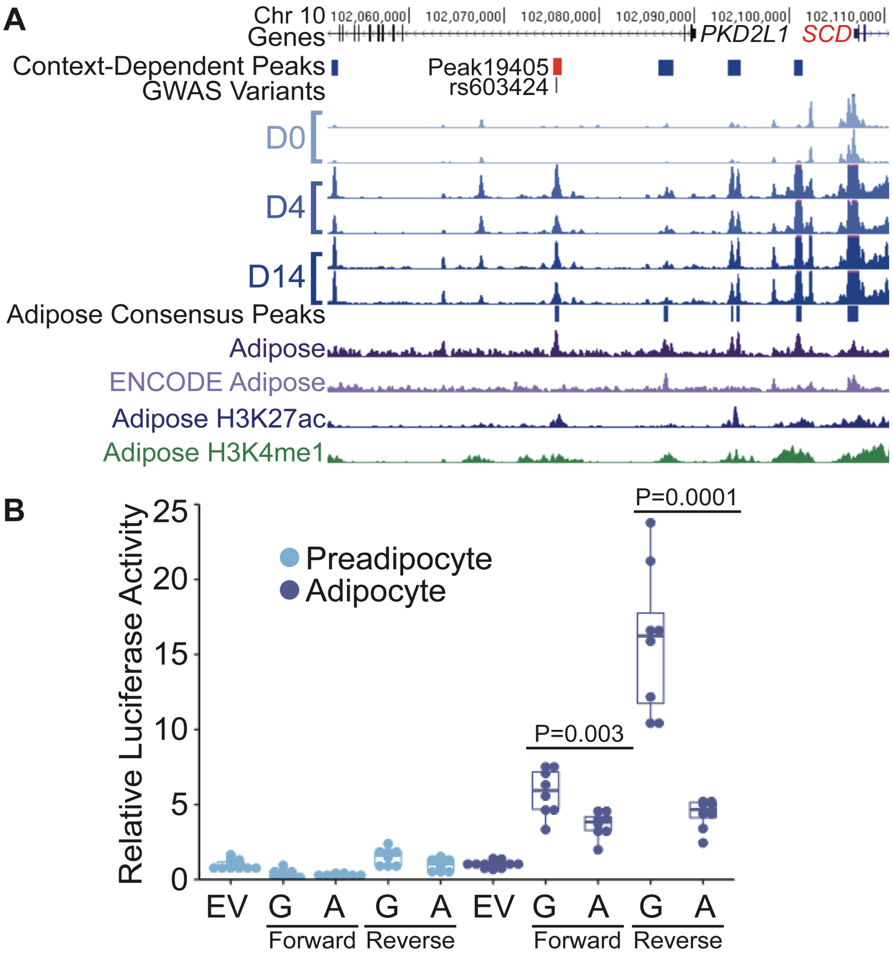


Figure 2.4: Allelic differences in transcriptional activity for a context-dependent regulatory variant in a context-dependent element at the *SCD* locus.

(A) Peak19405 (red) is more accessible in D4 and D14 adipocytes than D0 preadipocytes, overlaps an adipose tissue consensus peak (dark purple), and overlaps variant rs603424, which is associated with blood plasma levels of palmitoleic acid and adipose *SCD* expression. *SCD* is also more highly expressed at D4 and D14 compared to D0. Additional tracks show adipose tissue ATAC-seq from ENCODE (light purple) and adipose nuclei histone mark ChIP-seq from the Roadmap Epigenomics project (blue and green). (B) A 592-bp genomic region surrounding peak19405 containing the rs603424-G allele shows increased transcriptional activity compared to the rs603424-A allele in the forward and reverse orientations only in adipocytes (tested at day

12), the context in which chromatin was more accessible compared to preadipocytes. Dots represent two independent constructs assayed from four replicates each. Luciferase activity was normalized relative to an empty vector (EV). Values in S18 Table.

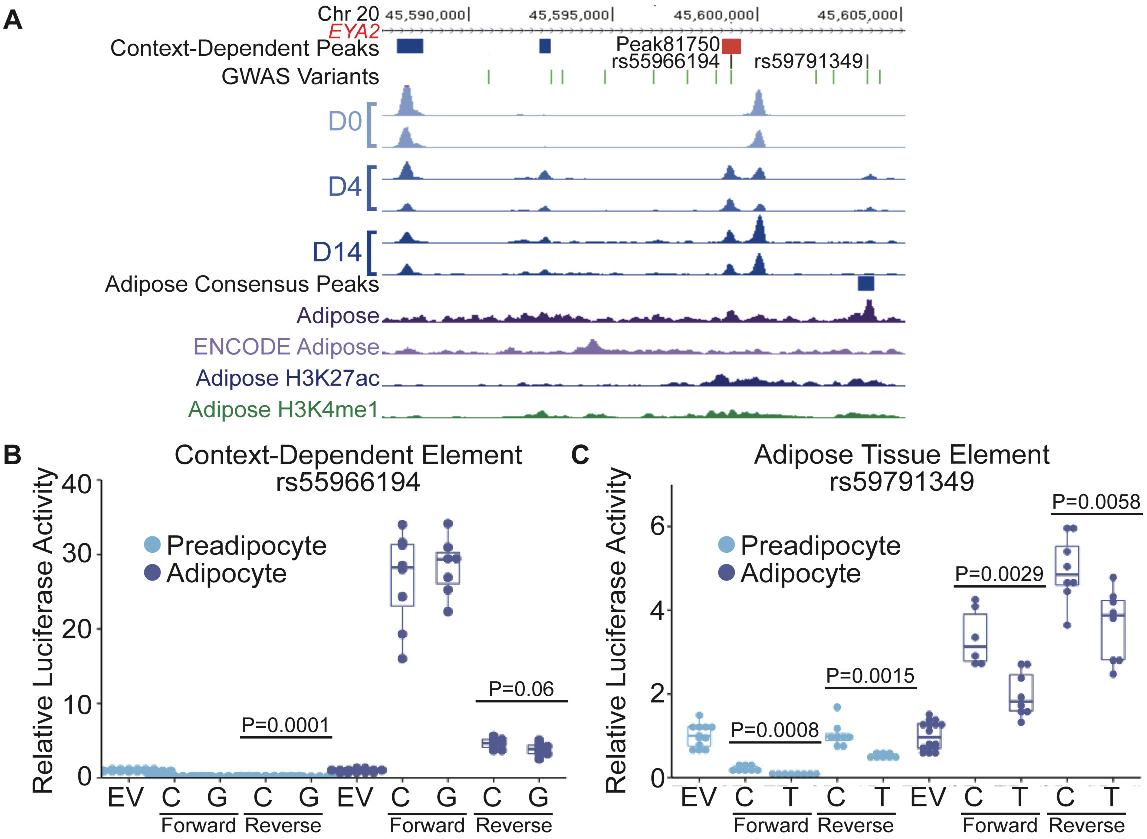


Figure 2.5: Allelic differences in transcriptional activity for variants in two regulatory elements at the *EYA2* locus.

(A) Peak81750 (red) is more accessible in D4 and D14 adipocytes and overlaps variant rs559066194, which is associated with increased risk of type 2 diabetes and increased *EYA2* expression. *EYA2* is more highly expressed at D4 and D14, compared to D0. A second variant at this locus, rs59791349, intersects a consensus adipose peak (dark purple) but not a context-dependent peak. Additional tracks as in Fig 4. (B-C) Values in S18 Table. (B) A 419-bp genomic region surrounding peak81750 containing the rs555966194-C allele shows modestly-increased transcriptional activity compared to the rs555966194-G allele in the reverse orientation, but not the forward, in adipocytes (tested at day 9), the context in which chromatin was more accessible compared to preadipocytes. Dots represent two independent constructs assayed from four replicates each. Luciferase activity was normalized relative to an empty vector (EV). (C) A 288-

bp genomic region containing the rs59791349-C allele shows increased transcriptional activity compared to the rs59791349-T allele in both orientations and in both preadipocytes and adipocytes (tested at day 9). Dots represent two independent constructs assayed from four replicates each. Luciferase activity was normalized relative to an EV.

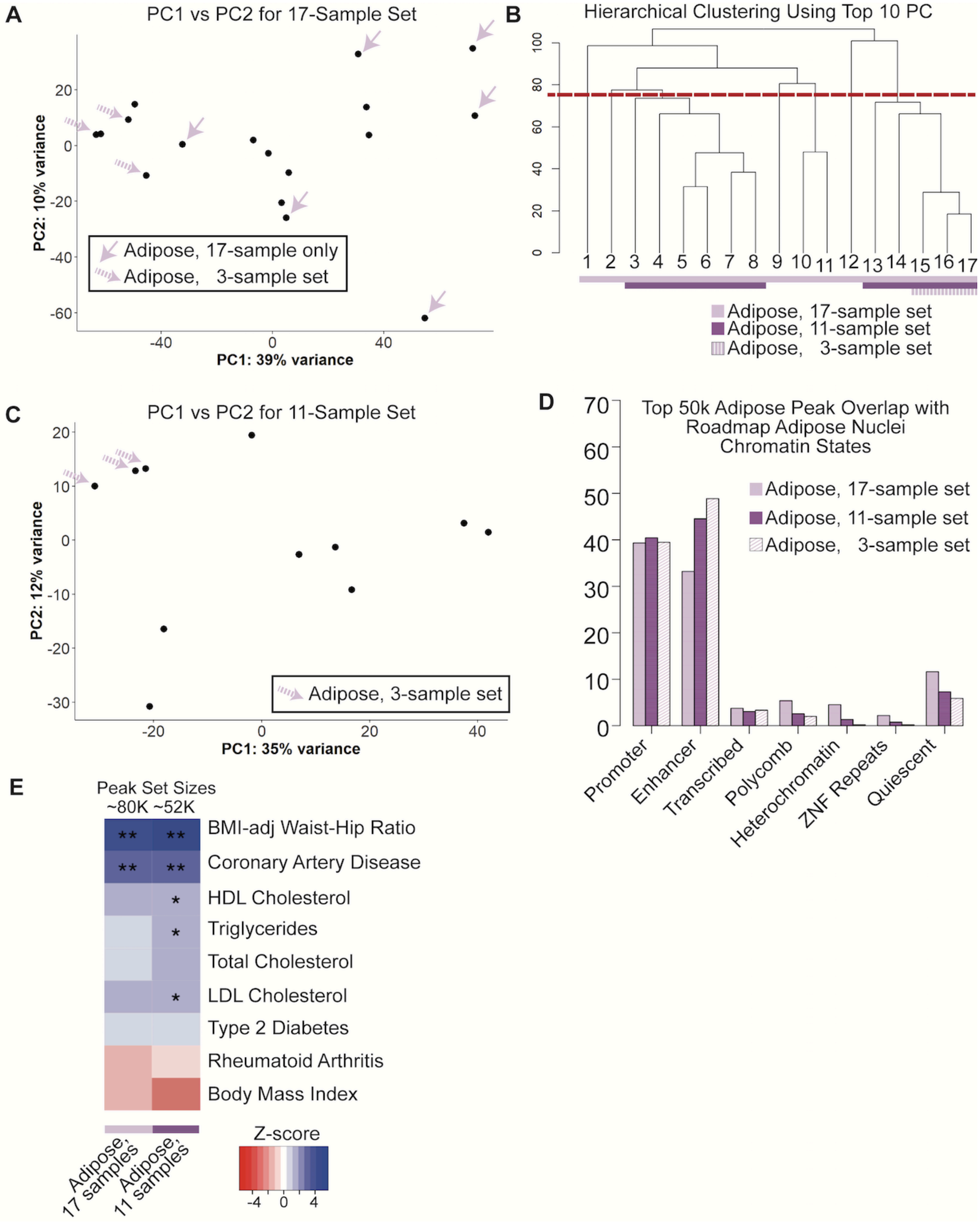

A
PC1 vs PC2 for 17-Sample Set
PC2: 10% variance
PC1: 39% variance
20
0
-20
-40
-60
-40
0
40
Adipose, 17-sample only
Adipose, 3-sample set
B
Hierarchical Clustering Using Top 10 PC
100
80
60
40
20
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
Adipose, 17-sample set
Adipose, 11-sample set
Adipose, 3-sample set
C
PC1 vs PC2 for 11-Sample Set
PC2: 12% variance
PC1: 35% variance
20
10
0
-10
-20
-30
-20
0
20
40
Adipose, 3-sample set
D
Top 50k Adipose Peak Overlap with Roadmap Adipose Nuclei Chromatin States
70
60
50
40
30
20
10
0
Adipose, 17-sample set
Adipose, 11-sample set
Adipose, 3-sample set
Promoter
Enhancer
Transcribed
Polycomb
Heterochromatin
ZNF Repeats
Quiescent
E
Peak Set Sizes
~80K ~52K
** **
** **
*
*
*
BMI-adj Waist-Hip Ratio
Coronary Artery Disease
HDL Cholesterol
Triglycerides
Total Cholesterol
LDL Cholesterol
Type 2 Diabetes
Rheumatoid Arthritis
Body Mass Index
Adipose, 17 samples
Adipose, 11 samples
Z-score
-4
0
4

Figure 2.6: Comparison of adipose tissue for three subsets of samples. (A) PCA for PC1 (principal component) vs PC2 for all 17 samples that met quality thresholds. Solid light purple arrows indicate samples that are unique to the 17-sample set (excluded from the 11-sample set). Dashed light purple arrows indicate three previously published samples that have been included in the 11- and 17- sample sets. (B) Hierarchical clustering using the top 10 PCs from PCA. The red dashed line indicates the cutoff used to exclude six samples from the 11-sample set. Dark purple indicates samples in the 11-sample set. Dashed light purple indicates samples in the 3-sample set. Sample numbers correspond to library quality metrics in Table S10. (C) PCA for PC1 vs PC2 for the 11-sample set. Dashed light purple arrows indicate three previously published samples that have been included. (D) Adipose peak overlap with chromatin states of Roadmap Epigenomics adipose nuclei for the three different sample subsets of adipose consensus peaks using the top 50k peaks for each set. (E) Heatmap of cardiometabolic trait GWAS locus enrichment; rheumatoid arthritis was selected for comparison. Peak sets include two sets of adipose tissue peaks. **, P < 0.005; *, P < 0.05.

Tables

Timepoint	Sample ID (identified in Fig 1B)	Batch I D	Final Reads	Number of Peaks	Percent Reads In Peaks	TSS Enrichment
D00	1	B1	57,474,834	147,944	64.27	5.9
D00	2	B1	45,276,054	133,522	51.68	5.9
D00	3	B4	102,126,542	162,246	48.14	5.4
D00	4	B4	97,154,896	191,773	49.03	7.2
D00	5	B2	43,365,932	116,318	38.47	5.8
D00	6	B2	40,894,284	118,621	43.37	6.2
D00	7	B2	33,634,560	155,237	55.52	9.6
D00	8	B2	50,964,904	132,142	48.36	6.5
D00	9	B2	77,285,432	157,596	57.87	5.9
D00	10	B2	97,528,124	150,754	49.25	5.9
D04	11	B3	80,807,596	163,193	50.2	6.7
D04	12	B3	72,362,194	154,669	49.1	5.8
D04	13	B4	112,892,568	172,115	53.33	5.9
D04	14	B4	34,294,124	124,029	48.95	5.9
D04	15	B2	37,724,142	156,758	52.16	11.3
D04	16	B2	67,860,150	141,195	45.16	7.1
D04	17	B2	62,275,188	144,845	47.73	6.8
D04	18	B2	67,595,282	154,912	49.61	6.3
D04	19	B2	79,694,554	165,682	51.96	6.4
D04	20	B2	106,937,716	164,104	49.2	7.4
D14	21	B1	137,878,360	167,758	44.66	8.5
D14	22	B1	152,136,306	170,636	47.64	6.8
D14	23	B1	156,493,884	171,050	51.5	8.4
D14	24	B3	62,563,568	144,757	43.27	6.2
D14	25	B3	45,696,682	106,454	23.08	5.1
D02	n/a	B3	77,897,392	164,274	52.01	6.3
D02	n/a	B3	47,221,196	113,441	26.23	5.5

Table 2-1: ATAC-seq library metrics for SGBS libraries.

ATAC-seq libraries of SGBS preadipocytes (D00), immature adipocytes (D02: not included in final analyses, and D04), and adipocytes (D14) with batch, sequencing, and alignment metrics.

	Total Consensus	
Differentiation Timepoint	**Peaks**	**Genes**
D00	127,297	13,513
D04	137,602	13,598
D14	144,052	13,323

Table 2-2: Summary of chromatin accessibility consensus peaks and genes for SGBS differentiation timepoints.

Consensus peaks were defined as the union of chromatin accessibility region accessible in majority of replicates for timepoint, overlapping by 1 or more base pairs and consensus genes were defined as those expressed in majority of replicates for timepoint (see methods for more information).

Differentiation Timepoints		**Early > Late**		**Early < Late**	
Early	**Late**	**Context-dependent Peaks***	**Context-dependent Genes**	**Context-dependent Peaks***	**Context-dependent Genes**
D00	D04	26,435	1,043	26,218	1,128
D00	D14	21,192	788	18,529	1,319
D04	D14	519	173	599	449
*Of top 100k consensus peaks from each day, merged					

Table 2-3: Summary of context-dependent peaks and genes.

Total context-dependent peaks and genes (DESeq2, LFC>1, FDR<5%) identified for each timepoint comparison.

Sample ID	Sample included in sets	Protocol	Final Filtered Reads	Number of Peaks	Percent Reads In Peaks	TSS Enrichment
1	17-sample	omni	81,206,302	35,912	2.58	4.04
2	17-sample	original	45,260,382	10,488	1.14	4.2
3	17-sample, 11-Sample	original	70,794,776	47,516	5.06	6.7
4	17-sample, 11-Sample	original	196,497,894	38,687	4.41	5.55
5	17-sample, 11-Sample	original	99,661,796	27,089	2.53	4.43
6	17-sample, 11-Sample	original	123,610,224	27,418	1.92	4.07
7	17-sample, 11-Sample	original	168,248,638	51,269	4.71	6.17
8	17-sample, 11-Sample	original	199,017,548	89,593	6.59	5.44
9	17-sample	omni	93,739,658	73,642	6.24	4.99
10	17-sample	omni	135,803,892	120,452	8.59	4.74
11	17-sample	omni	218,598,392	131,192	6.7	4
12	17-sample	original	66,562,322	34,196	6.26	7.14
13	17-sample, 11-Sample	original	50,367,150	35,367	4.55	6.08
14	17-sample, 11-Sample	original	46,965,542	43,063	6.76	7.26
15	17-sample, 11-Sample, 3-sample	original	95,600,037	41,351	4.56	4.93
16	17-sample, 11-Sample, 3-sample	original	90,614,097	58,340	6.88	5.54
17	17-sample, 11-Sample, 3-sample	original	104,977,421	65,312	7.38	5.23

Table 2-4: ATAC-seq library metrics for adipose tissue libraries.

Adipose tissue ATAC-seq library metrics with batch, sequencing and alignment metrics.

Differentiation Timepoint	Sample ID (as identified in Fig 1C)	Batch ID	Total Raw Reads	Reads Remaining After Adapter Trim
D0	1	B1	69,092,222	69,049,144
D0	2	B1	65,988,010	65,941,792
D0	3	B2	58,948,614	58,906,356
D0	4	B2	60,061,888	60,009,960
D0	5	B3	57,513,988	57,450,212
D0	6	B3	50,184,610	50,139,772
D4	7	B1	53,157,272	53,125,366
D4	8	B1	58,649,698	58,612,282
D4	9	B2	65,518,564	65,474,032
D4	10	B2	61,406,176	61,366,088
D4	11	B3	49,456,572	49,399,692
D4	12	B3	58,091,326	58,034,570
D14	13	B1	51,205,096	51,174,272
D14	14	B1	52,623,944	52,584,402
D14	15	B3	56,558,106	56,514,086
D14	16	B3	53,964,460	53,904,422
D2	n/a	B1	50,444,676	50,407,890
D2	n/a	B1	60,688,848	60,647,172

Table 2-5: RNA-seq library metrics for SGBS libraries.

RNA-seq libraries of SGBS preadipocytes (D00), immature adipocytes (D02: not included in final analyses, and D04), and adipocytes (D14) with batch, sequencing, and alignment metrics.

CHAPTER 3: CONTEXT-DEPENDENT CHROMATIN ACCESSIBILITY IN ADIPOCYTES UNDER DISEASE-RELEVANT CONDITIONS OF FREE FATTY ACIDS, HYPOXIA, AND INFLAMMATION

Introduction

Genome-wide association studies have identified thousands of loci associated with cardiometabolic traits[43,44]; however, the mechanisms of most loci remain unclear[53–55]. Factors such as colocalization of eQTL variants with GWAS loci suggest a regulatory mechanism at many of these noncoding loci[61–67]. Regulatory mechanisms can be cell type- and context-dependent[53], therefore, testing in disease relevant cell types and contexts can aid identification of mechanisms.

Adipose tissue is relevant to cardiometabolic traits through its roles in lipid storage[9,10]. Adipose tissue is heterogenous and composed of many cell types including preadipocytes, adipocytes, macrophages, and endothelial cells, among others[11]. Adipocytes are an important cell type within adipose tissue that are responsible for storing lipids[11]. During periods of excess nutrition, adipocytes store lipids through two primary pathways; hyperplasia, during which preadipocytes differentiate into mature adipocytes to store excess energy, or hypertrophy, during which existing adipocytes expand to store excess energy[12]. Adipocyte hypertrophy can be modeled by exposing adipocytes to stimuli such as excess free fatty acids which results in excess lipid accumulation within adipocytes[156]. Saturated free fatty acids such as palmitic acid and monounsaturated free fatty acids such as oleic acids have been shown to activate different transcriptional networks in mouse 3T3-L1 adipocyte cells[157]. Therefore, studying the contexts of hypertrophy during exposure to saturated free fatty acids or monounsaturated free fatty acids could identify different regulatory elements. Enlarged adipocytes experience hypoxia and inflammation, which are markers of dysfunctional adipocytes and metabolic disease[12]. Enlarged adipocytes have been shown to experience dysfunction such as insulin resistance independent of markers of inflammation in mouse 3T3-L1 adipocyte cells[156]. Therefore, we investigated regulatory mechanisms in the presence of excess free fatty acids in the Simpson-Golabi-Behmel Syndrome (SGBS) human adipocyte model[21,22]. SGBS cells are a human diploid preadipocyte cell model that can be differentiated into adipocytes to study adipocytes in disease-relevant contexts[21,22]. Comparison of gene expression changes in models of adipose dysfunction suggest that a combination of hypoxia and inflammation in *in vitro* mouse 3T3-L1 models most closely captures changes observed in diet induced obesity mouse models, compared to hypoxia or inflammation alone[158]. Therefore, we investigated regulatory mechanisms in the presence of hypoxia, inflammation, and combined hypoxia and inflammation in the SGBS human adipocyte model.

Identifying variants with regulatory effects after stimulation with excess free fatty acids or markers of metabolic disease, such as hypoxia and inflammation, may uncover additional mechanisms at GWAS loci for cardiometabolic traits. In this study, we profiled chromatin accessibility in the context of excess free fatty acids and produced high-quality profiles of accessibility. We also profiled chromatin accessibility and gene expression in the context of hypoxia, inflammation, and combined hypoxia and inflammation and produced quality profiles of gene expression changes.

Results

Chromatin accessibility in the context of free fatty acids

We profiled chromatin accessibility using ATAC-seq[69,88] in the context of excess free fatty acids, oleic acid or palmitic acid, and untreated controls in SGBS adipocytes[22] (Fig 3.1). We analyzed a final set of four replicates of day 20 (D20) adipocytes treated with 500 uM of oleic acid or 500 uM of palmitic acid for six days, compared to four replicates of an untreated control. These conditions were chosen based on previous studies of free fatty acid challenges in a 3T3-L1 mouse adipocyte model[157]. Final reads after quality filtering for our libraries ranged from ~16.9-86.8 million reads with an average of 44.8 million reads (Table 3.1). We identified ~105-191 thousand chromatin accessibility regions, hereafter referred to as peaks, per library and our libraries showed high quality, with an average transcription start site enrichment (TSS) of 5 and an average of 49% reads in peaks, in line with ENCODE standards for quality ATAC-seq libraries[74,76].

To identify a set of 111,996 peaks to test for differentially accessible peaks, we generated a set of consensus peaks present in a majority of each treatment (three out of four replicates) and merged the top 100,000 peaks from each consensus peak set. Principal component analysis (PCA) showed strong correlation with the technical quality measure of percent reads in peaks ($r^2 = 0.85$) on the first principal component, which explained 24% of the variance (Fig 3.2). When we corrected for percent reads in peaks, the first principal component reduced to explaining 16% of the variance, and libraries separated by treatment on the second principal component, which explained 15% of the variance (Fig 3.2). Based on these results, we proceeded with correcting for percent reads in peaks in downstream analyses.

To identify candidate regulatory elements, we tested for differentially accessible peaks between each treatment condition and the untreated control ($\log_2$ fold change (LFC) > 0; false discovery rate (FDR)<5%; Table 3.2). After correcting for percent reads in peaks, we identified only 37 significant peaks between oleic acid and the untreated control and 525 significant peaks between palmitic acid and the untreated control (Table 3.2). While these results could include interesting candidate regulatory elements, we did not proceed with further analyses due to the small numbers and an inability to rule out that the choices of treatment conditions resulted in few significant differences. However, these high-quality human adipocyte chromatin accessibility libraries could be a useful resource for future studies such as changes in chromatin accessibility at later timepoints of adipocyte maturity.

Optimizing treatment conditions of hypoxia and inflammation

To optimize treatment conditions for identifying regulatory elements that change with hypoxia and inflammation, we treated SGBS adipocytes at day five (D5) of differentiation and used a quantitative PCR (qPCR) array to measure changes in expression of genes relevant to hypoxia and inflammation. First, we treated adipocytes with exposure to 1% oxygen, 1% oxygen and 10 ng/mL of TNF-a, or 1% oxygen and 25 ng/mL of TNF-a and measured the expression levels of 42 hypoxia-relevant genes and four housekeeping control genes (Fig 3.3 and Table 3.7). For adipocytes treated for 24 hours with 1% oxygen alone, an expression level was measurable in 35 hypoxia-relevant genes, and 17 of those genes showed a LFC > 1 (Table 3.7). These results showed that exposure to 1% hypoxia affected hypoxia-relevant gene expression and we chose to proceed with this treatment. Next, we treated adipocytes for 24 hours with 10 ng/mL, 25 ng/mL, or 50 ng/mL of TNF-a and measured the expression levels of 92 inflammatory genes and four housekeeping control genes (Fig 3.3 and Table 3.8). For adipocytes treated with 10 ng/mL of TNF-a, an expression level was measurable for 60 inflammation-relevant genes, and 28 of those genes showed a LFC > 1 (Table 3.8). These results showed that treatment with TNF-a affected inflammatory-relevant genes. Due to similar changes in expression for cells treated with 10 ng/mL, 25 ng/mL, or 50 ng/mL of TNF-a, we proceeded with 10 ng/mL as the lowest concentration of TNF-a tested.

Identifying changes in chromatin accessibility due to hypoxia and inflammation

We profiled chromatin accessibility using ATAC-seq[69,88] in the context of hypoxia, inflammation, or hypoxia and inflammation, and untreated controls in SGBS adipocytes[22] (Fig 3.4). We analyzed replicates of day six (D6) adipocytes treated with 1% oxygen, 10 ng/mL TNF-a, or 1% oxygen and 10 ng/mL TNF-a for 24 hours, compared to replicates of untreated controls. After quality filtering, our libraries ranged from ~62.7 to 111.4 million reads, with an average of 85.7 million reads (Table 3.3), and we identified ~6.7 to 188.8 thousand peaks per library. To identify a final set of replicates to use for analysis, we filtered for quality control metrics of signal to noise (TSS enrichment > 2.5) and eliminated sample 4 from the control set as an outlier on PCA. Our final set of replicates showed improved quality metrics, with a range of ~62.7 to 111.4 million reads, with an average of 86.7 million reads (Table 3.3), and ~45.4 to 188.8 thousand peaks per library (Table 3.3). These libraries showed an average transcription start site enrichment (TSS) of 4 and an average of 11% reads in peaks. Despite the low signal-to-noise indicated by the low percentage of reads in peaks, we proceeded with analysis of differentially accessible peaks because we observed high (> 84%) overlap of the top 25,000 peaks with Roadmap Epigenomics[77] adipose nuclei enhancer and promoter regions (Table 3.3).

To identify a set of 64,830 peaks to test for differentially accessible peaks, we generated a set of consensus peaks present in a majority of each treatment (> 50%; 3/5 replicates for control and inflammation, 2/3 for hypoxia, and 2/2 for hypoxia and inflammation combined) and merged the top 40,000 peaks from each consensus peak set. PCA showed strong correlation with batch on the first principal component, which explained 42% of the variance (Fig 3.5). After batch correction, the first principal component reduced to explaining 35% of the variance with libraries separating by treatment with inflammation (Fig 3.5). Based on these results, we proceeded to correct for batch in downstream analyses.

To identify candidate regulatory elements, we tested for differentially accessible peaks between each treatment condition and the untreated control (LFC > 0; FDR < 5%; Table 3.4). After correcting for batch, we identified only 5,233 significant peaks that differ between inflammation and the untreated control, 17,610 significant peaks between combined hypoxia and inflammation and the untreated control, and no significant peaks between hypoxia and the untreated control (Table 3.4). These results could include interesting candidate regulatory elements, but we did not proceed with further analyses due to relatively poor quality of the ATAC-seq libraries as demonstrated by low TSS enrichment and percent reads in peaks.

Identifying changes in gene expression due to hypoxia and inflammation

We profiled gene expression using RNA-seq in the context of hypoxia, inflammation, or hypoxia and inflammation, and untreated controls in SGBS adipocytes[22] (Fig 3.4). We analyzed six replicates of each condition at day six (D6) of adipocyte differentiation treated with 1% oxygen, 10 ng/mL TNF-a, or 1% oxygen and 10 ng/mL TNF-a for 24 hours, compared to replicates of untreated controls. Final reads after quality filtering for all our libraries ranged from

~42.1 to 67.0 million reads, with an average of 54.6 million reads (Table 3.5). We identified 18,259 expressed genes (median normalized count >= 1 across all libraries). PCA showed strong correlation with batch (Fig 3.5). After batch correction, the first principal component separated by inflammation and explained 45% of the variance, while the second principal component separated by hypoxia and explained 13% of the variance (Fig 3.5). These results led us to correct for batch in downstream analyses.

To identify genes that change expression with exposure to hypoxia, inflammation, or hypoxia and inflammation, we identified genes differentially expressed between each treatment condition and the untreated controls (LFC > 1, FDR < 5%). With a higher threshold for significance than used for chromatin accessibility, we identified 573 differentially expressed genes between inflammation treatment and controls and 613 differentially expressed genes between the combined hypoxia and inflammation treatment and controls. Similar to the lack of significant results between hypoxia treatments and controls in chromatin accessibility, we only identified 4 genes that significantly differed between these conditions.

Discussion

In this study, we profiled chromatin accessibility and gene expression in a human adipocyte cell model treated with several contexts relevant to cardiovascular and metabolic disease. We produced high quality chromatin accessibility in adipocytes treated with oleic acid or palmitic acid, however, even using a lenient LFC threshold of LFC>0 we identified few changes between treatments and controls. We profiled chromatin accessibility in adipocytes treated with hypoxia, inflammation, or combined hypoxia and inflammation and we identified changes in gene expression between inflammation treatments and controls, however, our chromatin accessibility profiles showed low complexity, which complicated analysis. I will discuss some of the technical and biological factors that could have caused these results. Despite the lack of many context-dependent regulatory differences, these studies produced some high-quality chromatin accessibility and gene expression libraries that can be used for future studies.

In our study of regulatory elements that change with adipocyte exposure to oleic acid or palmitic acid, which should cause increased lipid storage, we produced high-quality profiles but identified few significant differences between treatments and untreated controls. One biological explanation for this outcome could be that mechanisms other than changes in chromatin accessibility, such as changes in transcription factor expression[159], could drive changes in cells. If we were to repeat this study, we could also perform RNA-seq to test for changes in gene expression including transcription factors during exposure to excess free fatty acids. We did identify changes in gene expression for other stimuli tested, including exposure to hypoxia and inflammation. One technical explanation for this outcome could be failure of uptake of the free fatty acids into the adipocytes, resulting in few differences between treatments and controls. If we were to repeat this study, we would test different concentrations and methods or sources of free fatty acid treatment and confirm enlargement of lipids within cells using assays such as oil red O staining. Despite the few significant results, these high-quality chromatin accessibility profiles identified consensus peaks in a mature human adipocyte cell model and could be used for future studies of adipocytes.

In our study of regulatory elements that change with adipocyte exposure to hypoxia, inflammation, and combined hypoxia and inflammation, which are markers of dysfunctional adipocytes, we were unable to produce high-quality chromatin accessibility profiles, but we did produce quality RNA-seq profiles and identified changes in gene expression between treatments and controls. One technical explanation for our low-quality chromatin accessibility profiles is the use of a hypoxia chamber in another building, which delayed time between nuclei isolation and library generation. In our experience with ATAC-seq, minimizing time between nuclei isolation and library generation has produced the highest quality libraries, possibly due to the quality of chromatin fixation. Despite the low quality of the ATAC-seq libraries, we proceeded with analyses due to high overlap with Roadmap Epigenomics[77] enhancer and promoter regions. However, with low quality, we could not be confident that the identified differentially accessible regions were representative of the treatment. We did produce high quality RNA-seq libraries between treatments that could be used in future studies of gene regulation under conditions of hypoxia and inflammation, however, we did not pursue further study without paired chromatin accessibility data due to limited novelty, as gene expression data exists for TNF-a treated adipocytes[160]. We also attempted to compare LFC of gene expression measured by qPCR and gene expression measured by RNA-seq, however, there were cases of disagreement in gene expression changes between methods. These differences could be caused by the qPCR and RNA-seq being measured in independent experiments. It is also possible that the primers used to assay gene expression by qPCR could affect the results and lead to differences compared to RNA-seq.

Although these studies did not produce significant results, they aided optimization of critical points in design of chromatin accessibility and gene regulation profiling of cells exposed to disease relevant contexts for future studies in our lab[83].

Methods

Cell culture:

For all treatments, SGBS cells[21] were generously provided by Dr. Martin Wabitsch (University of Ulm) and cultured as previously described[122]. Briefly, we cultured SGBS preadipocytes in serum-containing basal medium (DMEM:F12 + 33 uM biotin + 17 uM pantothenate) with 10% FBS until confluent, then rinsed in phosphate-buffered-saline (PBS) and differentiated for four days in medium supplemented with 0.01 mg/mL transferrin, 20 nM insulin, 200 nM cortisol, 0.4 nM triiodothyronine, 50 nM dexamethasone, 500 uM IBMX, and 2 uM rosiglitazone. After four days, we maintained differentiated SGBS cells in basal medium supplemented with 0.01 mg/mL transferrin, 20 nM insulin, 200 nM cortisol, 0.4 nM triiodothyronine until harvested.

Free fatty acid treatment:

Cells were treated with free fatty acids as previously described[156]. Briefly, oleic acid or palmitic acid (Sigma-Aldrich) were dissolved in ethanol and diluted in basal medium (DMEM:F12 + 33 uM biotin + 17 uM pantothenate) containing 1% FBS and 2% (wt/vol) BSA for 10 min at 55°C. Cells were maintained with BSA-conjugated free fatty acid containing media at 500 uM concentrations between D14 and D20.

Hypoxia and inflammation treatment:

Cells were treated with hypoxia by exposure to 1% oxygen for 24 hours in a controlled cell culture chamber. Cells were exposed to inflammation by final treatment with 10 ng/mL of TNF-a (Sigma-Aldrich) for 24 hours. 10 ng/mL, 25 ng/mL, and 50 ng/mL of TNF-a were used during optimization.

qPCR assessment of cells treated with hypoxia and inflammation:

Hypoxia and inflammation treatment conditions were optimized by testing treated cells for changes in expression using qPCR arrays for hypoxia (ThermoFisher, catalog: 4414090) and inflammation (ThermoFisher, catalog: 4414074). In brief, cells were treated with hypoxia and inflammation conditions, RNA was isolated using the Total RNA Purification Kit (product #17200) from Norgen Biotek (Ontario, Canada), cDNA was prepared using SuperScript (ThermoFisher, catalog: 11917010), and cDNA was added to each well of the array and cycled according to the conditions below. The ΔΔCt quantification method was used to analyze results[161].

Step	Temperature	Time	Cycles
UNG incubation	50°C	2 minutes	1
Enzyme activation	95°C	20 seconds	1
Denature	95°C	1 second	40
Anneal / Extend	60°C	20 seconds	

ATAC-seq Library Preparation:

For all treatments, we profiled chromatin accessibility in SGBS cells following the omni-ATAC-seq protocol[88] using unique, dual-barcoded indices. In brief, we isolated nuclei and used a cell countess to aliquot 50,000 nuclei per library and 5 uL of Tn5 per library. We cleaned the transposase reaction and final library with Zymo DNA Clean and Concentrator (D4029). We visualized and quantified libraries using a TapeStation, and sequenced with paired-end reads on Novaseq.

ATAC-seq Analysis:

We trimmed sequencing adapters and low quality base calls from the 3' ends of SGBS paired-end ATAC-seq reads using cutadapt[123] with parameters -q 20 –minimum-length 36. We aligned trimmed reads to the hg19 human genome[124] using bowtie2[125] with parameters –minins 36 –maxins 1000 –no-mixed –no-discordant –no-unal and selected nuclear chromosomal alignments with mapq>20 using samtools[125]. We removed alignments overlapping high-signal regions (Duke excluded and ENCODE/DAC exclusion list regions)[126] using BEDTools pairToBed[127] with the parameter -type notospan. We removed duplicate alignments using Picard MarkDuplicates (https://github.com/broadinstitute/picard) and generated ATAC-seq quality metrics using ataqv.[128]

We trimmed alignments so their 5' ends corresponded to the Tn5 binding site (+4 for + strand alignments and -5 for – strand alignments)[69] and smoothed signal by extending alignments 100 bp on either side of the Tn5 binding sites using BEDTools slop[127]. We called peaks (FDR<5%) with MACS2[129] with parameters -q 0.05 –nomodel –bdg and generated ATAC-seq signal bigwig files from MACS2 bedGraph files using the bedGraphToBigWig tool from ucsctools[130]. For free fatty acid treatments, we proceeded with analyses on a final set of libraries that met our signal-to-noise quality thresholds with a fraction of reads in peaks (FRiP) greater than 30% and a transcription start site enrichment greater than 4[76]. For hypoxia and inflammation treatments, we proceeded with analyses on a final set of libraries that met our signal-to-noise quality thresholds with a transcription start site enrichment greater than 2.5[76], and we additionally excluded sample 4 from the hypoxia and inflammation treatment controls as an outlier after PCA. For each analyzed treatment condition, we generated a set of consensus ATAC-seq peaks by merging peak genomic coordinates across replicates for a given treatment using BEDTools merge[127]. Then, we defined consensus peaks as merged peaks that overlapped individual replicate peaks in greater than 50% of replicates.

Identification of differentially accessible peaks:

We generated a set of merged peaks to test for differential chromatin accessibility for each treatment by merging the top 100,000 peaks for free fatty acid treatments and the top 40,000 consensus peaks for hypoxia and inflammation treatments (ranked by median peak p-value across replicates). We quantified the accessibility of these merged peaks in each library using featureCounts[131]. We computed the GC percent of each peak using BEDTools nuc[127] and generated within-library GC bias normalization factors using full quantile normalization with EDASeq[132]. We then used EDASeq GC bias normalization factors within DESeq2[133] and used DESeq2 size factors to control for differences in sequencing depth between libraries. We tested for differential chromatin accessibility using DESeq2[133] and classified significantly differential peaks with FDR < 5% and log fold change (LFC) > 1 or LFC > 0 as indicated between each treatment and the untreated control.

RNA-seq library preparation, read alignment, and identification of differentially expressed genes:

We isolated total RNA from SGBS cells exposed to hypoxia, inflammation, hypoxia and inflammation, or untreated controls using the Total RNA Purification Kit (product #17200) from Norgen Biotek (Ontario, Canada). Novogene (Beijing, China) generated poly-A RNA libraries and performed paired-end RNA sequencing (RNA-seq, read length 150 bp) using a NovaSeq 6000 (Illumina, California, USA). We trimmed sequencing adapters and low-quality base calls from the 3' ends of RNA-seq reads using cutadapt[123] with parameters -q 20 –minimum-length 36. We aligned reads to the hg19 human genome[124] using STAR[137] with parameters --sjdbOverhang 149 --twopassMode Basic --quantMode TranscriptomeSAM --outFilterMultimapNmax 20 --alignSJoverhangMin 8 --alignSJDBoverhangMin 1 --outFilterMismatchNmax 999 --outFilterMismatchNoverReadLmax 0.04 --alignIntronMin 20 --alignIntronMax 1000000 --alignMatesGapMax 1000000. We quantified expression of genes from GENCODE v29 lift37[138] and corrected for GC bias using salmon[140] with parameters –seqBias –gcBias –gencode. We generated RNA-seq quality metrics using the CollectRnaSeqMetrics tool from Picard (https://github.com/broadinstitute/picard). We used PCA to determine which replicates clustered. Within timepoint clusters, we observed additional clustering by batch that we corrected for in downstream analysis. To identify differentially expressed genes, we imported salmon transcript quantifications and collapsed to the gene level using tximport[139]. We retained genes with median DESeq2-normalized count >= 1 across all libraries. We tested for differential gene expression using DESeq2[133] and classified significantly different genes with FDR < 5% and LFC > 1 or LFC > 0 as indicated between each treatment and the untreated control.

Acknowledgements

Thank you to everyone who contributed to the work in this chapter, including Lianxin Hu and Qing Zhang for use of their hypoxia chamber, Martin Wabitsch for providing SGBS cells, Roadmap Epigenomics Consortium for data, Kevin Currin for guidance with data analysis, and Karen Mohlke for advising on study design, analyses, and interpretation.

Figures

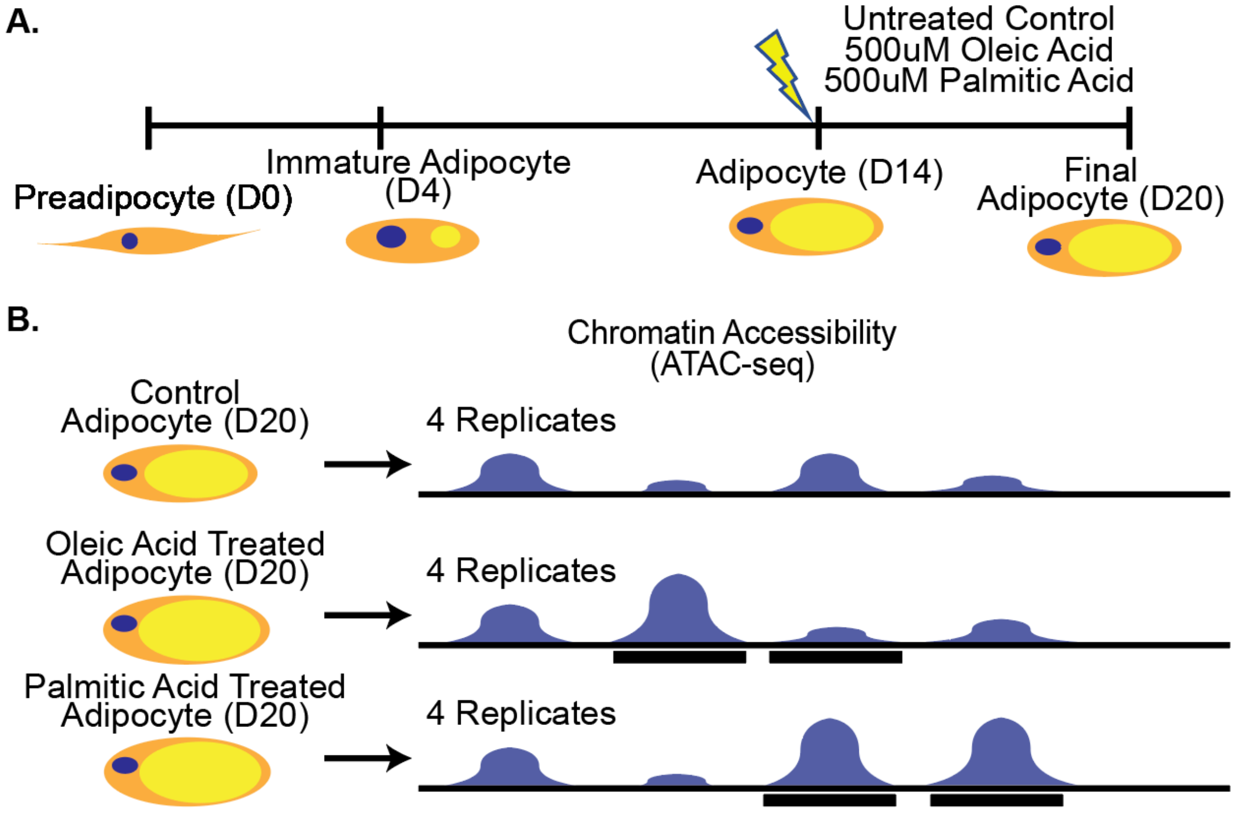
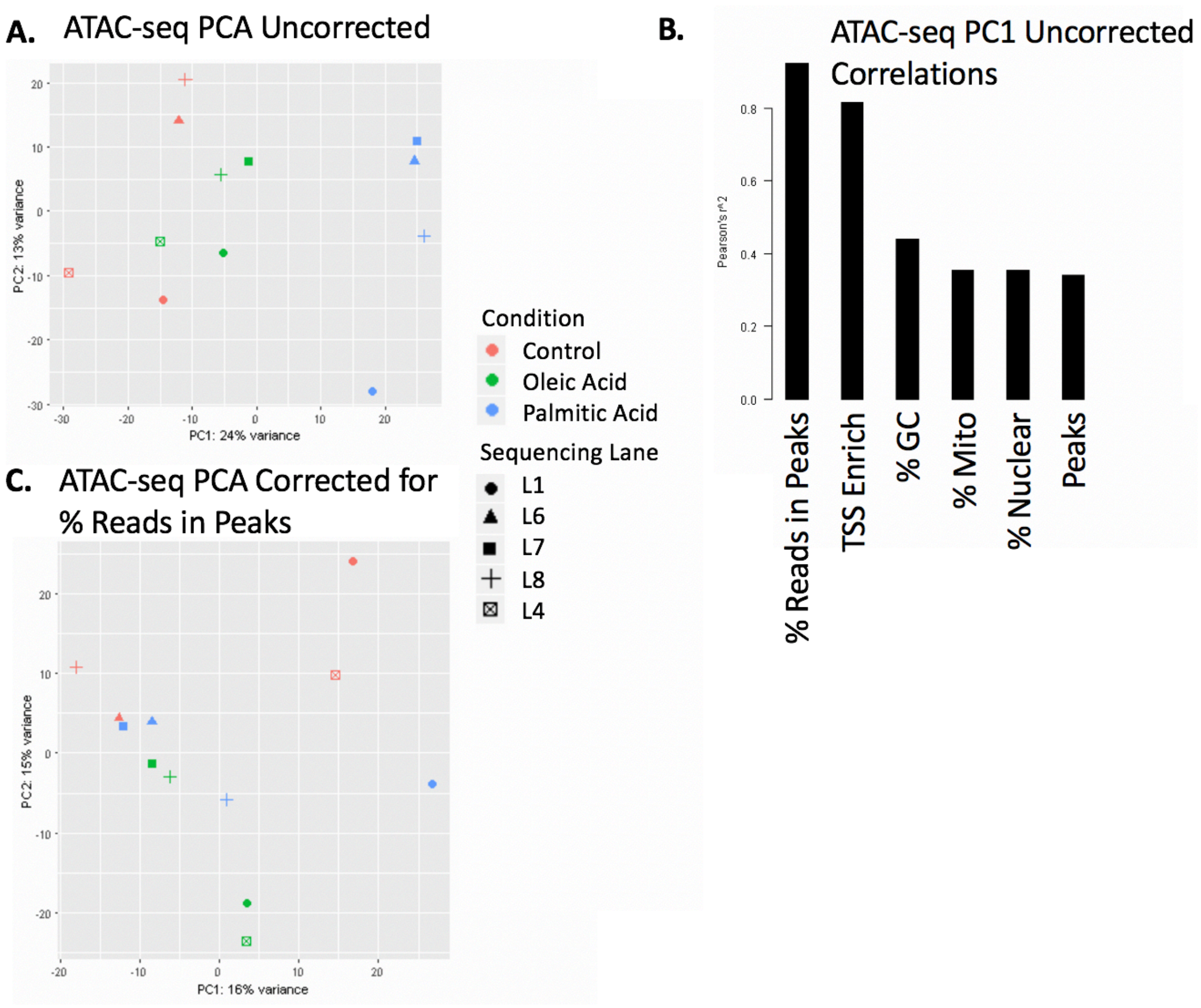


Figure 3.1: Genome-wide profiles of chromatin accessibility and gene expression in untreated controls and adipocytes treated with oleic or palmitic acid.

A. Schematic of experimental design. SGBS cells were started as preadipocytes (D0) and treated according to standard protocol with a differentiation medium for four days (D4) into immature adipocytes, and an adipocyte maintenance medium until harvested at day 5 (D20). At day 14 (D14) adipocytes were maintained untreated or treated with either 500 uM oleic acid or palmitic acid for six days. B. Chromatin accessibility (dark blue) and gene expression (green) profiles were generated on replicates from each treatment. Context-dependent peaks are shown as black bars. Chromatin accessibility and gene expression profiles were compared between each treatment and the untreated control.



A. ATAC-seq PCA Uncorrected

B. ATAC-seq PC1 Uncorrected Correlations

C. ATAC-seq PCA Corrected for % Reads in Peaks

Figure 3.2: PCA of ATAC-seq read count within peaks for free fatty acid treated and control adipocytes.

A. Plot of PCA of uncorrected ATAC-seq read counts within peaks for adipocytes for untreated controls (red), oleic acid treated (green), and palmitic acid treated (blue). Symbols are indicated in the legend for sequencing lane. B. Plot of Pearson's correlation for top six measured variables with principal component 1, showing a high correlation with percent reads in peaks. C. Plot of PCA of ATAC-seq read counts within peaks corrected for percent reads in peaks for adipocytes for untreated controls (red), oleic acid treated (green), and palmitic acid treated (blue). Symbols are indicated in the legend for sequencing lane.

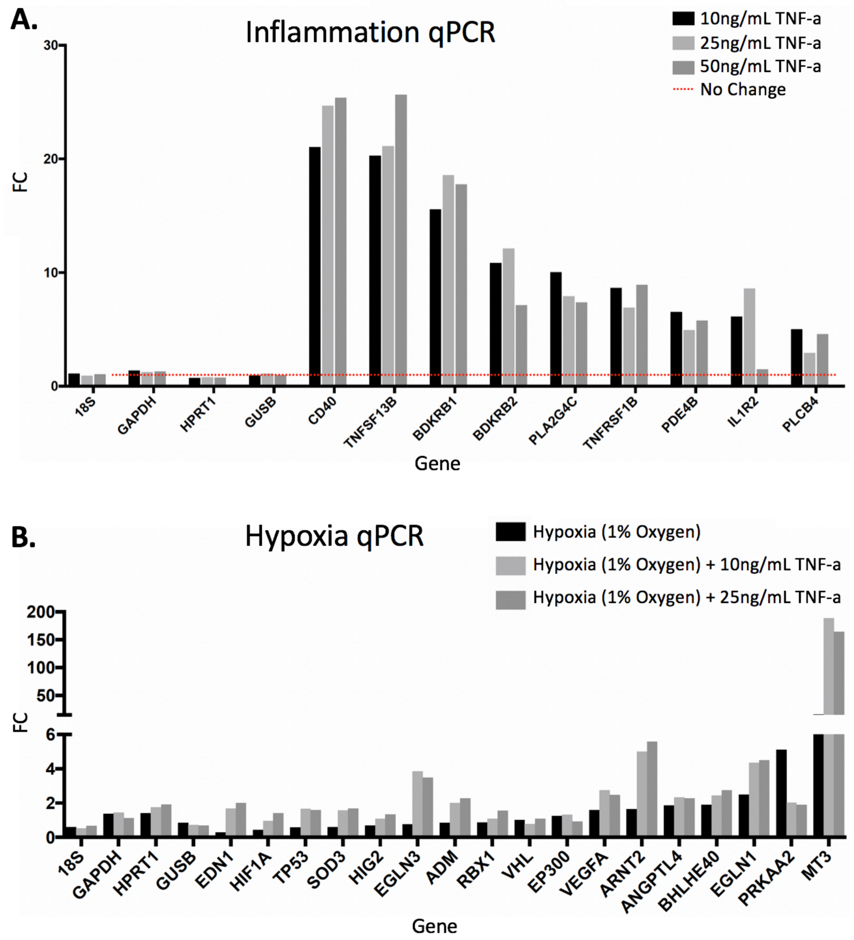


Figure 3.3: Gene expression measurements for inflammation and hypoxia treated adipocytes.

Bar plot of selected genes with the highest fold change (FC) compared to housekeeping genes for two qPCR arrays. Genes "18S, "GAPDH", "HPRT1", and "GUSB" are housekeeping genes. Full results are provided in Tables 3.7 and 3.8. A. Fold change for selected inflammation genes measured from a full set of 96 genes in an inflammation qPCR array. Three concentrations of TNF-a were tested, 10 ng/mL (black), 25 ng/mL (light grey), and 50 ng/mL (dark grey). B. Fold change for selected hypoxia genes measured from a full set of 46 genes in a hypoxia qPCR array. Three treatments were tested; hypoxia alone (black), hypoxia with 10 ng/mL of TNF-a (light grey), and hypoxia with 25 ng/mL of TNF-a (dark grey).

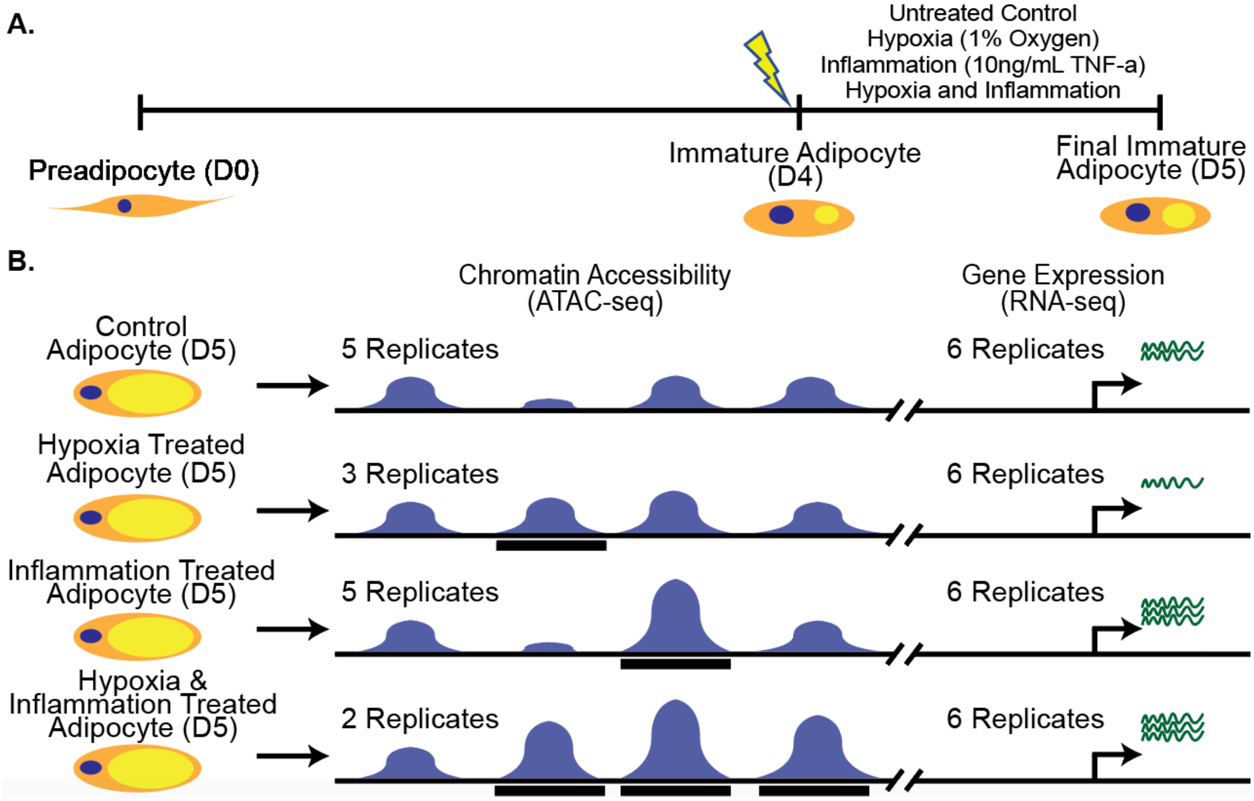

A.
Untreated Control
Hypoxia (1% Oxygen)
Inflammation (10ng/mL TNF-a)
Hypoxia and Inflammation
Preadipocyte (D0)
Immature Adipocyte (D4)
Final Immature Adipocyte (D5)
B.
Chromatin Accessibility (ATAC-seq)
Gene Expression (RNA-seq)
Control Adipocyte (D5)
5 Replicates
6 Replicates
Hypoxia Treated Adipocyte (D5)
3 Replicates
6 Replicates
Inflammation Treated Adipocyte (D5)
5 Replicates
6 Replicates
Hypoxia & Inflammation Treated Adipocyte (D5)
2 Replicates
6 Replicates

Figure 3.4: Genome-wide profiles of chromatin accessibility and gene expression in untreated controls and immature adipocytes treated with hypoxia and inflammation.

A. Schematic of experimental design. SGBS cells were started as preadipocytes (D0) and treated according to standard protocol with a differentiation medium for four days (D4) into immature adipocytes, and an adipocyte maintenance medium until harvested at day five (D5). At D4 immature adipocytes were maintained untreated or treated with either 1% oxygen, 10 ng/mL TNF-a, or 1% oxygen and 10 ng/mL TNF-a. B. Chromatin accessibility (dark blue) and gene expression (green) profiles were generated on replicates from each treatment. Context-dependent peaks are shown as black bars. Chromatin accessibility and gene expression profiles were compared between each treatment and the untreated control.

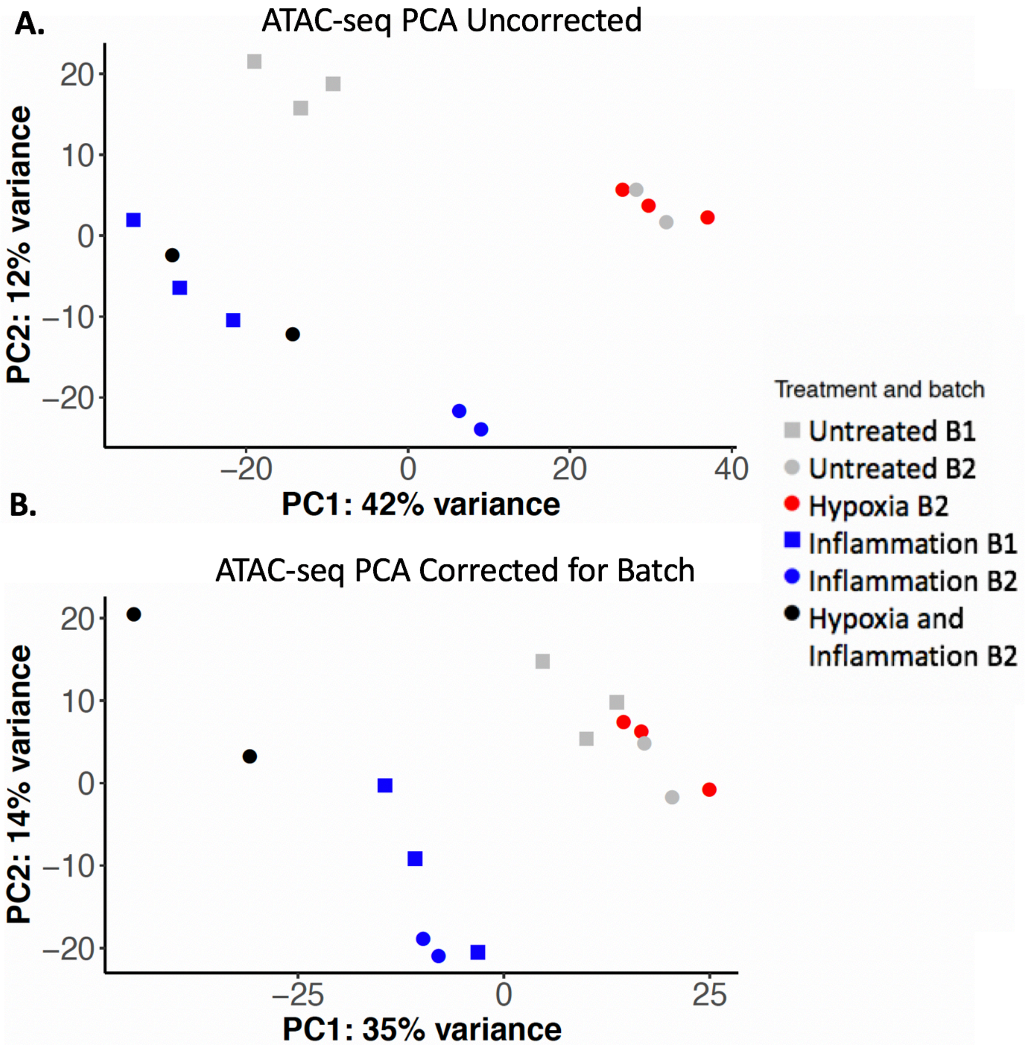


Figure 3.5: PCA of ATAC-seq read count within peaks for hypoxia and inflammation treated and control adipocytes.

Plots of PCA for ATAC-seq read counts within peaks for adipocytes for untreated controls (grey), 1% hypoxia treated (red), 10 ng/mL TNF-a treated (blue), and combined 1% hypoxia and 10 ng/mL treated (black). Symbols are indicated in the legend for batch 1 and batch 2. A. PCA for uncorrected ATAC-seq reads in peaks. B. PCA for ATAC-seq reads in peaks corrected for batch.

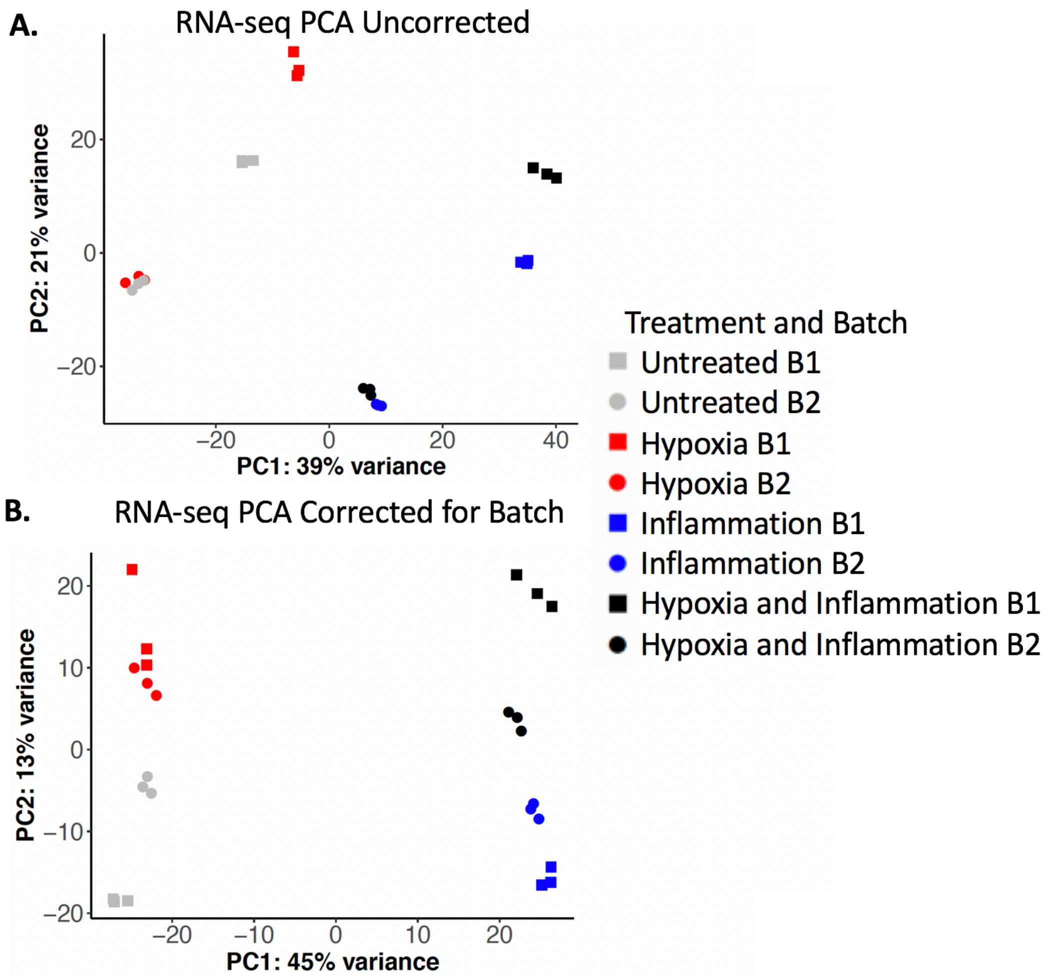


Figure 3.6: PCA of RNA-seq reads for hypoxia and inflammation treated and control adipocytes.

Plots of PCA for RNA-seq reads for adipocytes for untreated controls (grey), 1% hypoxia treated (red), 10 ng/mL TNF-a treated (blue), and combined 1% hypoxia and 10 ng/mL treated (black). Symbols are indicated in the legend for batch 1 and batch 2. A. PCA for uncorrected RNA-seq reads. B. PCA for RNA-seq reads corrected for batch.

Tables

Treatment Condition	Sample ID	Final Reads	Peaks	% Reads in Peaks	TSS Enrichment
Untreated	1	31,703,420	144,910	49.5	5.1
	2	66,230,646	165,895	54.6	4.8
	3	34,472,260	118,264	36.3	5.4
	4	16,869,394	105,182	44.6	4.8
Oleic Acid	5	51,581,504	155,558	49.1	5.1
	6	38,459,362	145,520	51.5	5.3
	7	86,765,094	184,391	55.6	5.5
	8	17,132,254	113,875	46.8	4.8
Palmitic Acid	9	49,187,804	149,274	40.9	5.9
	10	58,339,158	179,432	53.4	5.9
	11	22,614,914	125,490	47.5	6.2
	12	64,763,242	191,561	52.9	6.3

Table 3-1: Sequencing and alignment quality metrics for free fatty acid treatments ATAC-seq libraries.

Summary of sequencing and alignment quality metrics for libraries used in free fatty acid treatment analysis. “Final Reads” indicated the final number of reads used to call peaks for each library after quality filtering as described in methods.

Correction	FFA	Significant Peaks
Uncorrected	Oleic Acid	15
	Palmitic Acid	8,982
Corrected for % Reads in Peaks	Oleic Acid	37
	Palmitic Acid	525

Table 3-2: Summary of context-dependent peaks for free fatty acid treatments.

Counts of the number of significant peaks (DESeq2, $\log_2$ fold change (LFC) > 0, FDR < 5%) for each free fatty acid treatment compared to the untreated control from the top 100,000 peaks.

Treatment Condition	ID	Batch	Final Reads	Peaks	%Reads in Peaks	TSS Enrichment	Roadmap Overlap
Untreated	1	B1	62,656,480	61,422	5.1	3.1*	85.4
	2	B1	79,784,622	64,120	5.3	2.6*	90.8
	3	B1	76,472,046	131,772	10.6	3.8*	90.2
	4	B2	84,355,260	71,736	6.2	2.9	92.1
	5	B2	89,770,430	142,695	15.3	4.5*	92.8
	6	B2	94,145,918	169,431	16.5	4.4*	92.7
Hypoxia	7	B1	73,309,456	7,705	0.4	1.5	25.1
	8	B1	86,753,708	18,842	1.0	1.5	17.8
	9	B1	107,592,474	31,228	1.5	1.5	21.0
	10	B2	107,661,688	165,153	17.5	4.6*	92.6
	11	B2	95,030,782	188,758	17.1	4.7*	92.8
	12	B2	71,285,678	127,338	14.7	4.0*	93.2
Inflammation	13	B1	111,421,020	145,157	9.7	2.9*	84.5
	14	B1	91,061,910	66,646	5.5	2.5*	89.8
	15	B1	79,946,116	45,441	3.6	2.6*	85.0
	16	B2	82,649,692	139,183	13.6	4.0*	92.1
	17	B2	73,420,880	70,893	4.5	2.1	61.2
	18	B2	73,567,024	106,153	11.0	3.9*	92.3
Hypoxia and Inflammation	19	B2	81,776,294	6,678	0.4	1.4	18.8
	20	B2	96,598,812	36,508	1.9	1.6	22.8
	21	B2	66,864,866	26,333	1.6	1.8	32.3
	22	B2	97,223,278	90,429	6.0	2.8*	84.4
	23	B2	88,565,750	107,004	8.5	3.2*	91.3

Table 3-3: Sequencing and alignment quality metrics for hypoxia and inflammation treatments ATAC-seq libraries.

Summary of sequencing and alignment quality metrics for libraries sequenced in hypoxia and inflammation treatments. “ID” indicates an ID for each library prepped. “Batch” indicates libraries that were prepared in batch 1 (B1) and batch 2 (B2). “Final Reads” indicated the final number of reads used to call peaks for each library after quality filtering as described in methods. All sequenced libraries are summarized, but only samples with an “*” in the “TSS Enrichment” column were used in analysis, as a TSS enrichment greater than 2.5 was used as a primary quality filter to select final libraries (sample ID 2 was excluded as a PCA outlier despite a TSS enrichment of 2.6). “Roadmap Overlap” indicated the percent of the top 25,000 peaks in the library that overlap a Roadmap adipocyte nuclei promoter or enhancer region.

	Peaks (LFC>1)	Peaks (LFC>0)
Hypoxia	0	0
Inflammation	14	5,233
Hypoxia and Inflammation	1	17,610

Table 3-4: Summary of context-dependent peaks for hypoxia and inflammation treatments. Counts of the number of significant peaks (DESeq2, $\log_2$ fold change (LFC) > 0 of LFC > 1 as indicated, FDR < 5%) for each treatment compared to the untreated control.

Treatment	Sample	Total reads	Transcript reads
Untreated	1	41,496,230	36,107,570
	2	63,920,376	54,033,844
	3	64,406,330	47,965,522
	4	48,126,922	42,040,006
	5	67,042,866	58,163,240
	6	58,771,696	51,334,066
Hypoxia	7	56,255,264	48,294,026
	8	51,517,296	44,485,496
	9	42,080,716	36,183,506
	10	53,883,260	46,210,452
	11	51,870,658	44,555,260
	12	45,454,808	39,237,142
Inflammation	13	57,220,416	49,058,066
	14	58,819,082	51,307,148
	15	55,527,252	48,196,454
	16	55,318,696	48,240,304
	17	61,436,916	53,307,336
	18	44,938,146	38,228,458
Hypoxia and Inflammation	19	53,467,730	46,199,886
	20	53,843,152	45,577,320
	21	51,263,922	43,916,036
	22	55,982,774	48,175,134
	23	60,837,418	52,327,780
	24	43,787,248	37,832,012

Table 3-5: RNA sequencing and alignment quality metrics for hypoxia and inflammation treatments.

Summary of total final sequencing and transcript reads used for each RNA-seq library after quality filtering.

	Genes
Hypoxia	4
Inflammation	573
Hypoxia and Inflammation	613

Table 3-6: Summary of context-dependent genes for hypoxia and inflammation treatments. Counts of the number of significant genes (DESeq2, $\log_2$ fold change (LFC) > 1, FDR < 5%) for each treatment compared to the untreated control.

	qPCR LFC			RNA-seq LFC		
	Hypoxia	Hypoxia + Inf. (10 ng/mL TNF-a)	Hypoxia + Inf. (25 ng/mL TNF-a)	Hypoxia	Inf. (10 ng/mL TNF-a)	Hypoxia + Inf. (10 ng/mL TNF-a)
GAPDH	0.34	0.49	0.36	0.59	0.32	0.86
HPRT1	0.38	0.77	1.12	0.23	-0.05	0.33
GUSB	-0.34	-0.49	-0.36	-0.25	0.03	-0.17
ADM	-0.22	1.01	1.19	0.64	1.18	1.60
ANGPTL4	0.89	1.22	1.19	0.78	0.02	1.01
ARNT	-0.30	-0.31	-0.04	0.09	0.48	0.39
ARNT2	0.72	2.32	2.48	-0.11	-0.13	-0.61
ATP1B1	-1.31	-4.23	-3.91	-0.55	-2.26	-2.62
BHLHE40	0.93	1.28	1.46	0.92	0.16	1.07
CASP1	-0.09	-0.20	-0.13	-0.27	-0.11	-0.60
CREBBP	-1.71	-1.41	-11.03	0.18	0.18	0.27
DDIT4	-0.23	-0.79	-0.24	0.39	-0.27	0.32
DDIT4L	-1.68	-2.19	-1.97	0.96	-0.24	0.66
EDN1	-1.71	0.75	1.00	1.46	2.49	2.29
EGLN1	1.33	2.12	2.17	0.82	0.79	1.38
EGLN2	-2.15	-1.78	-1.91	0.12	-0.01	0.04
EGLN3	-0.39	1.94	1.80	-0.24	-0.67	NA
EP300	0.33	0.39	-0.11	0.33	0.30	0.59
EPAS1	-1.83	-2.56	-2.01	-0.40	-0.55	-1.02
EPO	NA	NA	NA	0.00	NA	NA
FRAP1	-1.56	-1.27	-0.93	NA	NA	NA
HIF1A	-1.22	-0.06	0.49	0.25	0.94	0.62
HIF1AN	-1.19	-1.08	-0.69	-0.26	-0.13	-0.39
HIF3A	-0.98	-1.70	-1.05	-1.44	0.38	-0.62
HIG2	-0.54	0.12	0.42	NA	NA	NA
HMOX1	-0.38	-0.30	-0.32	0.57	0.09	0.66
HYOU1	-1.91	-1.31	-0.73	-0.19	0.01	0.12
IGFBP1	NA	NA	NA	2.12	NA	NA
ING4	-1.39	-1.69	-1.39	-0.15	0.29	0.07
MB	3.97	7.56	7.36	-1.12	3.61	3.24
MT3	NA	NA	NA	0.94	0.98	3.00
NOS1	NA	NA	NA	0.00	NA	NA
NOS2	NA	NA	NA	-0.33	NA	NA
NOS3	NA	NA	NA	-0.24	-0.12	-0.29

NOTCH1	-1.94	-1.97	-1.58	-0.39	-0.33	-0.56
PIK3CA	-0.77	-1.02	-0.66	-0.17	-0.68	-0.74
PRKAA1	-0.85	-0.46	-0.30	0.29	0.28	0.55
PRKAA2	2.35	1.02	0.92	0.28	-0.48	-0.06
PTEN	-0.77	-1.37	-0.99	-0.16	0.49	0.17
SLC2A8	NA	NA	NA	-1.59	0.19	-1.54
SOD3	NA	NA	NA	-0.14	0.91	0.89
TGFBR2	-1.08	-1.07	-0.57	-0.13	-0.01	-0.10
TP53	-0.80	0.74	0.67	0.05	1.02	1.07
VEGFA	0.67	1.45	1.31	0.59	0.39	0.78
VHL	0.01	-0.38	0.11	0.06	0.25	0.34
CUL2	-1.49	-1.12	-0.85	-0.08	-0.11	-0.06
RBX1	-0.21	0.10	0.63	0.20	0.20	0.39

Table 3-7: LFC of treatments analyzed by hypoxia qPCR array and context-dependent RNA-seq analysis.

Three combinations of hypoxia and inflammation (“Inf.”) were tested for expression of 46 genes in a hypoxia qPCR array. qPCR LFC reports the LFC measured for three treatments by the qPCR array. A negative LFC indicates a decrease in expression compared to the housekeeping controls and a positive LFC indicates an increase in expression compared to the housekeeping controls. RNA-seq LFC reports the LFC for the three treatment conditions compared to the untreated controls for genes also measured by qPCR array. A negative LFC indicates a decrease in expression compared to the untreated controls and a positive LFC indicates an increase in expression compared to the untreated controls. “NA” indicates that expression was not accurately measured for that condition.

	qPCR LFC			RNAseq LFC		
Gene	Inf. (10 ng/mL TNF-a)	Inf. (25 ng/mL TNF-a)	Inf. (50 ng/mL TNF-a)	Hypoxia	Inf. (10 ng/mL TNF-a)	Hypoxia + Inf. (10 ng/mL TNF-a)
GAPDH	0.46	0.30	0.38	0.59	0.32	0.86
HPRT1	-0.48	-0.36	-0.40	0.23	-0.05	0.33
GUSB	-0.10	0.17	-0.05	-0.25	0.03	-0.17
A2M	-1.52	-0.88	-1.29	-0.30	0.15	0.06
ADRB1	NA	NA	NA	-0.39	-1.30	-0.35
ADRB2	-1.94	-1.57	-1.63	-0.58	-1.48	-1.81
ALOX12	NA	NA	NA	0.39	0.26	-1.50
ALOX5	NA	NA	NA	NA	NA	NA
ANXA1	-0.13	-0.20	-0.04	0.40	-0.04	0.30
ANXA3	1.28	1.15	0.83	0.60	0.98	0.69
ANXA5	0.96	0.94	0.93	0.13	0.26	0.47
KLK3	NA	NA	NA	NA	NA	NA
BDKRB1	3.96	4.22	4.15	-0.06	4.48	4.81
BDKRB2	3.44	3.60	2.83	0.71	3.66	3.57
CACNA1C	0.76	0.51	0.38	0.20	0.19	-0.20
CACNA1D	NA	NA	NA	-0.85	-3.07	NA
CACNA2D1	-0.32	-0.10	-0.70	-0.73	-0.59	-0.81
CACNB2	-2.28	NA	-1.56	-1.20	0.08	-0.71
CACNB4	-1.63	0.04	0.95	-0.24	-0.06	0.45
CASP1	-0.83	-0.85	-0.58	-0.27	-0.11	-0.60
CD40	4.40	4.63	4.67	-0.26	2.44	2.05
CD40LG	NA	NA	NA	-0.32	NA	NA
CES1	0.09	-0.06	0.04	-0.50	-0.13	-0.18
LTB4R	0.15	0.77	0.27	-0.56	-1.04	-0.92
MAPK14	-0.09	-0.18	-0.13	-0.05	-0.13	-0.08
NR3C1	-0.44	-0.27	-0.43	0.03	0.05	-0.08
HPGD	-2.00	-2.63	-2.47	0.30	-1.90	-1.53
HRH1	-0.13	-0.02	0.47	0.78	1.17	0.77
HRH2	NA	NA	NA	0.45	0.06	NA
HTR3A	NA	NA	NA	0.76	NA	NA
ICAM1	7.53	7.74	7.52	0.74	7.27	7.11
IL1R1	-0.10	-0.27	-0.25	-0.20	-0.30	-0.50
IL2RA	NA	NA	NA	NA	NA	NA
IL2RB	NA	NA	NA	-1.11	2.11	3.17

IL2RG	NA	NA	NA	-0.84	NA	NA
IL13	NA	NA	NA	1.36	NA	NA
ITGAL	NA	NA	NA	0.09	2.02	-0.45
ITGAM	NA	NA	NA	0.00	NA	NA
ITGB1	0.36	0.06	0.26	-0.13	-0.07	-0.23
ITGB2	-0.68	-0.19	0.34	-0.07	0.31	0.88
KLK1	NA	NA	NA	0.65	NA	NA
KLK2	NA	NA	NA	-2.04	NA	NA
KLKB1	NA	NA	0.57	0.18	-0.46	NA
KNG1	1.30	NA	NA	-0.32	NA	NA
LTA4H	-0.16	-0.56	-0.36	-0.38	0.02	-0.38
LTC4S	0.31	1.27	0.35	1.66	1.41	1.95
MC2R	NA	NA	NA	NA	NA	NA
NFKB1	1.69	1.61	1.44	0.13	1.42	1.52
NOS2	NA	NA	NA	-0.33	NA	NA
PDE4A	0.99	1.16	0.92	0.14	0.27	0.18
PDE4B	2.71	2.30	2.53	-0.22	2.16	1.52
PDE4C	NA	NA	NA	1.76	-0.65	1.80
PDE4D	-1.15	-1.14	-0.68	0.40	-0.05	0.55
PLA2G1B	0.08	2.56	0.30	-1.26	0.06	NA
PLA2G2A	0.17	-0.03	-0.52	-0.15	-0.45	-0.68
PLA2G5	-1.34	NA	0.60	-0.74	-1.60	-3.51
PLCB2	NA	NA	NA	-0.30	-2.29	-0.31
PLCB3	-0.42	-0.82	-0.53	-0.21	0.03	-0.24
PLCB4	2.33	1.55	2.19	1.81	1.98	2.53
PLCD1	-0.62	-0.45	-0.59	-0.20	0.10	-0.18
PLCG1	-0.45	-0.67	-0.69	-0.03	-0.17	-0.23
PLCG2	0.28	-0.75	-1.33	-0.58	0.56	0.48
MAPK1	0.05	0.07	-0.02	0.12	0.14	0.17
MAPK3	0.89	0.31	0.71	-0.10	0.50	0.44
MAPK8	-0.62	-0.56	-0.39	0.41	-0.08	-0.03
PTAFR	NA	NA	-1.29	-0.76	-1.52	-1.06
PTGDR	NA	NA	NA	-0.33	3.93	4.36
PTGER2	1.56	1.67	1.78	0.35	1.71	2.25
PTGER3	-1.34	-1.46	-1.33	-0.05	-0.17	-0.17
PTGFR	2.06	1.85	1.64	-0.57	2.15	1.61

PTGIR	0.35	0.94	1.29	2.43	4.20	4.74
PTGIS	0.28	1.09	0.35	-0.46	0.62	0.00
PTGS1	-0.46	-0.44	-1.08	-0.51	-0.74	-1.20
PTGS2	-0.75	-0.79	-0.70	0.64	-0.65	-0.13
TBXA2R	-1.16	1.38	-1.03	-0.25	1.68	1.18
TBXAS1	-0.56	0.25	-0.47	0.15	-0.21	-0.73
TNF	1.18	2.31	1.56	1.10	1.91	NA
TNFRSF1A	0.19	0.01	-0.13	-0.03	0.47	0.41
TNFRSF1B	3.11	2.79	3.16	-0.45	3.37	2.79
VCAM1	9.03	9.25	9.37	-1.22	7.28	6.17
IL1R2	2.61	3.10	0.58	0.26	-0.08	-0.22
PLA2G7	NA	NA	NA	NA	NA	NA
PLA2G10	NA	NA	NA	-0.32	NA	NA
PLA2G4C	3.33	2.98	2.88	-1.51	2.22	1.16
IL1RL1	0.34	2.29	0.01	-0.30	2.20	1.64
HTR3B	NA	NA	NA	-2.11	NA	NA
TNFSF13B	4.34	4.40	4.68	-0.48	2.92	2.59
CYSLTR1	NA	NA	NA	0.00	NA	NA
HRH3	NA	NA	NA	NA	NA	NA
PLA2G2D	NA	NA	NA	NA	NA	NA
IL1RAPL2	1.94	1.96	0.63	1.36	3.61	NA
KLK14	-1.01	NA	0.69	-0.09	-1.26	NA
PLCE1	0.25	0.19	0.03	-0.25	0.17	-0.18
KLK15	NA	NA	NA	NA	NA	NA
LTB4R2	NA	NA	-0.64	-0.43	-0.42	-0.56

Table 3-8: LFC of treatments analyzed by inflammation qPCR array and context-dependent RNA-seq analysis.

Three concentrations of inflammation ("Inf.") were tested for expression of 96 genes in an inflammation qPCR array. qPCR LFC reports the LFC measured for three treatments by the qPCR array. A negative LFC indicates a decrease in expression compared to the housekeeping controls and a positive LFC indicates an increase in expression compared to the housekeeping controls. RNA-seq LFC reports the LFC for the three treatment conditions compared to the untreated controls for genes also measured by qPCR array. A negative LFC indicates a decrease in expression compared to the untreated controls and a positive LFC indicates an increase in expression compared to the untreated controls. “NA” indicates that expression was not accurately measured for that condition.

CHAPTER 4: SEX-BIASED CHROMATIN ACCESSIBILITY IN LIVER

Introduction

Genome-wide association studies (GWAS) have identified thousands of loci associated with cardiometabolic traits, including loci with differential effects by sex[51,162–166], however identifying the mechanisms at these loci remains challenging[54]. The mechanisms remain particularly challenging due to the large number of noncoding loci[53]. Many noncoding cardiometabolic GWAS loci colocalize with expression quantitative trait (eQTL) loci in disease-relevant tissues, suggesting a regulatory mechanism[61,62,64,65]. Active regulatory elements are found in accessible regions of the genome[58], therefore chromatin accessibility profiles in disease-relevant tissues and contexts will aid identification of regulatory elements that alter gene expression to affect cardiometabolic traits.

Liver plays an important role in cardiometabolic traits through biological processes such as lipid metabolism, drug metabolism, and glucose storage[23]. Liver eQTL have been identified in multiple studies, and a subset of liver eQTL colocalize with cardiometabolic trait loci[62,67,167]. Liver QTL have been identified for histone markers of active regulatory regions such as H3K27ac and H3K4me3[167] as well as for chromatin accessibility[28] and a subset of QTL colocalize with cardiometabolic trait loci. While some GWAS loci colocalize with QTL in disease-relevant tissues, others only colocalize in disease-relevant contexts[68]. Therefore, further study of chromatin accessibility in disease-relevant contexts could identify additional regulatory mechanisms.

Sex is a disease-relevant context for cardiometabolic traits[24–26]. Many cardiometabolic diseases display sex differences in prevalence, including those in liver such as non-alcoholic fatty liver disease (NAFLD), which is significantly more prevalent in men than in pre-menopausal women[25]. Furthermore, sex differences in drug metabolism can impact treatment and health outcomes for many cardiometabolic diseases[25,168,169]. Sex-stratified GWAS analyses have identified traits demonstrating sexual dimorphism including seven loci for measures of body fat distribution[49] and 64 for blood lipids[51]. Sex-biased gene expression has also been identified in many tissues, including liver[62,170]. Sex-biased chromatin accessibility has been identified in cell types such as peripheral blood mononuclear cells[171]. Analyses of sex-biased chromatin accessibility compared to chromatin accessibility at the promoters of sex-biased genes suggests that sex-biased genes are likely altered by distal regulatory elements[172]. Therefore, identification of sex-biased chromatin accessibility may identify candidate regulatory elements and those associated with cardiometabolic traits could reveal key mechanisms and improve health outcomes.

Genetic and environmental factors can affect gene regulation and disease risk[173]. Identification of chromatin accessibility in a large number of samples can capture more genetic and environmental variation that contributes to disease risk. Environmental factors that can introduce variability between samples includes age, drug use, disease, cause of death, and hormonal statuses such as puberty and menopause. In this study, I used samples of liver tissue from deceased organ donors not selected for any known disease. Only limited data was available on environmental factors that could contribute to variability between samples.

In this study we identified consensus chromatin accessibility in 139 human liver samples and identified chromatin accessibility regions that differ between males and females. We linked these regions of sex-biased chromatin accessibility to eQTL[62] in liver and to GWAS traits[44].

Results

Sex-biased liver chromatin accessibility identified candidate regulatory elements

We profiled chromatin accessibility using the Assay for Transposase Accessible Chromatin (ATAC-seq) in human liver tissue samples from 93 male and 46 female organ donors aged 2-81 years (Fig 4.1, Table 4.1) for which ATAC-seq data met sequencing quality thresholds (Tables 4.2 and 4.3; Methods: TSS enrichment >= 4, percent reads in peaks >= 10). Each sample was prepared in triplicate or quadruplicate libraries, and the best library, determined by highest TSS enrichment, was used for analyses. The libraries had an average of ~54.5 million filtered reads and demonstrated high quality in line with ENCODE standards[74,76] with an average TSS enrichment of 8.0 and an average percent reads in peaks of 30.0% (Table 4.1). Sex of genotype data from the same samples was verified to match reported sex using PLINK[142], all genotype samples were found to correctly match to ATAC-seq profiles using verifyBamID[174175], and sex of ATAC-seq profiles was further verified through inspection of Y chromosome signals (Fig 4.2). In the 139 samples, we identified 231,736 autosomal consensus liver tissue peaks by merging genomic coordinates for peaks present in a liberal definition of at least 5% or more of the samples (n>7). We also considered a more stringent set of 172,813 autosomal consensus liver tissue peaks present in 10% or more of the samples (n>14) (Fig 4.1).

Principal component analysis (PCA) of the 5% consensus peak set showed that 19% of variance was explained by PC1, which demonstrated moderate correlation with the data quality metric percent reads in peaks (Fig 4.3A, Pearson's $r^2 = 0.65$). After adjusting for percent reads in peaks, PC1 explained 11% of the variance and PC1's highest, but modest correlation was with TSS enrichment (Fig 4.3B, Pearson's $r^2 = 0.25$). Based on these results, we decided to adjust for percent reads in peaks in downstream analyses. Despite evidence that chromatin accessibility differences can increase with age[176], we did not observe correlation with age and any of the top five PCs (Pearson's $r^2 < 0.1$, Fig 4.4B-F), therefore we decided not to adjust for age in downstream analyses. Additionally, none of the top five PCs were highly corelated with reported ancestry (Pearson's $r^2 < 0.1$), therefore we decided not to adjust for ethnicity in downstream analyses.

To predict regulatory elements in liver tissue that contribute to sex differences, hereafter referred to as sex-biased peaks, we identified differentially accessible peaks between males and females ($\log_2$ fold change (LFC)>0; false discovery rate (FDR) < 5%; Table 4.4). We defined male-biased peaks as peaks that are significantly more accessible in males compared to females and female-biased peaks as peaks that are significantly more accessible in females compared to males. Using the 10% consensus peaks adjusted for percent reads in peaks, we identified 774 sex-biased peaks (0.45% of 172,813 total peaks), including 384 male-biased and 390 female-biased (Table 4.4, Fig 4.5). These 774 sex-biased peaks spanned all 22 autosomal chromosomes (Table 4.5). We considered alternate thresholds of LFC and FDR. At a more stringent LFC threshold (LFC > 1, FDR < 5%) we did not observe any significant results. Our maximum significant LFC observed using the threshold LFC > 0 was 1.7, with an average significant LFC of 0.5. These results indicate that these sex-biased peaks do not represent strong differences in accessibility between sexes. Next, for comparison with a study on sex-biased chromatin accessibility in peripheral blood mononuclear cells that identified 577 sex-biased regions (0.69%

of tested regions)[171], we applied a less stringent threshold (LFC > 0, FDR < 10%), and identified a more comparable percent of tested peaks as sex biased (1300 autosomal sex-biased regions, which represents 0.75% of our 172,813 tested regions). However, we proceeded with analyses using the more stringent threshold (FDR < 5%) due to the weak effects at sex-biased peaks. We observed similar numbers of significant results with the more liberal definition of consensus peaks (5% consensus = 741, 10% consensus = 774; Table 4.4), 89% (662) of which overlapped between analyses. Therefore, we used the 774 sex-biased peaks from the stringent definition of consensus peaks adjusted for percent reads in peaks in our downstream analyses.

Linking sex-biased liver chromatin accessibility to genes

To link sex-biased peaks to genes, we identified sex-biased regions that overlap liver eQTL signals[62], with a signal defined as all variants in high linkage disequilibrium with a lead eQTL variant (methods, $r^2 > 0.8$). Of 774 sex-biased liver peaks, 71 overlapped a liver eQTL signal linked to 81 unique genes (Table 4.5, Figure 4.6). These 71 peaks spanned an average width of 1304 base pairs each, compared to an average of 998 base pairs each for the full set of 774 sex-biased peaks. An increase in average width of peaks overlapping a liver tissue eQTL variant compared to the average width of sex-biased peaks could indicate increased risk of variants overlapping a peak by chance.

Linking sex-biased liver chromatin accessibility to GWAS traits

To identify genetic variants that may have a sex-biased mechanism on disease traits, we identified GWAS variants in high linkage disequilibrium with a lead GWAS variant (methods, $r^2 > 0.8$) that overlap a sex-biased peak. Of 774 sex-biased liver peaks, 71 overlapped a GWAS variant (Table 4.6, Figure 4.6). Of the 71 sex-biased peaks linked to a GWAS signal, 48 were female-biased and 23 were male-biased. Of the 71 sex-biased peaks linked to a GWAS signal, 30 overlapped variants for cardiometabolic trait including 3 associated with diabetes, 5 associated with body mass index, 3 associated with liver enzyme levels, and 5 associated with cholesterol. Some of the sex-biased peaks were also associated with less obviously cardiometabolic but potentially relevant traits such as lung function, that has been shown to be decreased in individuals with metabolic syndrome[103,104]. Of the 71 sex-biased peaks linked to a GWAS signal, 20 were also linked to 28 genes by liver eQTL. These 71 peaks spanned an average width of 1281 base pairs each, compared to an average of 998 base pairs each for the full set of 774 sex-biased peaks and 992 base pairs for the full testing set of 172,813 consensus peaks. An increase in average width of peaks overlapping a GWAS variant compared to the average width of sex-biased peaks could indicate increased risk of variants overlapping a peak by chance.

At one sex-biased peak (peak1441) a female-biased peak overlapped variants rs12562207, rs12057175, and rs12057222 which are linked to differential expression of protein kinase receptor *EPHA2*[62] and gamma glutamyl transferase levels[177], an important marker for liver function (Figure 4.7). *EPHA2* has been linked to NAFLD[178,179]. A nearby peak at this locus was also identified as a caQTL in liver tissue[28].

Discussion

Sex differences are known to influence disease risk and drug metabolism[25–27]. Identifying mechanisms behind these sex differences could aid diagnosis and treatment to improve healthcare. In this project, we profiled chromatin accessibility in 139 human liver samples and among 172,813 consensus liver chromatin accessibility regions identified 774 regions of sex-biased chromatin accessibility between males and females. Of the 774 sex-biased regions we identified, 390 were female-biased and 384 were male-biased, suggesting an even representation of sex-biased traits in each direction. We linked these 774 sex-biased chromatin accessibility regions to gene expression using eQTL and/or GWAS traits, including 24 regions linked to both a gene and a trait. Of 71 sex-biased regions linked to GWAS traits, 30 peaks linked to cardiometabolic traits including diabetes, cholesterol, and liver enzyme levels. Sex-biased gene expression has been identified in liver tissue[62] and we observed a sex-biased chromatin accessibility region (peak19490) that overlapped a variant associated with sex-biased expression (Table 4-6). This variant is associated with expression of *HKDC1*, a hexokinase protein with known roles in glucose metabolism[180], and glycemic traits during pregnancy[181]. These sex-biased chromatin accessibility regions are a resource that can guide future studies into the mechanism of relevant cardiometabolic traits in liver.

Some limitations in the current study design can be addressed in future analyses. A larger sample size would increase power to detect sex-biased chromatin accessibility. The subset of liver samples for this study were chosen based on criteria of existing genotype and gene expression data. However, the liver bank includes hundreds of additional tissue samples which could be analyzed. The majority of samples in the current study were also of European ancestry, and analyses did not consider ancestry, so sex-differential peaks that differ by ancestry may have been missed. Samples selected in this study were also biased towards males (67%, Table 4-3), which could limit our ability to detect differences (Table 4-4). Analyses also did not consider differences due to disease status or body mass, for which data was missing for most samples. Additionally, our samples ranged in age from 2-81 and age can play a role in chromatin accessibility[176] through several mechanisms including changes in hormones[25]. Therefore, future studies could include age as a covariate or be performed on a subset of samples from a narrower age range. Accounting for the effect of age could aid identification of sex-biased chromatin accessibility due to hormonal changes. Also, while environmental variables such as age, drug use, disease, cause of death, and hormonal status could affect chromatin accessibility, limited data were available about these liver samples and organ donors used in this study. Due to this limitation, we were not able to adjust for potential environmental sources of variation. We recently obtained some additional data on known variables, such as known drug use which could be used to adjust for or exclude individuals. Finally, liver tissue is heterogenous, and we would have missed cell type-dependent differences in chromatin accessibility.

Although we identified a similar number of sex-biased peaks compared to other studies of sex-biased chromatin accessibility[171], we identified few sex-biased peaks, and our sex-biased peaks demonstrated small differences in LFC between sexes, which suggest weak sex-biased chromatin accessibility differences. Weak identification of sex-biased regions could be due to technical or biological reasons. Technical reasons that could lead to weak identification of sex-biased regions are low power due to small sample size or insufficient sequencing depth. We have additional samples available to increase sample size and additional libraries prepared that could be combined to increase sequencing depth. I produced chromatin accessibility profiles in triplicate for each liver sample but only used one library per sample for these initial analyses. Future studies that combine reads from replicate libraries would improve sequencing depth and power to detect sex-biased regions or other features such as chromatin accessibility QTL[182]. Some biological mechanisms of sex differences in traits include genetic differences due to sex chromosomes, epigenetic differences, differences in gene regulation, differences in environmental exposures, and differences in endogenous factors such as hormones[27]. Sex-biased gene expression has been identified in liver tissue[62] but we observed only one sex-biased chromatin accessibility region (peak19490) that overlapped a variant associated with sex-biased expression. It is possible that our sample set did not include individuals with environmental exposures that cause differences in expression. Another possibility is that the age range and lack of relevant phenotype data that could affect hormones such as pregnancy status for females could have affected our power to detect differences due to hormonal factors.

The liver chromatin accessibility profiles and candidate sex-biased regulatory regions identified in this study will be a useful resource for future studies in regulatory mechanisms of disease in liver. Sex-biased chromatin accessibility regions could be tested for differences in transcription factor binding site enrichment[134], which could help identify mechanisms of sex differences. They can be used to identify regulatory elements that correspond to sex-specific liver eQTL variants[62]. Sex-biased chromatin accessibility could also more thoroughly be linked to genes using additional datasets such as chromosome conformation capture profiles[183]. Linking regulatory elements to candidate genes remains challenging due to distances between noncoding elements and genes[53], therefore linking a gene by multiple methods can increase confidence in the association.

Due to the heterogeneity of liver tissue, these accessible chromatin regions reflect a mixture of liver cell types[184]. Single nucleus chromatin accessibility and gene expression profiling would also allow us to identify cell-type-specific regulatory elements and more generally differentiate between regulatory mechanisms in different cell types within the tissue. I have re-optimized nucleus isolation and we have started single nucleus multiomic chromatin accessibility and gene expression profiling on a subset of 40 samples from the 139 samples described in this study. Analyses of these data may identify additional cell-type-specific sex differences. In addition, the single nucleus data may prove useful as a reference for deconvolution of cell type in other bulk liver tissue chromatin and gene expression studies.

Overall, these analyses show some promising initial results. They identify hundreds of sex-biased regions that may help explain regulatory effects on gene expression. Future analyses of these sex-biased regions could focus on genes that have shown sex-biased gene expression or are known to be involved in response to sex hormones. Furthermore, these libraries can be used to identify genetic variants that influence chromatin accessibility in a larger sample size than previous studies[28]. These liver chromatin accessibility profiles will be a valuable resource for future studies on gene regulation in liver.

Methods

Liver tissue:

Human liver tissue was collected as previously described[28]. Briefly tissue was collected from deceased organ donors without known disease through the National Institutes of Health Liver Tissue Cell Distribution System (LTCDS). Tissue was obtained from LTCDS and approved for use in this study as non-human subjects research by the Institutional Review Boards (IRBs) at St Jude Children's Research Hospital (Memphis, TN) and the University of North Carolina (Chapel Hill, NC). Tissue was flash frozen and stored at -80°C until use.

Ethnicity for samples was reported as "Black" or "White" at time of sample collection. "Black" is here reported as African Ancestry (AFR) and "White" as European Ancestry (EUR).

Nuclei isolation:

We isolated human liver tissue nuclei as previously described[28]. All steps were performed on ice unless otherwise stated. Briefly, we crushed 50-mg pieces of frozen sample in liquid nitrogen using a Cell Crusher (CellCrusher, Cork Island), homogenized the sample in a 1 mL dounce for 40 strokes in nuclei isolation buffer (NIB: 20 mM Tris-HCl, 50 mM EDTA, 5 mM spermidine, 0.15 mM spermine, 0.1% mercaptoethanol, 40% glycerol, pH 7.5) and rotated for 5 minutes at 4°C. We filtered the solution through a Miracloth (Calbiochem, San Diego, Ca USA), centrifuged at 1100g for 10 minutes at 4°C, resuspended the pellet in 250 uL NIB containing 0.5% Triton-X, centrifuged at 500g for 5 minutes at 4°C, and finally resuspended the pellet in 250 uL of resuspension buffer (10 mM Tris-HCl, 10 mM NaCl, 3 mM MgCl2, pH 7.4). We quantified nuclei concentration using a cell countess to aliquot 50,000 nuclei for each library preparation.

ATAC-seq library preparation:

We profiled chromatin accessibility as previously described[28] following the ATAC-seq protocol[69]. An ATAC-seq library was prepared in triplicate or quadruplicate for each nuclei isolation prep for a sample. Briefly, we used Nextera (Illumina) kits with 5uL of Tn5 per library and unique, dual-barcoded indices. We cleaned the Tn5 transposase reaction and final library after PCR with Zymo DNA Clean and Concentrator (D4029). We visualized and quantified libraries using TapeStation, and sequenced with paired-end reads on a Novaseq.

ATAC-seq read alignment and peak calling:

We aligned ATAC-seq reads and called peaks as previously described[83]. Briefly, we trimmed sequencing adapters using cutadapt[123]. We aligned trimmed reads to the hg19 human genome[124] using bowtie2[125] and selected nuclear chromosomal alignments with mapq>20 using samtools[125]. We removed alignments overlapping high-signal regions (Duke excluded and ENCODE/DAC exclusion list regions)[126] using BEDTools pairToBed[127]. We removed duplicate alignments using Picard MarkDuplicates (https://github.com/broadinstitute/picard) and generated ATAC-seq quality metrics using ataqv.[128] After filtering libraries retained 54.6 million reads on average (Table 4.2). We trimmed alignments so their 5' ends corresponded to the Tn5 binding site (+4 for + strand alignments and -5 for – strand alignments)[69] and smoothed signal by extending alignments 100 bp on either side of the Tn5 binding sites using BEDTools slop[127].

We called peaks (FDR<5%) with MACS2[129] and generated ATAC-seq signal bigwig files from MACS2 bedGraph files using the bedGraphToBigWig tool from ucsctools[130]. We verified that ATAC-seq libraries matched genotypes with verifyBamID[174] and verified that sex based on genotype data matched reported sex with PLINK sex check[142]. We proceeded with analysis on libraries that had TSS enrichment >= 4 and percent reads in peaks >= 10. The best replicate, determined by highest TSS enrichment, was used for downstream analyses.

Identification of liver consensus and sex-biased peaks:

We generated a set of consensus liver peaks by merging peak genomic coordinates across libraries using BEDTools merge[127]. We defined consensus peaks as merged peaks that overlapped peaks in 5% or more of individual liver samples (at least 7 out of 139 samples). We used all of the consensus peaks to test for sex-biased peaks. We quantified accessibility of consensus peaks using featureCounts[131]. We tested for differential chromatin accessibility using DESeq2[133] and defined peaks with FDR < 5% and log fold change (LFC) > 0 as differentially accessible or sex-biased.

Identification of genes linked to sex-biased peaks:

We linked sex-biased chromatin accessibility to genes using overlap with liver eQTL variants. We identified sex-biased peaks that overlapped eQTL proxy variants (r^2>0.8 with the eQTL lead, 1000G phase 3 EUR LD calculated using SniPA[185]) using previously published liver (n = 2.3 million) and sex-biased liver (n = 1,683) eQTL[62] using BedTools[127]. We listed all eQTL variants that intersected a peak.

Overlap of GWAS signals with sex-biased peaks:

We performed overlap of GWAS signals with sex-biased peaks as previously described[83]. Briefly, we downloaded the NHGRI-EBI GWAS catalog[44] on January 17, 2020 and lifted variant positions from hg38 to hg19 using pyliftover (https://github.com/konstantint/pyliftover), a python implementation of the UCSC liftOver tool[148]. We performed LD-clumping using swiss (https://github.com/statgen/swiss)[61]. We identified sex-biased peaks that overlapped GWAS proxy variants (LD r^2>0.8 with the signal lead variant, 1000G phase 3 EUR, calculated with PLINK v1.9[142]) using BedTools[127].

Acknowledgements

Thank you to everyone who contributed to the work in this chapter, including Michael Erdos for guidance in optimizing our nuclei isolation protocol, Amy Etheridge for assistance in

sample acquisition, Swarooparani Vadlamudi for assistance in ATAC-seq library preparation, Shelley Moxley for assistance with single nucleus preps, Kevin Currin for guidance with data analysis, and Karen Mohlke for advising on study design, analyses, and interpretation.

Figures

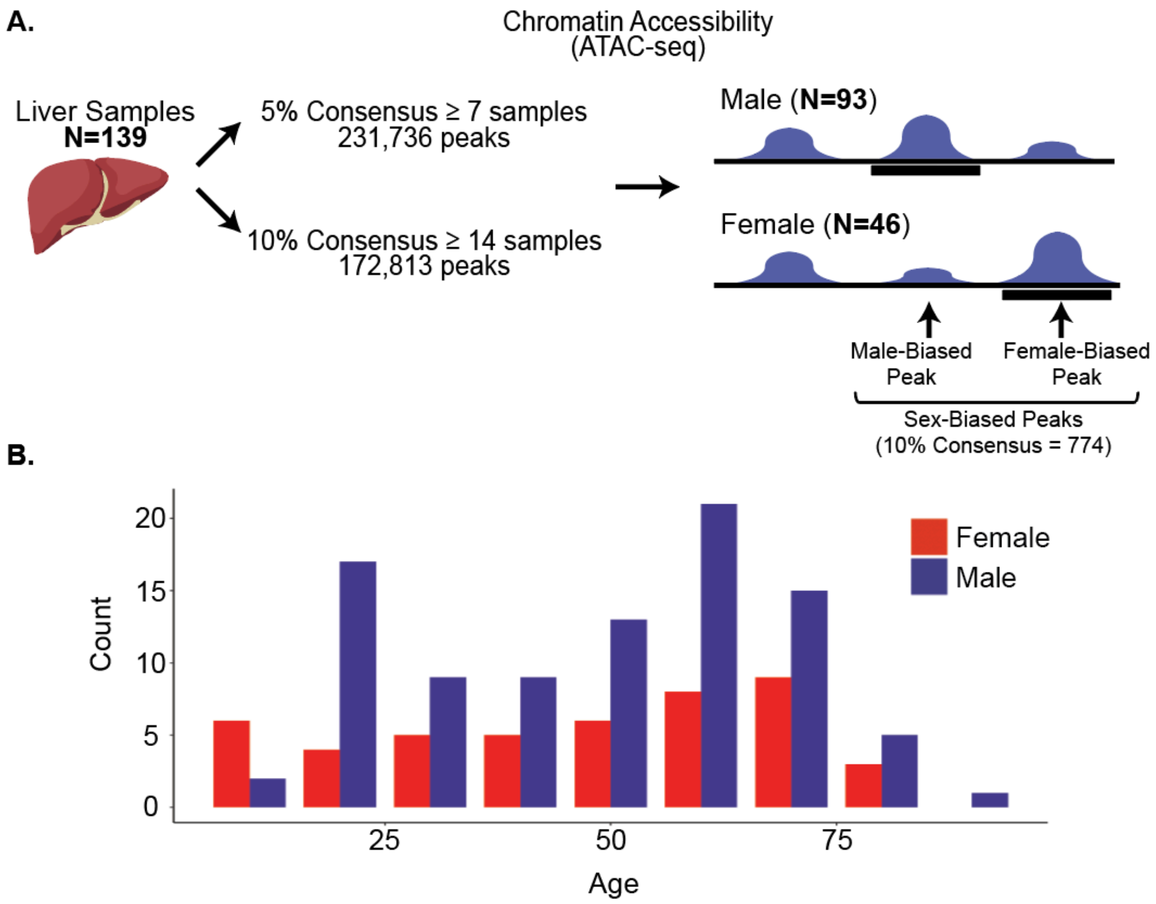


Figure 4.1: Genome-wide profiles of chromatin accessibility in human liver samples to identify sex-biased peaks.

A. Schematic of experimental design. Human liver samples were obtained and ATAC-seq profiles were generated. Consensus peaks (blue) were called in two sets; the first for peaks present in 5% of total samples and a second for peaks present in 10% of total samples. We identified sex-biased peaks (black bars) as differential peaks between male and female samples (DeSEQ2, LFC > 0, FDR < 5%). Male-biased peaks were more accessible in males and female-biased peaks were more accessible in females. B. Histogram plotting distribution of ages for 139 liver samples. The average age was 43 with a range of 2 to 81. Females are indicated in red and males are indicated in blue.

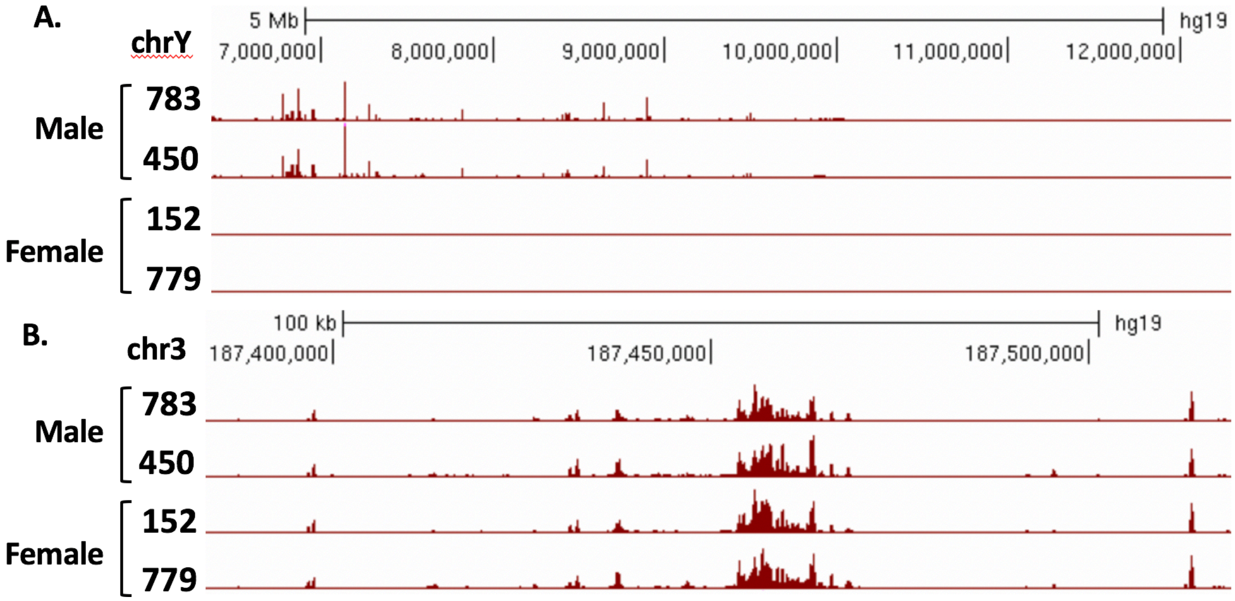


Figure 4.2: Peaks in sex chromosomes correspond with reported sex.

A. Chromatin accessibility profiles for two male samples and two female samples on the y chromosome, showing lack of signal for the female samples. B. Chromatin accessibility profiles for two male samples and two female samples on an autosome, showing comparable signal between sexes.

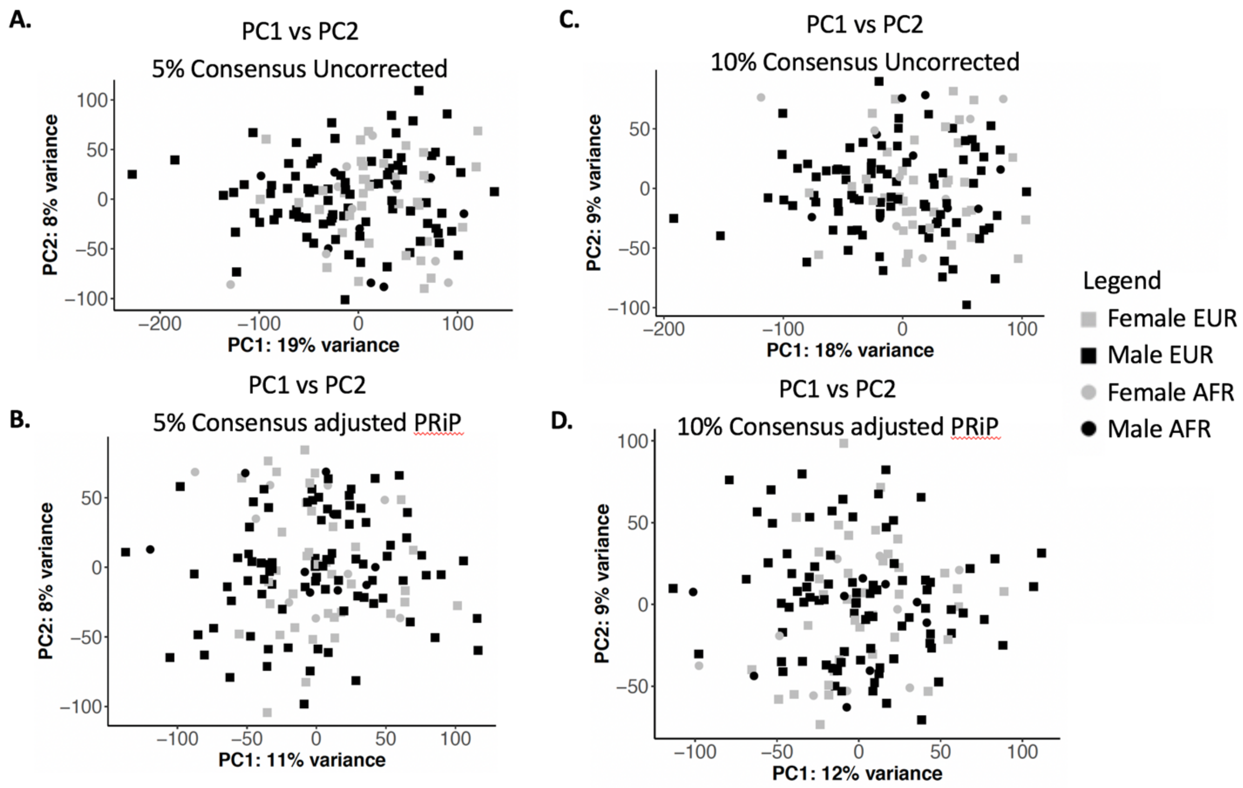


Figure 4.3: PCA of ATAC-seq read count within peaks for 139 liver samples.

A-D. Plot of PCA of ATAC-seq read counts within peaks for two sets of consensus peaks showing variance between female (grey) and male (black) samples. European ancestry samples (EUR) are represented by squares and African ancestry (AFR) samples are represented by circles. A. PCA for unadjusted read counts for the 5% consensus peak set. B. PCA for read counts within peaks adjusted for percent reads in peaks for the 5% consensus peak set. C. PCA for unadjusted read counts for the 10% consensus peak set. D. PCA for read counts adjusted for percent reads in peaks for the 10% consensus peak set. While there is not a clear separation by sex, adjusting for percent reads in peaks reduced variance of PC1.

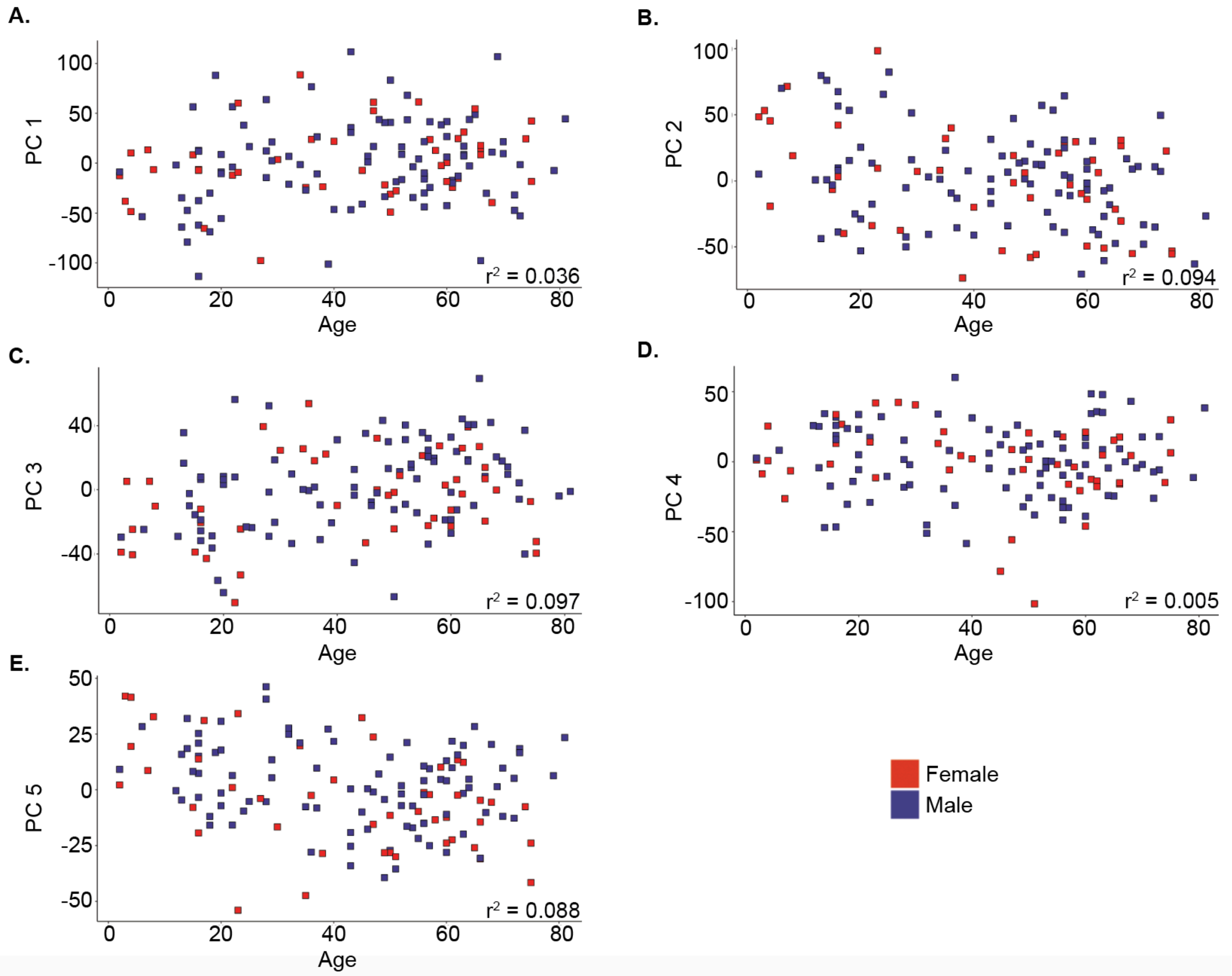


Figure 4.4: Distribution of ages for 139 liver samples and correlation with first five principal components.

B-F. Plot of age compared to principal components one through five with the Pearson's correlation (r^2) with age for each. Females are indicated in red and males are indicated in blue.

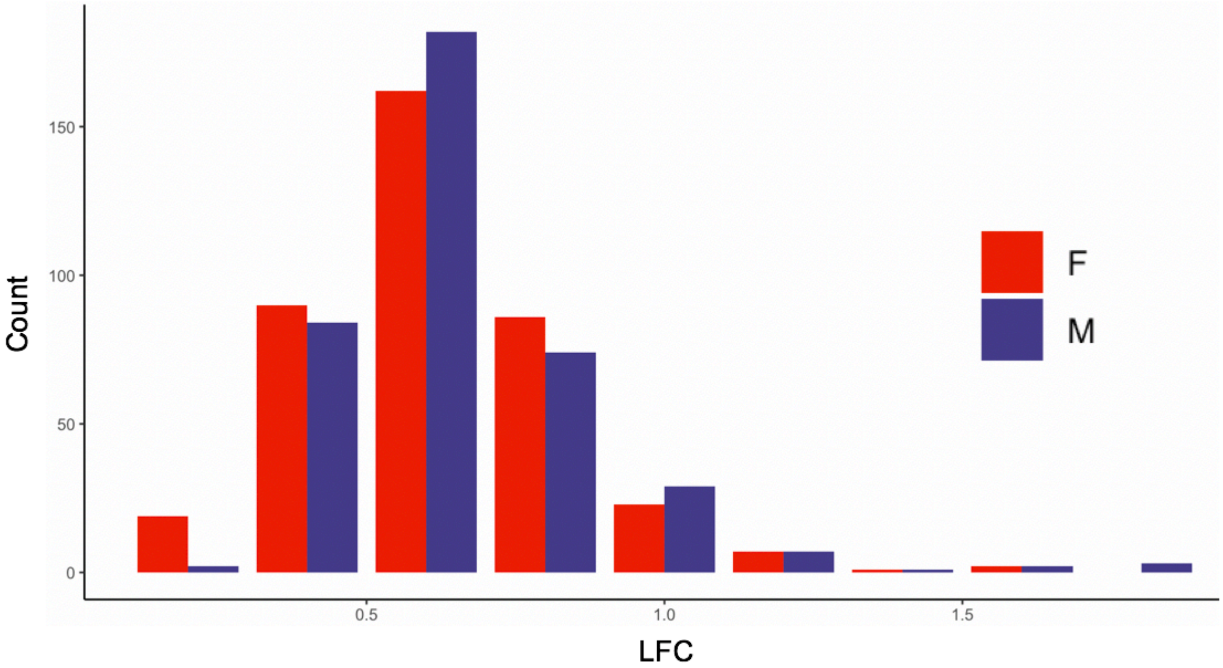


Figure 4.5: Distribution of LFC for female-biased and male-biased peaks.

Histogram of LFC showing the distribution of the 774 sex-biased peaks from the 10% consensus peak set adjusted for percent reads in peaks for 93 male and 46 female samples. Female-biased results “F” are shown in red and male-biased results “M” are shown in blue.

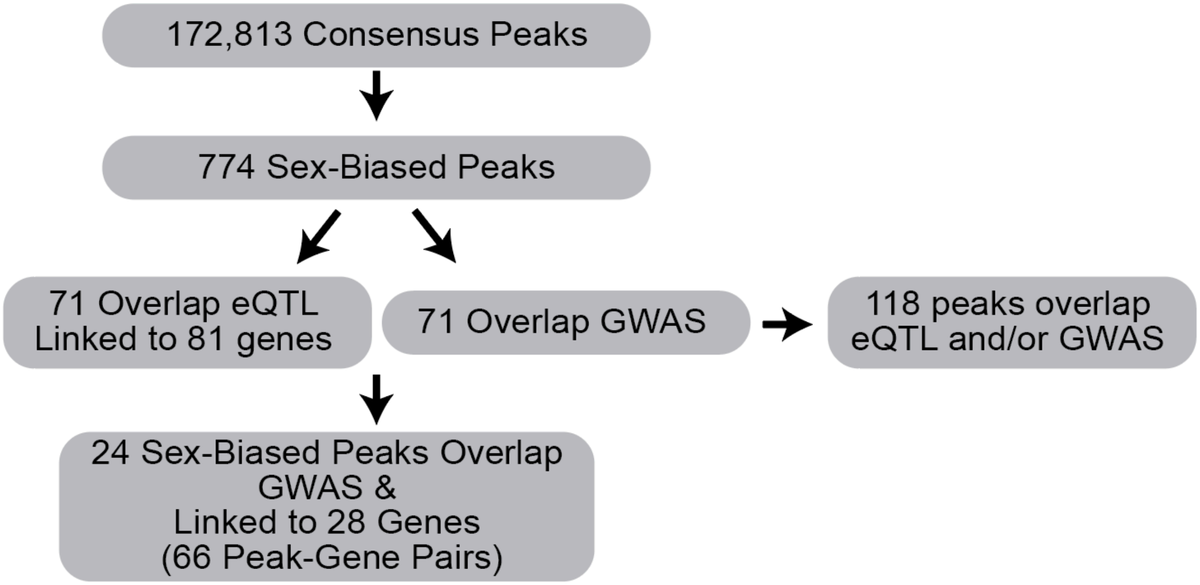

172,813 Consensus Peaks
774 Sex-Biased Peaks
71 Overlap eQTL Linked to 81 genes
71 Overlap GWAS
118 peaks overlap eQTL and/or GWAS
24 Sex-Biased Peaks Overlap GWAS & Linked to 28 Genes (66 Peak-Gene Pairs)

Figure 4.6: Summary of sex-biased liver peaks linked to GWAS traits and differential gene expression.

Flowchart identifying sex-biased peaks overlapping GWAS signals and linked to genes through overlap with liver eQTL signals for the 10% consensus peak set.

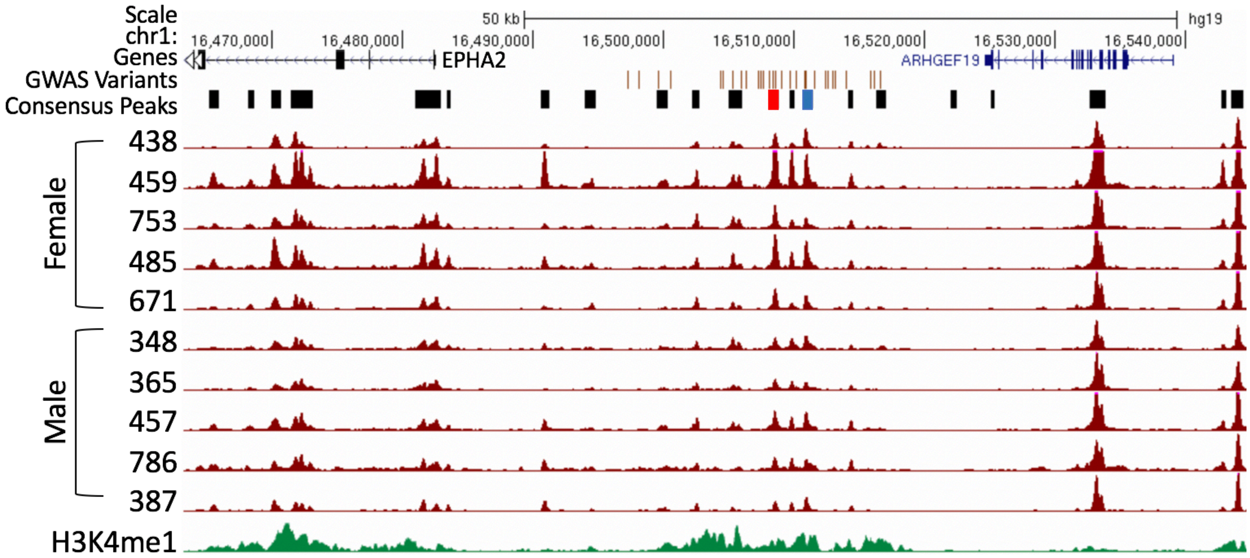


Figure 4.7: Sex-biased peak identified with variant associated with expression of *EPHA2* and levels of gamma glutamyl transferase levels.

Variants rs12562207, rs12057175, and rs12057222 overlap female-biased peak1441 (red) and are linked to expression of protein kinase receptor *EPHA2* and gamma glutamyl transferase levels, an important marker for liver function. The blue consensus peak indicates a peak identified as a chromatin accessibility QTL. H3K4me1 adult liver histone modifications (green) from the Roadmap Epigenomics Consortium[77].

Tables

	Male	Female	Total
EUR	84	37	121
AFR	9	9	18
Total	93	46	139

Table 4-1: Sex and ethnicity demographics for 139 liver samples.

Counts of samples by sex and ancestry of the 139 liver samples donor. “AFR” indicates African ancestry, “EUR” indicates European ancestry.

	Final Reads (million)	Peaks	% Reads in Peaks	TSS Enrichment	BMI*
Average	54.6	109,295	30.0	8.0	28
Std Dev	±23.5	± 24,727	± 9.1	± 2.4	± 7.3
Minimum	3.4	38,442	10.6	4.0	13.1
Maximum	132.9	204,881	49.6	16.4	62.9

Table 4-2: Summary of ATAC-seq library and sample metrics for 139 samples. Selected metrics summarizing 139 ATAC-seq libraries including the average, standard deviation (Std Dev), minimum value, and maximum value. Final reads are the total reads used for peak calling after quality filtering as described in the methods. *BMI metrics are calculated from 108 samples with a reported height and weight.

Sample ID	Reads After Filtering	Peaks	% Reads in Peaks	TSS Enrichment	Ethnicity*	Sex	Age	BMI
1	83,110,670	143,225	49.6	9.6	AFR	M	2	NA
39	3,659,042	45,628	26.3	16.4	EUR	M	56	NA
122	35,316,708	81,426	14.4	6.3	EUR	M	67	NA
151	27,317,546	70,865	24.5	6.4	AFR	M	32	23.6
152	61,118,334	115,926	35.8	7.3	EUR	F	40	29.7
156	47,993,186	93,845	33.0	5.7	EUR	M	20	NA
162	46,322,516	128,372	32.9	10.6	EUR	M	18	NA
172	40,462,484	100,430	30.6	8.3	EUR	M	16	22.9
174	60,712,830	155,400	44.7	11.6	EUR	F	23	19.1
175	92,798,678	129,402	46.6	9.2	EUR	M	32	NA
177	36,506,984	104,321	45.8	14.9	AFR	M	39	NA
200	19,007,920	92,879	33.9	11.0	EUR	F	2	NA
201	35,895,952	116,545	41.2	15.0	EUR	M	6	14.7
204	42,132,022	84,934	26.6	8.5	EUR	F	8	13.1
217	90,646,340	144,347	44.5	7.2	EUR	M	20	26.5
221	41,361,276	101,750	26.2	6.9	EUR	M	16	25.0
223	34,396,472	101,259	30.5	10.0	EUR	M	14	20.0
238	57,252,620	120,800	47.0	10.7	AFR	F	4	NA
253	50,314,904	106,720	10.6	7.2	AFR	F	45	29.9
323	79,198,844	119,175	16.6	6.0	EUR	M	43	41.1
325	22,614,520	78,468	19.2	5.8	EUR	M	60	22.8
331	34,464,376	80,345	27.0	6.7	EUR	F	62	26.4
332	63,360,300	115,477	32.7	6.0	EUR	F	65	30.0
333	46,138,582	98,274	34.8	6.7	EUR	M	59	22.6
334	84,483,840	126,616	36.9	5.7	EUR	M	63	34.7
335	57,827,568	111,423	32.6	5.5	EUR	M	36	28.3
336	57,705,794	89,233	18.5	5.2	EUR	M	70	37.9
337	34,809,246	75,511	15.3	5.2	EUR	M	34	30.8
340	28,935,678	101,961	25.5	10.8	EUR	M	52	32.3
342	23,837,186	89,849	29.4	6.2	EUR	M	43	30.7
343	36,775,276	88,069	14.8	8.9	EUR	M	35	20.8
344	62,041,426	136,413	37.6	7.1	EUR	M	63	32.2
345	28,652,532	84,098	33.3	8.6	EUR	M	60	34.4
346	33,009,090	106,232	41.3	9.0	EUR	M	24	29.8
347	61,150,678	116,838	36.4	9.4	EUR	F	4	15.3
348	70,080,502	119,641	30.2	6.8	EUR	M	43	22.5

350	82,638,414	138,344	38.2	7.6	EUR	F	75	29.5
351	37,103,670	109,534	22.1	9.2	EUR	M	49	31.3
352	81,592,510	137,131	38.6	10.4	EUR	M	72	27.6
356	22,682,876	74,796	22.7	6.2	EUR	M	37	22.4
357	132,882,982	166,751	44.8	8.2	EUR	M	62	36.5
358	41,839,902	101,712	30.5	6.9	EUR	M	53	30.9
360	53,392,272	109,214	30.1	8.9	EUR	M	54	28.3
363	56,303,886	105,272	28.1	6.4	EUR	M	46	NA
365	66,023,240	108,552	36.3	5.3	EUR	M	28	23.3
366	51,754,096	114,465	36.6	7.7	EUR	F	60	32.0
368	44,607,564	90,015	33.7	5.6	EUR	F	23	22.1
369	51,491,458	98,725	25.9	6.9	EUR	F	66	27.5
372	48,285,422	55,946	10.8	4.0	EUR	M	37	28.5
374	3,379,040	38,442	16.4	9.5	EUR	M	72	26.9
378	75,803,202	133,963	46.3	6.6	EUR	M	81	24.3
381	29,042,412	92,597	12.3	7.7	EUR	M	14	16.4
382	45,904,450	123,506	47.6	11.3	EUR	M	56	25.8
383	67,613,438	141,478	31.9	9.3	EUR	M	61	NA
387	31,334,274	123,808	37.7	13.6	EUR	M	66	NA
390	69,212,094	119,608	21.4	6.3	EUR	F	59	NA
399	56,580,292	100,640	26.2	6.6	EUR	F	22	NA
401	106,697,692	133,718	41.4	6.9	EUR	F	38	NA
403	28,404,136	83,751	21.1	6.6	EUR	M	73	20.6
414	62,360,206	89,341	14.4	7.5	EUR	M	29	13.5
418	44,828,962	131,689	41.3	8.5	EUR	F	16	22.7
421	103,135,120	143,810	11.2	6.5	EUR	M	18	21.7
431	18,985,096	63,207	17.5	6.1	EUR	M	56	NA
433	45,933,202	107,404	25.9	8.2	EUR	M	46	26.1
434	31,829,762	80,248	28.1	6.8	EUR	M	64	16.8
435	68,748,300	135,056	31.6	11.0	AFR	F	58	34.4
436	42,572,884	123,802	27.3	12.2	EUR	F	49	39.0
437	47,958,298	102,377	30.2	9.6	EUR	F	62	26.2
438	41,912,874	107,956	33.8	10.7	EUR	F	7	15.9
439	58,914,314	117,278	32.1	7.2	EUR	M	48	36.1
440	34,128,294	78,069	15.5	6.9	AFR	M	29	42.2
444	61,146,302	78,739	12.7	5.6	EUR	M	12	18.1
450	76,951,990	140,264	34.8	8.2	EUR	M	40	35.4
457	63,917,430	119,942	36.3	7.8	AFR	M	13	34.4
458	49,174,114	101,455	18.6	6.6	AFR	F	27	NA

459	114,019,114	161,133	45.7	9.7	EUR	F	17	22.7
465	34,431,392	95,559	18.4	9.8	EUR	F	61	33.1
469	82,834,156	112,936	36.7	6.1	EUR	M	28	26.6
470	51,107,544	102,920	25.1	7.4	EUR	F	68	29.1
476	72,139,166	135,136	31.8	8.2	EUR	F	66	30.3
479	50,608,338	106,898	33.5	8.6	AFR	F	57	22.0
480	82,639,252	117,893	44.9	7.9	AFR	F	51	23.3
484	76,648,838	122,164	40.5	6.8	EUR	F	34	32.9
485	74,296,956	146,982	43.9	10.6	EUR	F	50	27.5
489	105,903,992	115,920	21.4	5.1	EUR	M	28	24.4
492	57,221,890	90,513	20.9	5.3	AFR	F	63	23.3
493	73,027,038	122,655	19.3	6.7	EUR	M	63	25.8
617	53,948,340	120,890	35.9	6.5	EUR	M	16	23.0
618	20,483,848	90,432	34.8	14.1	EUR	M	56	25.1
619	33,611,144	80,894	21.3	5.9	EUR	M	55	31.2
620	60,269,378	115,560	38.6	7.6	AFR	M	43	15.5
623	34,551,396	121,255	34.6	12.0	EUR	M	13	32.2
627	31,016,478	92,507	28.9	8.0	EUR	F	74	22.7
629	28,399,842	91,515	30.5	7.8	EUR	F	50	NA
630	45,074,214	131,871	33.4	10.6	EUR	M	45	34.8
631	37,282,718	87,286	31.1	5.6	EUR	M	65	27.8
632	47,378,534	119,922	29.4	9.1	EUR	M	54	26.8
633	27,736,278	73,166	22.8	5.6	EUR	M	15	NA
634	48,409,990	98,097	20.1	7.8	EUR	M	60	25.8
636	44,239,748	103,233	23.8	6.5	EUR	M	47	33.4
639	63,864,848	124,149	35.2	8.4	EUR	M	60	26.1
644	54,209,058	107,970	46.5	12.2	EUR	F	60	NA
649	31,203,188	104,351	27.9	5.6	EUR	M	22	NA
657	63,241,484	123,262	36.1	9.7	EUR	F	15	NA
659	36,366,080	97,192	26.1	7.8	AFR	M	52	24.5
662	57,174,968	145,129	38.7	12.1	EUR	F	35	40.5
669	42,854,316	107,642	41.1	10.3	EUR	F	66	NA
671	58,514,812	128,182	35.5	11.2	EUR	F	3	NA
687	35,389,742	97,656	27.4	8.5	EUR	M	52	NA
711	74,307,346	153,560	38.0	10.2	EUR	M	73	NA
713	64,835,824	90,023	18.4	5.1	EUR	M	20	26.6
720	35,124,838	110,430	32.8	9.2	EUR	M	51	29.5
724	95,761,940	151,693	27.4	6.4	EUR	M	61	27.4
730	91,558,200	149,470	26.3	8.1	EUR	M	57	33.4

733	67,290,698	109,691	34.9	6.2	EUR	M	49	38.0
741	114,020,806	138,741	28.7	5.8	EUR	M	68	22.1
744	30,070,406	96,191	22.9	6.6	EUR	M	22	22.0
750	68,163,608	105,316	29.2	7.2	EUR	M	50	23.9
751	46,770,746	86,383	26.8	4.5	EUR	M	19	27.1
753	82,061,212	134,963	40.7	9.6	EUR	F	16	NA
755	65,233,570	89,688	19.6	4.6	EUR	F	75	22.5
765	29,763,754	88,226	32.7	6.0	EUR	M	69	28.1
767	45,660,222	118,355	32.8	7.5	EUR	F	30	40.2
770	72,558,668	103,002	20.6	5.3	AFR	M	79	26.5
773	86,253,288	101,096	41.7	6.0	EUR	F	47	NA
775	36,679,348	81,085	24.9	5.7	EUR	M	53	23.7
778	27,404,944	97,425	42.0	15.8	EUR	M	56	NA
779	87,406,220	111,478	34.3	6.5	EUR	F	47	62.9
780	76,582,292	122,535	39.1	7.0	EUR	M	70	33.0
783	34,996,376	109,249	31.1	12.1	EUR	M	25	23.0
786	89,681,118	204,881	28.0	8.2	EUR	M	16	23.0
791	83,122,894	120,656	28.1	6.6	AFR	F	55	26.6
793	24,780,878	61,291	15.9	7.4	EUR	F	36	40.4
794	48,280,568	143,965	31.3	11.4	EUR	M	16	39.0
795	50,035,908	84,827	26.8	5.0	EUR	M	50	NA
796	79,771,748	102,458	30.4	6.5	AFR	M	57	NA
798	41,205,566	97,062	35.8	6.9	EUR	M	64	NA
800	80,848,314	127,812	31.1	8.2	AFR	F	56	NA

Table 4-3: ATAC-seq library metrics for 139 liver samples.

ATAC-seq libraries of 139 human liver samples used in these analyses with sequencing and alignment metrics and sex, ethnicity, age, and BMI where known. In the Sex column "M" indicates a male sample and "F" indicated a female sample. In the BMI column "NA" indicates height and weight were not reported at sample collection.

% Consensus	Model	Male-Biased	Female-Biased	Total Sex-Biased
5	Unadjusted	1,396	1,379	2,775
	Adjusted for PRiP	361	380	741
10	Unadjusted	1,386	1,391	2,777
	Adjusted for PRiP	384	390	774

Table 4-4: Summary of identified sex-biased liver peaks.

Counts summarizing sex-biased peaks (DeSEQ2, LFC > 0, FDR < 5%), separated by male-biased and female-biased for the liberal and stringent consensus peak sets that required a peak to be present in 5% (n>7) or 10% (n>14) individuals, respectively. "PrIP" indicates percent reads in peaks, a quality metric.

Chromosome #	Count
chr1	58
chr2	69
chr3	56
chr4	56
chr5	47
chr6	50
chr7	43
chr8	42
chr9	29
chr10	34
chr11	35
chr12	45
chr13	23
chr14	35
chr15	22
chr16	23
chr17	12
chr18	37
chr19	15
chr20	23
chr21	11
chr22	9

Table 4-5: Summary of counts of sex-biased liver peaks by chromosome.

Counts summarizing sex-biased peaks, separated by chromosome. “Chromosome #” indicates the chromosome being counted. “Count” indicates the total number of significant peaks located on that chromosome.

Sex-Biased Peak ID	Chr	Start position	Stop position	LFC	Sex-Bias	Q-Value	Proxy rsID	r^2	Allele	Gene	GWAS Overlap
16719	10	18971524	18974772	0.71	F>M	4E-03	rs10741132	0.97	A/G	*ARL5B*	yes
16719	10	18971524	18974772	0.71	F>M	4E-03	rs4748509	0.95	G/A	*ARL5B*	yes
16719	10	18971524	18974772	0.71	F>M	4E-03	rs10829213	0.97	T/G	*ARL5B*	yes
16719	10	18971524	18974772	0.71	F>M	4E-03	rs10764691	0.97	A/G	*ARL5B*	yes
16719	10	18971524	18974772	0.71	F>M	4E-03	rs7909570	0.97	T/C	*ARL5B*	yes
16719	10	18971524	18974772	0.71	F>M	4E-03	rs7909845	0.97	T/A	*ARL5B*	yes
16719	10	18971524	18974772	0.71	F>M	4E-03	rs10741131	0.82	T/C	*CACNB2*	yes
16719	10	18971524	18974772	0.71	F>M	4E-03	rs10741131	0.89	T/C	*NSUN6*	yes
16719	10	18971524	18974772	0.71	F>M	4E-03	rs7084709	0.82	C/T	*NSUN6*	yes
52125	14	102543276	102546295	0.68	F>M	4E-02	rs7153402	0.94	C/T	*DYNC1H1*	yes
52125	14	102543276	102546295	0.68	F>M	4E-02	rs56198816	0.94	T/G	*DYNC1H1*	yes
52125	14	102543276	102546295	0.68	F>M	4E-02	rs7152877	0.81	G/A	*DYNC1H1*	yes
19490*	10	70979797	70980602	0.67	F>M	2E-03	rs10823318	0.96	T/A	*HKDC1*	yes
19490*	10	70979797	70980602	0.67	F>M	2E-03	rs10823318	0.96	T/A	*HKDC1*	yes
8603	1	150544095	150545106	0.67	F>M	2E-02	rs7550234	0.95	G/A	*PRUNE*	yes
8603	1	150544095	150545106	0.67	F>M	2E-02	rs112564154	0.90	A/G	*PRUNE*	yes
8602	1	150539762	150542939	0.65	F>M	1E-05	rs7549723	0.82	T/C	*ADAMTSL4*	yes
8602	1	150539762	150542939	0.65	F>M	1E-05	rs6655975	0.83	G/A	*ADAMTSL4*	yes
8602	1	150539762	150542939	0.65	F>M	1E-05	rs4971044	0.95	A/G	*PRUNE*	yes

85867	2	100916062	100916411	0.64	F>M	2E-02	rs11123823	0.83	A/G	*LONRF2*	yes
168994	9	95004259	95005536	0.47	F>M	2E-02	rs72750433	0.96	G/A	*IARS*	yes
168994	9	95004259	95005536	0.47	F>M	2E-02	rs10125474	0.88	C/T	*OGN*	yes
19491*	10	70981671	70983070	0.46	F>M	3E-02	rs4746822	1.00	T/C	*HKDC1*	yes
19491*	10	70981671	70983070	0.46	F>M	3E-02	rs4746822	1.00	T/C	*HKDC1*	yes
103629	22	41856938	41857969	0.46	F>M	4E-02	rs5751107	0.82	G/A	*NHP2L1*	yes
103629	22	41856938	41857969	0.46	F>M	4E-02	rs5751107	0.97	G/A	*TOB2*	yes
97740	20	47314202	47315371	0.44	F>M	2E-02	rs1040559	0.81	A/G	*PREX1*	yes
97740	20	47314202	47315371	0.44	F>M	2E-02	rs761274	0.81	T/C	*PREX1*	yes
28958	11	68657459	68659541	0.41	F>M	3E-02	rs514833	1.00	C/T	*IGHMBP2*	yes
28958	11	68657459	68659541	0.41	F>M	3E-02	rs488363	1.00	C/G	*IGHMBP2*	yes
28958	11	68657459	68659541	0.41	F>M	3E-02	rs660614	1.00	G/A	*IGHMBP2*	yes
28958	11	68657459	68659541	0.41	F>M	3E-02	rs514833	0.95	C/T	*MRPL21*	yes
28958	11	68657459	68659541	0.41	F>M	3E-02	rs488363	0.95	C/G	*MRPL21*	yes
28958	11	68657459	68659541	0.41	F>M	3E-02	rs660614	0.95	G/A	*MRPL21*	yes
28958	11	68657459	68659541	0.41	F>M	3E-02	rs514833	0.99	C/T	*MTL5*	yes
28958	11	68657459	68659541	0.41	F>M	3E-02	rs488363	0.99	C/G	*MTL5*	yes
28958	11	68657459	68659541	0.41	F>M	3E-02	rs660614	0.99	G/A	*MTL5*	yes
85866	2	100907449	100909015	0.37	F>M	4E-02	rs764828	0.83	A/C	*LONRF2*	yes
1441	1	16508024	16508884	0.36	F>M	4E-02	rs12562207	0.82	A/G	*EPHA2*	yes
1441	1	16508024	16508884	0.36	F>M	4E-02	rs12057175	0.82	G/A	*EPHA2*	yes
1441	1	16508024	16508884	0.36	F>M	4E-02	rs12057222	0.82	G/A	*EPHA2*	yes
105871	3	17794733	17795621	0.35	F>M	5E-02	rs13079096	0.89	A/T	*TBC1D5*	yes

138463	6	28129157	28130003	0.32	F>M	7E-03	rs9357065	0.90	T/C	*ZNF165*	yes
138463	6	28129157	28130003	0.32	F>M	7E-03	rs9357066	0.90	T/C	*ZNF165*	yes
84694	2	73963871	73965175	0.19	F>M	5E-02	rs11894953	0.84	T/C	*SEMA4F*	yes
84694	2	73963871	73965175	0.19	F>M	5E-02	rs17350188	0.84	C/G	*SEMA4F*	yes
84694	2	73963871	73965175	0.19	F>M	5E-02	rs11894953	0.98	T/C	*TPRKB*	yes
84694	2	73963871	73965175	0.19	F>M	5E-02	rs17350188	0.97	C/G	*TPRKB*	yes
151111	7	66056407	66057775	0.17	F>M	1E-03	rs1968127	1.00	C/T	*RABGEF1*	yes
151111	7	66056407	66057775	0.17	F>M	1E-03	rs6977632	0.81	A/G	*RABGEF1*	yes
37710	12	69978944	69980012	0.16	F>M	5E-02	rs11177730	0.96	T/C	*CCT2*	yes
37710	12	69978944	69980012	0.16	F>M	5E-02	rs11177731	0.91	C/T	*CCT2*	yes
147195	7	1095631	1096472	0.24	M>F	2E-02	rs2363279	0.80	G/A	*C7orf50*	yes
3825	1	44015530	44016175	0.30	M>F	4E-02	rs2819336	0.93	C/T	*PTPRF*	yes
3825	1	44015530	44016175	0.30	M>F	4E-02	rs11580258	0.95	G/A	*PTPRF*	yes
166965	9	33828335	33828969	0.37	M>F	4E-02	rs12376951	0.90	G/A	*UBAP2*	yes
166965	9	33828335	33828969	0.37	M>F	4E-02	rs12376951	0.82	G/A	*UBE2R2*	yes
41097	12	123617506	123618546	0.49	M>F	3E-03	rs1790116	0.84	T/G	*C12orf65*	yes
41097	12	123617506	123618546	0.49	M>F	3E-03	rs1790116	0.86	T/G	*CDK2AP1*	yes
41097	12	123617506	123618546	0.49	M>F	3E-03	rs1790116	0.86	T/G	*SBNO1*	yes
96265	20	25260512	25261441	0.51	M>F	6E-04	rs2261720	0.99	T/G	*ABHD12*	yes
96265	20	25260512	25261441	0.51	M>F	6E-04	rs2227890	0.99	A/G	*ABHD12*	yes
96265	20	25260512	25261441	0.51	M>F	6E-04	rs2261720	0.96	T/G	*RP4-691N24.1*	yes

96265	20	25260512	25261441	0.51	M>F	6E-04	rs2227890	0.96	A/G	*RP4-691N24.1*	yes
111383	3	122162543	122164131	0.64	M>F	1E-03	rs4677952	0.99	T/C	*WDR5B*	yes
111383	3	122162543	122164131	0.64	M>F	1E-03	rs4677952	0.99	T/C	*WDR5B*	yes
133430	5	137805073	137807120	1.03	F>M	3E-03	rs34885420	0.92	C/T	*CDC25C*	no
61845	16	68368819	68369741	0.74	F>M	2E-04	rs3785113	0.94	T/C	*PRMT7*	no
48563	14	51495552	51496661	0.66	F>M	2E-02	rs73288876	0.91	A/G	*TRIM9*	no
48563	14	51495552	51496661	0.66	F>M	2E-02	rs6572720	1.00	T/C	*TRIM9*	no
48563	14	51495552	51496661	0.66	F>M	2E-02	rs61093844	1.00	C/T	*TRIM9*	no
72742	18	48000332	48000985	0.59	F>M	6E-04	rs2969978	0.84	A/G	*MAPK4*	no
107043	3	39178064	39180991	0.46	F>M	5E-02	rs784495	0.87	T/C	*VILL*	no
107043	3	39178064	39180991	0.46	F>M	5E-02	rs784496	0.82	G/A	*VILL*	no
107043	3	39178064	39180991	0.46	F>M	5E-02	rs704959	0.82	A/G	*VILL*	no
169442	9	100465428	100466715	0.44	F>M	3E-02	rs2805831	0.94	G/A	*NCBP1*	no
86726	2	113422458	113423631	0.38	F>M	5E-02	rs2048873	0.91	A/G	*SLC20A1*	no
39457	12	103343947	103345172	0.37	F>M	1E-04	rs7314012	1.00	C/T	*STAB2*	no
119416	4	53039708	53040318	0.37	F>M	4E-02	rs1482110	0.87	T/A	*SPATA18*	no
119416	4	53039708	53040318	0.37	F>M	4E-02	rs1482110	0.87	T/A	*SPATA18*	no
64257	17	4478778	4479765	0.34	F>M	2E-02	rs76919645	0.82	G/A	*PSMB6*	no
47117	14	21457709	21458755	0.22	F>M	6E-03	rs2783781	0.84	C/G	*METT11D1*	no
47117	14	21457709	21458755	0.22	F>M	6E-03	rs2771344	0.88	T/C	*METT11D1*	no
99976	21	33983809	33985701	0.20	F>M	2E-03	rs866412	0.96	A/T	*C21orf59*	no
117420	4	8441818	8443173	0.20	F>M	3E-02	rs2688247	0.80	C/T	*CPZ*	no

135132	5	159435561	159436998	0.18	F>M	4E-02	rs10044313	1.00	C/T	*TTC1*	no
135132	5	159435561	159436998	0.18	F>M	4E-02	rs34655688	0.84	T/C	*TTC1*	no
70119	18	3261386	3263119	0.18	F>M	4E-02	rs34603712	0.85	G/A	*MRLC2*	no
70119	18	3261386	3263119	0.18	F>M	4E-02	rs35012773	0.86	T/C	*MRLC2*	no
46215	13	103497715	103499345	0.17	F>M	5E-02	rs2296147	0.91	T/C	*BIVM*	no
47170	14	21978792	21980098	0.16	F>M	4E-02	rs3762163	0.87	C/T	*METTL3*	no
42430	13	27825029	27826385	0.15	F>M	5E-02	rs3118727	1.00	C/G	*RPL21*	no
139827	6	43422042	43424099	0.18	M>F	5E-02	rs3734687	0.98	T/C	*ABCC10*	no
75701	19	8398520	8400776	0.25	M>F	3E-02	rs2241589	0.99	C/T	*KANK3*	no
75701	19	8398520	8400776	0.25	M>F	3E-02	rs710949	0.99	G/A	*KANK3*	no
103212	22	38027582	38028294	0.26	M>F	4E-02	rs117316616	1.00	C/T	*SSTR3*	no
27866	11	58701179	58702961	0.28	M>F	5E-02	rs10896879	1.00	G/A	*GLYATL1*	no
27866	11	58701179	58702961	0.28	M>F	5E-02	rs10896880	1.00	T/C	*GLYATL1*	no
27866	11	58701179	58702961	0.28	M>F	5E-02	rs11229704	1.00	A/C	*GLYATL1*	no
27866	11	58701179	58702961	0.28	M>F	5E-02	rs12272494	0.89	C/A	*GLYATL1*	no
27866	11	58701179	58702961	0.28	M>F	5E-02	rs11229705	1.00	G/C	*GLYATL1*	no
6512	1	92011533	92012789	0.32	M>F	4E-02	rs17501512	0.98	G/C	*CDC7*	no
6512	1	92011533	92012789	0.32	M>F	4E-02	rs11164928	0.91	A/G	*CDC7*	no
156556	7	158673542	158674994	0.36	M>F	2E-02	rs28679564	0.91	C/G	*WDR60*	no
146743	6	168216069	168216820	0.41	M>F	3E-02	rs142852243	0.87	A/G	*LOC441179*	no
146743	6	168216069	168216820	0.41	M>F	3E-02	rs552410256	0.87	G/A	*LOC441179*	no
18916	10	60802789	60803759	0.42	M>F	2E-02	rs12249399	0.95	G/A	*PHYHIPL*	no
106724	3	33896868	33898303	0.44	M>F	2E-03	rs4679093	1.00	G/A	*PDCD6IP*	no

106724	3	33896868	33898303	0.44	M>F	2E-03	rs4679094	1.00	C/T	*PDCD6IP*	no
106724	3	33896868	33898303	0.44	M>F	2E-03	rs9834071	0.99	A/G	*PDCD6IP*	no
106724	3	33896868	33898303	0.44	M>F	2E-03	rs9816786	0.99	T/C	*PDCD6IP*	no
90764	2	189463157	189463679	0.45	M>F	3E-02	rs34323074	1.00	A/T	*GULP1*	no
75327	19	4979107	4980436	0.46	M>F	7E-03	rs56123572	0.95	C/T	*JMJD2B*	no
75327	19	4979107	4980436	0.46	M>F	7E-03	rs62114274	0.96	A/G	*JMJD2B*	no
75327	19	4979107	4980436	0.46	M>F	7E-03	rs62114275	0.95	C/T	*JMJD2B*	no
75197	19	4171453	4174720	0.50	M>F	2E-02	rs350847	0.90	G/C	*SIRT6*	no
75197	19	4171453	4174720	0.50	M>F	2E-02	rs350846	0.90	G/C	*SIRT6*	no
86808	2	114020420	114021337	0.53	M>F	7E-03	rs2305131	1.00	G/T	*LOC440900*	no
86808	2	114020420	114021337	0.53	M>F	7E-03	rs1466018	1.00	G/T	*LOC440900*	no
6948	1	97813992	97814803	0.53	M>F	1E-02	rs10875070	0.86	A/G	*FLJ35409*	no
4943	1	59942767	59943322	0.53	M>F	2E-02	rs17119503	0.93	G/C	*FGGY*	no
9387	1	159927658	159928482	0.57	M>F	9E-04	rs2820555	0.92	A/T	*DARC*	no
32501	11	124837726	124838563	0.57	M>F	3E-04	rs12276990	0.99	T/C	*OR10D1P*	no
32501	11	124837726	124838563	0.57	M>F	3E-04	rs12270044	0.99	G/A	*OR10D1P*	no
32501	11	124837726	124838563	0.57	M>F	3E-04	rs12277044	0.99	T/C	*OR10D1P*	no
32501	11	124837726	124838563	0.57	M>F	3E-04	rs12277161	0.99	T/C	*OR10D1P*	no
124774	4	162840338	162842399	0.58	M>F	3E-02	rs1470137	0.83	A/G	*FSTL5*	no
124774	4	162840338	162842399	0.58	M>F	3E-02	rs1470139	0.83	A/G	*FSTL5*	no
124774	4	162840338	162842399	0.58	M>F	3E-02	rs1549828	0.81	A/G	*FSTL5*	no
40942	12	122340405	122342098	0.59	M>F	2E-02	rs74421874	1.00	G/A	*PSMD9*	no
40942	12	122340405	122342098	0.59	M>F	2E-02	rs3825172	1.00	C/T	*PSMD9*	no

40942	12	122340405	122342098	0.59	M>F	2E-02	rs58379261	1.00	A/G	*PSMD9*	no
146823	6	168723545	168723997	0.60	M>F	2E-02	rs7770790	0.96	T/C	*LOC340178*	no
141180	6	74104112	74105158	0.61	M>F	1E-02	rs558198	1.00	G/C	*DDX43*	no
141180	6	74104112	74105158	0.61	M>F	1E-02	rs558198	0.90	G/C	*OOEP*	no
234	1	2340193	2340748	0.62	M>F	1E-03	rs12092052	0.95	C/T	*PEX10*	no
13623	1	224330323	224331343	0.63	M>F	2E-03	rs4653986	1.00	G/A	*FBXO28*	no
13623	1	224330323	224331343	0.63	M>F	2E-03	rs4653986	0.82	G/A	*NVL*	no
25228	11	9124867	9126186	0.65	M>F	7E-03	rs34605207	0.97	C/A	*SCUBE2*	no
25899	11	18278079	18278833	0.69	M>F	1E-02	rs11024589	0.98	A/C	*SAA1*	no
25899	11	18278079	18278833	0.69	M>F	1E-02	rs11024589	0.98	A/C	*SAA1*	no
25899	11	18278079	18278833	0.69	M>F	1E-02	rs11024589	0.98	A/C	*SAA1*	no
70772	18	12551504	12553615	0.69	M>F	4E-03	rs12961966	1.00	T/A	*SPIRE1*	no
75329	19	4991287	4992414	0.72	M>F	1E-03	rs197141	0.87	G/C	*JMJD2B*	no
75329	19	4991287	4992414	0.72	M>F	1E-03	rs197140	0.86	T/G	*JMJD2B*	no
57330	15	91608684	91609377	0.81	M>F	8E-04	rs77232286	1.00	C/A	*ZNF774*	no
115081	3	180085094	180086262	0.96	M>F	3E-14	rs116410405	0.87	G/C	*GNB4*	no
115081	3	180085094	180086262	0.96	M>F	3E-14	rs116027546	0.87	G/A	*GNB4*	no

Table 4-6: Liver eQTL overlap with sex-biased peaks.

List of liver eQTL variants overlapping a sex-biased liver peak. 71 unique sex-biased peaks link to 81 unique genes. “Sex-Biased Peak ID” is a unique peak identifier. An “*” after the peak ID indicates that the eQTL signal was sex-biased. “Chr”, “Start position”, and “Stop position” identify the chromosome coordinates of the peak. “LFC” reports the log fold change for the sex-biased peak. “Sex-bias” identifies direction of the sex-biased peak where “M>F” indicates male-biased peaks and “F>M” indicates female-biased peaks. “Q-Value” reports the FDR-adjusted p-value. “Proxy rsID” reports the eQTL variant located within the peak. “r^2” reports the linkage disequilibrium between the proxy and lead variant at the eQTL signal, based on 1000 Genomes European reference. “Allele” reports the allele at the proxy variant in the format “major/minor”. “Gene” reports the differentially expressed gene corresponding to the eQTL signal. “GWAS Overlap” reports whether the peak also overlaps a variant associated with a GWAS trait from Table 4-7.

Sex-Biased Peak ID	Chr	Start position	Stop position	Sex-Bias	LFC	Q-Value	Proxy Variant	Proxy Variant Allele	r^2	Significant Traits	eQTL Overlap
16719	10	18971524	18974772	F>M	0.71	4E-03	rs12355670	CT/GC	0.96	Adolescent idiopathic scoliosis	*ARL5B, CACNB2, NSUN6*
16719	10	18971524	18974772	F>M	0.71	4E-03	rs12356702	CG/GT	0.96	Adolescent idiopathic scoliosis	*ARL5B, CACNB2, NSUN6*
16719	10	18971524	18974772	F>M	0.71	4E-03	rs111648476	CG/GA	0.96	Adolescent idiopathic scoliosis	*ARL5B, CACNB2, NSUN6*
52125	14	102543276	102546295	F>M	0.68	4E-02	rs6575894	CC/TT	0.81	Body mass index, Chronic obstructive pulmonary disease or resting heart rate (pleiotropy)	*DYNC1H1*
19490	10	70979797	70980602	F>M	0.67	2E-03	rs10823318	CA/TT	0.97	Glycemic traits (pregnancy), Birth weight, Offspring birth weight	*HKDC1*
8603	1	150544095	150545106	F>M	0.67	2E-02	rs11204669	GG/AA	0.99	Lung function (FVC), FEV1	*PRUNE*
8603	1	150544095	150545106	F>M	0.67	2E-02	rs12406712	GC/AA	0.92	Lung function (FVC), FEV1	*PRUNE*
8603	1	150544095	150545106	F>M	0.67	2E-02	rs12131809	TT/GC	0.92	Monocyte count, Monocyte percentage of white cells, Granulocyte percentage of myeloid white cells, Serum total protein level	*PRUNE*
8602	1	150539762	150542939	F>M	0.65	1E-05	rs932054	GC/AG	0.89	Lung function (FVC), FEV1	*ADAMTSL4, PRUNE*
8602	1	150539762	150542939	F>M	0.65	1E-05	rs72700829	AT/GC	0.86	Keratinocyte cancer (MTAG), Schizophrenia	*ADAMTSL4, PRUNE*
85867	2	100916062	100916411	F>M	0.64	2E-02	rs2033748	AA/GG	0.99	Intelligence (MTAG), Body mass index, Urinary sodium excretion	*LONRF2*
168994	9	95004259	95005536	F>M	0.47	2E-02	rs10125474	TT/AC	0.99	Blood protein levels	*IARS, OGN*

19491	10	70981671	70983070	F>M	0.46	3E-02	rs4746822	CC/TT	1.00	Glycemic traits (pregnancy), Birth weight, Offspring birth weight	*HKDC1*
103629	22	41856938	41857969	F>M	0.46	4E-02	rs5751107	CA/TG	0.93	Bitter alcoholic beverage consumption, Allergic disease, Respiratory diseases, Lifetime smoking index, Alcohol consumption, Bitter non-alcoholic beverage consumption, Smoking cessation (MTAG)	*NHP2L1, TOB2*
97740	20	47314202	47315371	F>M	0.44	2E-02	rs1040559	TG/CA	0.90	Diastolic blood pressure	*PREX1*
97740	20	47314202	47315371	F>M	0.44	2E-02	rs761274	TC/CT	0.89	Diastolic blood pressure	*PREX1*
28958	11	68657459	68659541	F>M	0.41	3E-02	rs514833	GT/AC	0.97	High density lipoprotein cholesterol levels	*IGHMBP2, MRPL21, MTL5*
28958	11	68657459	68659541	F>M	0.41	3E-02	rs488363	GG/AC	0.97	High density lipoprotein cholesterol levels	*IGHMBP2, MRPL21, MTL5*
28958	11	68657459	68659541	F>M	0.41	3E-02	rs660614	GA/AG	0.97	High density lipoprotein cholesterol levels	*IGHMBP2, MRPL21, MTL5*
85866	2	100907449	100909015	F>M	0.37	4E-02	rs4851285	AA/GG	0.98	Intelligence (MTAG), Body mass index, Urinary sodium excretion	*LONRF2*
1441	1	16508024	16508884	F>M	0.36	4E-02	rs12562207	CG/TA	0.83	Total cholesterol levels, Gamma glutamyl transferase levels, Lung function (FEV1/FVC), Liver enzyme levels (gamma-glutamyl transferase)	*EPHA2*
1441	1	16508024	16508884	F>M	0.36	4E-02	rs12057175	CA/TG	0.82	Total cholesterol levels, Gamma glutamyl transferase levels, Lung function (FEV1/FVC), Liver enzyme levels (gamma-glutamyl transferase)	*EPHA2*
1441	1	16508024	16508884	F>M	0.36	4E-02	rs12057222	CA/TG	0.82	Total cholesterol levels, Gamma glutamyl transferase levels, Lung function (FEV1/FVC), Liver enzyme levels (gamma-glutamyl transferase)	*EPHA2*
105871	3	17794733	17795621	F>M	0.35	5E-02	rs13079096	CT/TA	0.85	Self-reported risk-taking behavior	*TBC1D5*

138463	6	28129157	28130003	F>M	0.32	7E-03	rs13197176	GT/AC	0.80	Autism spectrum disorder or schizophrenia, Highest math class taken, Lung cancer in smokers, Well-being spectrum	*ZNF165*
138463	6	28129157	28130003	F>M	0.32	7E-03	rs9357065	AC/GT	0.81	Life satisfaction, Positive affect, Neuroticism	*ZNF165*
138463	6	28129157	28130003	F>M	0.32	7E-03	rs9357066	AC/GT	0.81	Life satisfaction, Positive affect, Neuroticism	*ZNF165*
138463	6	28129157	28130003	F>M	0.32	7E-03	rs1150668	GG/TT	1.00	Smoking initiation, Smoking status	*ZNF165*
138463	6	28129157	28130003	F>M	0.32	7E-03	rs1225618	CC/TA	0.98	Autism spectrum disorder or schizophrenia	*ZNF165*
138463	6	28129157	28130003	F>M	0.32	7E-03	rs13197175	CT/TC	0.89	Feeling nervous, Squamous cell lung carcinoma, Smoking initiation, Number of sexual partners, Triglycerides, Positive affect, QRS complex (Cornell), Asthma and major depressive disorder, Well-being spectrum, Autism spectrum disorder or schizophrenia, Intelligence (MTAG), Blood protein levels, Pulse pressure, Cardiometabolic and hematological traits, Depression (broad), Lung function (low FEV1 vs high FEV1), Lung adenocarcinoma, Depressive symptoms, Sarcoidosis, Estimated glomerular filtration rate in diabetes, Schizophrenia, Cognitive performance, Breast cancer, Urate levels, Worry	*ZNF165*
138463	6	28129157	28130003	F>M	0.32	7E-03	rs13197176	CT/TC	0.90	Autism spectrum disorder or schizophrenia, Highest math class taken, Lung cancer in ever smokers, Well-being spectrum	*ZNF165*
84694	2	73963871	73965175	F>M	0.19	5E-02	rs11894953	GC/AT	0.94	Serum metabolite levels	*SEMA4F, TPRKB*
84694	2	73963871	73965175	F>M	0.19	5E-02	rs17350188	GG/AC	0.94	Serum metabolite levels	*SEMA4F, TPRKB*
151111	7	66056407	66057775	F>M	0.17	1E-03	rs1968127	GT/AC	0.93	Calcium levels, Corneal structure, Estimated glomerular filtration rate	*RABGEF1*
37710	12	69978944	69980012	F>M	0.16	5E-02	rs2601007	TC/CG	0.98	Urinary albumin excretion, Urinary albumin excretion (no hypertensive medication), Urinary albumin-to-creatinine ratio	*CCT2*

37710	12	69978944	69980012	F>M	0.16	5E-02	rs2601006	TT/CC	1.00	Urinary albumin excretion, Urinary albumin excretion (no hypertensive medication), Urinary albumin-to-creatinine ratio	*CCT2*
147195	7	1095631	1096472	M>F	0.24	2E-02	rs76713558	GG/AA	0.97	Cholesterol, total, C-reactive protein levels or total cholesterol levels (pleiotropy)	*C7orf50*
147195	7	1095631	1096472	M>F	0.24	2E-02	rs79658522	GG/AA	0.97	Cholesterol, total, C-reactive protein levels or total cholesterol levels (pleiotropy)	*C7orf50*
147195	7	1095631	1096472	M>F	0.24	2E-02	rs76525951	GT/AC	0.97	Cholesterol, total, C-reactive protein levels or total cholesterol levels (pleiotropy)	*C7orf50*
147195	7	1095631	1096472	M>F	0.24	2E-02	rs79683221	GG/AA	0.97	Cholesterol, total, C-reactive protein levels or total cholesterol levels (pleiotropy)	*C7orf50*
147195	7	1095631	1096472	M>F	0.24	2E-02	rs78185801	GA/AG	0.97	Cholesterol, total, C-reactive protein levels or total cholesterol levels (pleiotropy)	*C7orf50*
3825	1	44015530	44016175	M>F	0.30	4E-02	rs2819336	GT/AC	0.91	Smoking initiation (MTAG), Smoking initiation, Smoking cessation (MTAG), Height	*PTPRF*
3825	1	44015530	44016175	M>F	0.30	4E-02	rs11580258	GA/AG	0.93	Smoking initiation (MTAG), Smoking initiation, Smoking cessation (MTAG), Height	*PTPRF*
166965	9	33828335	33828969	M>F	0.37	4E-02	rs12376951	TA/CG	0.97	Body mass index	*UBAP2*
41097	12	123617506	123618546	M>F	0.49	3E-03	rs6488864	TG/CC	0.83	Schizophrenia, Multiple sclerosis	*C12orf65, CDK2AP1, SBNO1*
41097	12	123617506	123618546	M>F	0.49	3E-03	rs6488864	AG/GC	0.90	Schizophrenia, Multiple sclerosis	*C12orf65, CDK2AP1, SBNO1*
41097	12	123617506	123618546	M>F	0.49	3E-03	rs1790116	AG/GT	0.87	Lymphocyte percentage of white cells, Diastolic blood pressure, Uterine fibroids, Highest math class taken, Self-reported math ability, Heel bone mineral density, Hair color, Intelligence (MTAG), Eosinophil counts, Type 2 diabetes, Allergy, Height, Insomnia, Venous	*C12orf65, CDK2AP1, SBNO1*

										thromboembolism, Head circumference (infant), Knee osteoarthritis	
96265	20	25260512	25261441	M>F	0.51	6E-04	rs2261720	AG/GT	0.84	Lung function (FVC), Liver enzyme levels (alkaline phosphatase)	*ABHD12, RP4-691N24.1*
96265	20	25260512	25261441	M>F	0.51	6E-04	rs2227890	AG/GA	0.84	Lung function (FVC), Liver enzyme levels (alkaline phosphatase)	*ABHD12, RP4-691N24.1*
111383	3	122162543	122164131	M>F	0.64	1E-03	rs7620827	AT/GC	1.00	Mean corpuscular hemoglobin, LDL cholesterol levels	*WDR5B*
142791	6	109625263	109626179	F>M	0.89	1E-07	rs1546722	AA/GG	0.92	Platelet crit, Basophil percentage of white cells, Mean corpuscular hemoglobin, High light scatter reticulocyte count, White blood cell count (basophil), Platelet count	NA
142791	6	109625263	109626179	F>M	0.89	1E-07	rs1546722	AG/GA	0.84	Platelet crit, Basophil percentage of white cells, Mean corpuscular hemoglobin, High light scatter reticulocyte count, White blood cell count (basophil), Platelet count	NA
142791	6	109625263	109626179	F>M	0.89	1E-07	rs1546723	AA/GG	1.00	Mean corpuscular hemoglobin concentration, Lymphocyte percentage of white cells, White blood cell count, Red cell distribution width, mean corpuscular hemoglobin, High light scatter reticulocyte percentage of red cells, Neutrophil percentage of white cells, Red blood cell traits, Mosaic loss of chromosome Y, Red blood cell count, Mean corpuscular volume	NA
142791	6	109625263	109626179	F>M	0.89	1E-07	rs1006081	CC/TT	0.85	HDL cholesterol, Heel bone mineral density, HDL cholesterol levels	NA
15599	10	3815754	3818722	F>M	0.89	1E-05	rs10795075	TT/CC	1.00	Red blood cell count, Heel bone mineral density, Height	NA
15599	10	3815754	3818722	F>M	0.89	1E-05	rs7071909	TT/CA	0.98	Red blood cell count, Heel bone mineral density, Height	NA
142792	6	109629216	109629734	F>M	0.84	8E-04	rs11153168	GT/AA	0.98	Mean corpuscular hemoglobin, Immature fraction of reticulocytes, Mean corpuscular volume, Red blood cell count	NA

152030	7	84582381	84583000	F>M	0.75	2E-03	rs2141415	GT/AC	0.86	Waist-to-hip ratio adjusted for BMI, Peak expiratory flow, FEV1, Waist-hip ratio, Lung function (FVC)	NA
152030	7	84582381	84583000	F>M	0.75	2E-03	rs17159470	GC/AT	0.86	Waist-to-hip ratio adjusted for BMI, Peak expiratory flow, FEV1, Waist-hip ratio, Lung function (FVC)	NA
87252	2	121985358	121985912	F>M	0.73	9E-05	rs34773350	TT/CC	0.99	Hemoglobin concentration, Hematocrit, Red blood cell count, Estimated glomerular filtration rate	NA
143509	6	122101284	122102441	F>M	0.69	2E-02	rs2091624	TC/CG	0.98	Diastolic blood pressure, Atrial fibrillation, Resting heart rate, RR interval (heart rate), Heel bone mineral density, Biomedical quantitative traits, Pulse pressure	NA
109804	3	85609257	85610007	F>M	0.64	2E-03	rs59073108	GA/AG	0.90	Body mass index, Systolic blood pressure, feeling worry, Anxiety/tension, Neuroticism, Worry, Smoking initiation, Information processing speed, Alcohol consumption (MTAG), Adventurousness, Automobile speeding propensity, Smoking status, Self-reported risk-taking behavior, Pulse pressure	NA
109804	3	85609257	85610007	F>M	0.64	2E-03	rs62250467	GT/AC	0.90	Body mass index, Systolic blood pressure, feeling worry, Anxiety/tension, Neuroticism, Worry, Smoking initiation, Information processing speed, Alcohol consumption (MTAG), Adventurousness, Automobile speeding propensity, Alcohol consumption, Smoking status, Self-reported risk-taking behavior, Smoking status, Pulse pressure	NA
109804	3	85609257	85610007	F>M	0.64	2E-03	rs17459563	GG/AA	0.90	Body mass index, Systolic blood pressure, feeling worry, Anxiety/tension, Neuroticism, Worry, Smoking initiation, Information processing speed, Alcohol consumption (MTAG), Adventurousness, Automobile speeding propensity, Alcohol consumption, Smoking status, Self-reported risk-taking behavior, Smoking status, Pulse pressure	NA
109804	3	85609257	85610007	F>M	0.64	2E-03	rs72615727	GG/AC	0.89	Body mass index, Systolic blood pressure, feeling worry, Anxiety/tension, Neuroticism,	NA

										Worry, Smoking initiation, Information processing speed, Alcohol consumption (MTAG), Adventurousness, Automobile speeding propensity, Alcohol consumption, Smoking status, Self-reported risk-taking behavior, Pulse pressure	
109804	3	85609257	85610007	F>M	0.64	2E-03	rs77852438	GC/AT	0.90	Body mass index, Systolic blood pressure, feeling worry, Anxiety/tension, Neuroticism, Worry, Smoking initiation, Information processing speed, Alcohol consumption (MTAG), Adventurousness, Automobile speeding propensity, Smoking status, Self-reported risk-taking behavior, Pulse pressure	NA
109804	3	85609257	85610007	F>M	0.64	2E-03	rs78288623	GC/AA	0.90	Body mass index, Systolic blood pressure, feeling worry, Anxiety/tension, Neuroticism, Worry, Smoking initiation, Information processing speed, Alcohol consumption (MTAG), Adventurousness, Automobile speeding propensity, Smoking status, Self-reported risk-taking behavior, Pulse pressure	NA
109804	3	85609257	85610007	F>M	0.64	2E-03	rs1449378	GC/AT	0.90	Body mass index, Systolic blood pressure, feeling worry, Anxiety/tension, Neuroticism, Worry, Smoking initiation, Information processing speed, Alcohol consumption (MTAG), Adventurousness, Automobile speeding propensity, Smoking status, Self-reported risk-taking behavior, Pulse pressure	NA
109804	3	85609257	85610007	F>M	0.64	2E-03	rs1449379	GT/AC	0.90	Body mass index, Systolic blood pressure, feeling worry, Anxiety/tension, Neuroticism, Worry, Smoking initiation, Information processing speed, Alcohol consumption (MTAG), Adventurousness, Automobile speeding propensity, Smoking status, Self-reported risk-taking behavior, Pulse pressure	NA
15195	1	247431835	247433554	F>M	0.63	1E-04	rs1771918	CA/TG	0.95	Monocyte percentage of white cells	NA
15195	1	247431835	247433554	F>M	0.63	1E-04	rs6698384	CA/TG	0.99	Monocyte percentage of white cells	NA
172544	9	139281523	139282457	F>M	0.60	3E-02	rs34826348	AC/GA	0.95	Chronic inflammatory diseases (pleiotropy), Ulcerative colitis, Inflammatory bowel disease,	NA

										Lung function (FVC), Granulocyte percentage of myeloid white cells, Crohn's disease	
172544	9	139281523	139282457	F>M	0.60	3E-02	rs78428995	AC/GT	0.97	Chronic inflammatory diseases (pleiotropy), Ulcerative colitis, Inflammatory bowel disease, Lung function (FVC), Granulocyte percentage of myeloid white cells, Crohn's disease	NA
32179	11	120300448	120300822	F>M	0.60	5E-02	rs11217866	AA/GT	0.99	Intraocular pressure, Glaucoma	NA
91281	2	198346630	198350409	F>M	0.58	4E-02	rs6745660	CA/TG	0.98	Morning person	NA
145677	6	154911878	154912437	F>M	0.55	5E-02	rs4458701	TT/GC	0.94	Lipoprotein (a) levels	NA
120445	4	76928125	76928656	F>M	0.54	2E-03	rs2276886	TT/CC	0.86	Blood protein levels, Neurological blood protein biomarker levels	NA
120445	4	76928125	76928656	F>M	0.54	2E-03	rs2276886	GT/TC	0.87	Blood protein levels, Neurological blood protein biomarker levels	NA
19224	10	65125715	65126430	F>M	0.53	2E-02	rs72835389	CC/TT	0.97	Immature fraction of reticulocytes, High light scatter reticulocyte count	NA
162041	8	95933943	95937370	F>M	0.52	1E-03	rs12548874	TC/CA	0.84	Hemoglobin concentration, Eosinophil counts, Type 2 diabetes	NA
154126	7	120724744	120725465	F>M	0.51	1E-02	rs1024743	TA/GC	0.94	Heel bone mineral density	NA
87251	2	121977359	121978002	F>M	0.49	3E-02	rs34764243	TA/CC	0.96	Hemoglobin concentration, Hematocrit, Red blood cell count, Estimated glomerular filtration rate	NA
36570	12	53289062	53290882	F>M	0.47	9E-04	rs4919707	AG/GA	0.84	Prostate cancer, Cancer (pleiotropy)	NA
167318	9	37966399	37967160	F>M	0.46	8E-03	rs7862130	CG/AA	0.85	Lung function (FEV1/FVC)	NA
21653	10	98623589	98624592	F>M	0.45	4E-02	rs7921885	CT/GG	0.98	Height	NA
119917	4	67900796	67902221	F>M	0.43	2E-02	rs1435434	TG/GT	0.94	Intelligence (MTAG), Highest math class taken (MTAG), Educational attainment (years of education), Educational attainment (MTAG)	NA

157080	8	8093070	8096615	F>M	0.33	1E-03	rs9693219	AC/GG	0.87	Systemic lupus erythematosus, Neuroticism, General factor of neuroticism, Red cell distribution width, Estimated glomerular filtration rate, Schizophrenia	NA
157080	8	8093070	8096615	F>M	0.33	1E-03	rs2945251	AA/GG	0.92	Systemic lupus erythematosus, Neuroticism, General factor of neuroticism, Red cell distribution width, Estimated glomerular filtration rate, Schizophrenia	NA
157080	8	8093070	8096615	F>M	0.33	1E-03	rs2945250	AC/GT	0.88	Systemic lupus erythematosus, Neuroticism, General factor of neuroticism, Red cell distribution width, Estimated glomerular filtration rate, Schizophrenia	NA
157080	8	8093070	8096615	F>M	0.33	1E-03	rs2945249	AA/GG	0.91	Systemic lupus erythematosus, Neuroticism, General factor of neuroticism, Red cell distribution width, Estimated glomerular filtration rate, Schizophrenia	NA
157080	8	8093070	8096615	F>M	0.33	1E-03	rs2980437	AC/GT	0.92	Systemic lupus erythematosus, Neuroticism, General factor of neuroticism, Red cell distribution width, Estimated glomerular filtration rate, Schizophrenia	NA
157080	8	8093070	8096615	F>M	0.33	1E-03	rs2980438	AC/GT	0.88	Systemic lupus erythematosus, Neuroticism, General factor of neuroticism, Red cell distribution width, Estimated glomerular filtration rate, Schizophrenia	NA
157080	8	8093070	8096615	F>M	0.33	1E-03	rs2980439	AA/GG	1.00	Systemic lupus erythematosus, Neuroticism, General factor of neuroticism, Red cell distribution width, Estimated glomerular filtration rate, Schizophrenia	NA
157080	8	8093070	8096615	F>M	0.33	1E-03	rs2948294	AG/GA	0.90	Systemic lupus erythematosus, Neuroticism, General factor of neuroticism, Red cell distribution width, Estimated glomerular filtration rate, Schizophrenia	NA
157080	8	8093070	8096615	F>M	0.33	1E-03	rs2980440	AA/GG	0.91	Systemic lupus erythematosus, Neuroticism, General factor of neuroticism, Red cell distribution width, Estimated glomerular filtration rate, Schizophrenia	NA

157080	8	8093070	8096615	F>M	0.33	1E-03	rs2980441	AC/GG	0.86	Systemic lupus erythematosus, Neuroticism, General factor of neuroticism, Red cell distribution width, Estimated glomerular filtration rate, Schizophrenia	NA
157080	8	8093070	8096615	F>M	0.33	1E-03	rs2945248	AT/GC	0.81	Systemic lupus erythematosus, Neuroticism, General factor of neuroticism, Red cell distribution width, Estimated glomerular filtration rate, Schizophrenia	NA
157080	8	8093070	8096615	F>M	0.33	1E-03	rs2945247	AA/GG	0.83	Systemic lupus erythematosus, Neuroticism, General factor of neuroticism, Red cell distribution width, Estimated glomerular filtration rate, Schizophrenia	NA
157080	8	8093070	8096615	F>M	0.33	1E-03	rs73199790	GC/AT	0.86	Morning person	NA
157080	8	8093070	8096615	F>M	0.33	1E-03	rs13258063	CA/TC	0.80	Systemic lupus erythematosus, Neuroticism, General factor of neuroticism, Red cell distribution width, Estimated glomerular filtration rate, Schizophrenia	NA
129097	5	55273175	55274626	F>M	0.31	1E-02	rs13183065	GA/AG	0.83	Blood protein levels	NA
129097	5	55273175	55274626	F>M	0.31	1E-02	rs13170520	GC/AT	1.00	Blood protein levels	NA
129097	5	55273175	55274626	F>M	0.31	1E-02	rs11744301	GT/AC	1.00	Blood protein levels	NA
146174	6	160675587	160676390	F>M	0.31	2E-02	rs3798156	AT/TC	0.81	Chronic kidney disease, Estimated glomerular filtration rate	NA
146174	6	160675587	160676390	F>M	0.31	2E-02	rs316008	TG/GT	0.93	Estimated glomerular filtration rate	NA
146174	6	160675587	160676390	F>M	0.31	2E-02	rs316009	AT/GC	0.99	Blood metabolite levels, Glomerular filtration rate, Serum metabolite levels, Blood metabolite ratios, Estimated glomerular filtration rate, Creatinine levels	NA
96385	20	30292203	30292918	F>M	0.29	5E-02	rs6060812	AT/TC	0.90	Mean corpuscular hemoglobin, Mean corpuscular volume, Red blood cell count, Height	NA

112763	3	139146191	139146905	F>M	0.22	4E-02	rs9881973	CT/TC	1.00	Adolescent idiopathic scoliosis	NA
89227	2	160080545	160083047	M>F	0.26	4E-02	rs34344829	CA/TG	0.93	Red cell distribution width	NA
89227	2	160080545	160083047	M>F	0.26	4E-02	rs13419975	CT/TA	0.87	Red cell distribution width	NA
89227	2	160080545	160083047	M>F	0.26	4E-02	rs9678339	CG/TA	0.94	Red cell distribution width	NA
89227	2	160080545	160083047	M>F	0.26	4E-02	rs9678325	CC/TT	0.87	Red cell distribution width	NA
155239	7	139404385	139405470	M>F	0.33	2E-02	rs141212865	CC/AA	1.00	Systolic blood pressure	NA
172409	9	137852492	137853953	M>F	0.34	3E-02	rs4842200	AC/GT	0.96	Blood protein levels, White blood cell count	NA
52361	14	104568846	104573522	M>F	0.34	4E-02	rs8012505	TG/CC	0.80	Plasma free amino acid levels (adjusted for one other PFAA), Plasma free amino acid levels (adjusted for twenty other PFAAs), Hair color, Plasma free asparagine levels	NA
52361	14	104568846	104573522	M>F	0.34	4E-02	rs8012505	TG/CC	0.86	Plasma free amino acid levels (adjusted for one other PFAA), Plasma free amino acid levels (adjusted for twenty other PFAAs), Hair color, Plasma free asparagine levels	NA
157329	8	10587001	10588667	M>F	0.36	4E-03	rs6995692	GG/CC	1.00	Number of sexual partners, Diastolic blood pressure x smoking status (current vs non-current) interaction (2df test), Systolic blood pressure x smoking status (current vs non-current) interaction (2df test), Systolic blood pressure x smoking status (ever vs never) interaction (2df test), Diastolic blood pressure x smoking status (ever vs never) interaction (2df test), Carotid intima media thickness	NA
14070	1	229717711	229718457	M>F	0.40	3E-02	rs6687883	AT/GC	0.84	Eosinophil counts, Blood protein levels	NA
14070	1	229717711	229718457	M>F	0.40	3E-02	rs12049351	AG/GC	0.82	Eosinophil counts, Blood protein levels	NA
32799	11	128499575	128500091	M>F	0.41	3E-02	rs7945677	CC/TT	1.00	Chronic lymphocytic leukemia or systemic lupus erythematosus, Systemic lupus erythematosus	NA

48055	14	37832426	37833436	M>F	0.42	3E-03	rs1950520	AT/GC	0.85	Mean corpuscular hemoglobin, Red blood cell count	NA
48055	14	37832426	37833436	M>F	0.42	3E-03	rs12896270	AT/GC	0.85	Mean corpuscular hemoglobin, Red blood cell count	NA
172410	9	137854222	137855085	M>F	0.43	3E-02	rs11103602	AA/GG	1.00	Blood protein levels, White blood cell count	NA
78798	19	48650923	48651454	M>F	0.47	4E-02	rs16981519	CT/TC	0.88	Height	NA
78798	19	48650923	48651454	M>F	0.47	4E-02	rs12461815	CA/TC	0.87	Height	NA
77545	19	35769435	35770304	M>F	0.51	3E-03	rs34832786	TC/CT	1.00	Mean corpuscular hemoglobin, Mean corpuscular volume, Red cell distribution width	NA
77545	19	35769435	35770304	M>F	0.51	3E-03	rs28365136	TG/CT	1.00	Mean corpuscular hemoglobin, Mean corpuscular volume, Red cell distribution width	NA
77545	19	35769435	35770304	M>F	0.51	3E-03	rs2280747	TT/CC	0.99	Mean corpuscular hemoglobin, Mean corpuscular volume, Red cell distribution width	NA
77545	19	35769435	35770304	M>F	0.51	3E-03	rs1051995	TT/CC	1.00	Mean corpuscular hemoglobin, Mean corpuscular volume, Red cell distribution width	NA
130307	5	76551251	76551904	M>F	0.51	5E-02	rs12697863	CG/TA	1.00	Eosinophil counts	NA
74526	18	77647030	77648417	M>F	0.51	2E-03	rs56071379	GA/AG	0.93	Schizophrenia	NA
128272	5	37754737	37756085	M>F	0.52	1E-02	rs9292669	AC/CG	0.82	Diverticular disease	NA
130299	5	76434591	76435368	M>F	0.54	4E-02	rs7732130	GG/AA	1.00	Waist-to-hip ratio adjusted for BMI, Type 2 diabetes	NA
60960	16	54114323	54115493	M>F	0.80	2E-05	rs16953002	AA/GG	1.00	Melanoma, Hair color	NA
60960	16	54114323	54115493	M>F	0.80	2E-05	rs16953002	AA/GG	1.00	Melanoma, Hair color	NA
47635	14	30605821	30606822	M>F	0.85	7E-05	rs8019932	GG/TT	1.00	Educational attainment (years of education), Educational attainment (MTAG)	NA

Table 4-7: GWAS variants overlapping sex-biased peaks.

List of GWAS variants overlapping a sex-biased liver peak. 71 sex-biased peaks overlap a GWAS signal. “Sex-Biased Peak ID” is a unique peak identifier. “Chr”, “Start position”, and “Stop position” identify the hg19 chromosome coordinates of the peak. “Sex-bias” identifies direction of the sex-biased peak where “M>F” indicates male-biased peaks and “F>M” indicates female-biased peaks. “LFC” reports the log fold change for the sex-biased peak. “Q-Value” reports the FDR adjusted p-value. “Proxy rsID” reports the variant within the peak. “Proxy Variant Allele” reports the allele at the lead and proxy variant in the format “Lead major Proxy major/Lead minor Proxy minor”. “r^2” reports the linkage disequilibrium between the proxy and lead variants. “Significant Traits” reports the traits associated with the signal. “eQTL Overlap” reports the gene(s) associated if a liver eQTL variant was also identified within the indicated peak. “NA” indicates no eQTL overlap at the indicated peak

CHAPTER 5: DISCUSSION

GWAS have identified thousands of loci associated with cardiometabolic traits[44]. However, GWAS associations do not identify the mechanisms at these largely noncoding loci, including which variants and genes are involved, which cell types and contexts they are active in, and the molecular mechanisms of the functional variants[53]. Noncoding GWAS loci can have regulatory mechanisms that can be studied by profiling chromatin accessibility and linking accessible regions to genes[58]. These regulatory mechanisms can be cell type- and context-dependent[58,68,76,118]. Therefore, identification of regulatory elements in disease-relevant cell types and contexts can aid identification of the regulatory mechanisms. Furthermore, identification of cell type- and context-dependent genetic effects can inform treatment for cardiometabolic traits.

In this dissertation, I identified and described regulatory mechanisms in cardiometabolic disease-relevant tissues, cell types, and contexts. I profiled chromatin accessibility in adipose tissue, adipocytes under several cardiometabolic disease-relevant contexts, and liver tissue. Despite many challenges of working with high lipid adipose tissue, Chapter 2 describes a consensus human adipose chromatin accessibility map from 11 individuals, one of the largest human adipose sample sizes to date. In the SGBS adipocyte cell model, Chapter 2 describes regions of context-dependent chromatin accessibility during adipocyte differentiation, links of these candidate regulatory elements to genes and traits, and allele- and context-dependent effects of elements on transcriptional activity. I also investigated context-dependent chromatin accessibility of other disease-relevant contexts: exposure to excess lipids, hypoxia, and inflammation, described in Chapter 3. Finally, in Chapter 4, I described sex-biased chromatin accessibility in 139 human liver tissue samples. The work in this dissertation identifies hundreds of candidate regulatory mechanisms for noncoding GWAS loci. Furthermore, these chromatin accessibility profiles in disease-relevant tissues, cell types, and contexts will be an excellent resource for future work on elucidating regulatory mechanisms of disease.

I studied chromatin accessibility and regulatory mechanisms in both human adipose tissues and a human adipocyte cell model. There are advantages and disadvantages to each approach. While adipocytes are a major component cell type of adipose tissue, chromatin accessibility from adipose tissue may better capture regulatory elements relevant to adipose biology and disease compared to adipocytes from a cell model. This can be seen in our enrichment analyses in Chapter 2 (Fig 2.3), where adipose peaks are enriched for additional traits such as triglycerides and HDL-cholesterol compared to preadipocyte and adipocyte context-dependent peaks. Furthermore, chromatin accessibility in adipose tissue can be assayed across many individuals to capture more genetic and environmental variation.

Obtaining quality chromatin accessibility in adipose tissue across many individuals proved challenging. I planned to profile chromatin accessibility in ~400 adipose samples available from the METSIM study[15]. An advantage of the METSIM study is that individuals have genotype, gene expression, and cardiometabolic trait phenotyping[15,61,6515,61,65], however the individuals are all Finnish males, which limits genetic diversity. When initial chromatin accessibility profiles demonstrated inconsistent quality, I tested many factors to optimize nuclei isolation and ATAC-seq library preparation from frozen adipose tissue including buffers, detergents, filtering steps, transposase Tn5-to-nuclei ratio, and the Omni ATAC-seq protocol[88]. After optimization, we produced a consensus map of human adipose chromatin accessibility from 11 individuals, however the quality remained inconsistent enough to proceed with the larger sample set. Based on my experience with optimizing chromatin accessibility profiling in frozen adipose tissue, I suspect that the lipid content adversely affects the nuclei isolation and/or library preparation or that the freezing protocol adversely affected the chromatin structure. Evidence to support the lipid content adversely affecting the library preparation includes that data from mature adipocytes of the cell model were less consistent in quality than data from preadipocytes and that adipose tissue samples from a separate source with a different freezing protocol also showed inconsistent quality. Inconsistent quality of chromatin accessibility libraries in our mature adipocytes contributed to difficulties in identifying context-dependent regulatory elements due to additional disease-relevant contexts such as free fatty acids, hypoxia, and inflammation described in Chapter 3. Future testing of fresh adipose tissue or adipocytes could help identify the cause of the inconsistent quality. While chromatin accessibility profiling in adipose tissue proved challenging, producing maps in larger sample sizes will capture additional genetic and environmental variation. The 11-sample consensus map we developed represents more genetic diversity than any existing adipose dataset.

Adipocyte cell models are also useful for studying regulatory mechanisms of disease because they can provide a consistent genetic background to compare changes due to environmental perturbation. Adipose tissue is heterogenous and composed of many cell types, including preadipocytes and adipocytes[11], but regulatory mechanisms can act in cell type- and context-dependent manners[58,68,118]. Therefore, profiling chromatin accessibility in relevant cell types and contexts can identify context-dependent regulatory mechanisms that could be missed in heterogenous tissue samples that may lack relevant context. Chapter 2 described regulatory mechanisms of disease in adipocyte-dependent regions. Our functional tests of variants at two loci (*SCD* and *EYA2*) showed context-dependent regulatory mechanisms and identified allele-

and context-dependent transcriptional effects at the *SCD* locus. The results at *EYA2* were more complex, as we identified both a context-dependent regulatory element and an element only present in adipose tissue. We tested both elements and only identified allele-dependent transcriptional effects at the adipose tissue element. This result demonstrates that, while context-dependent regulatory elements can identify molecular mechanisms at GWAS loci, the identified variant may not be causal and other mechanisms or contexts may be involved. Another possibility is that allelic effects on transcriptional activity for the context-dependent region were not detectable in our *in vitro* reporter assay. Additionally, this result demonstrates the utility of our consensus adipose map, which may better represent biologically-relevant regulatory elements.

Future work on elucidating the molecular mechanisms at these and other loci we identified in Chapter 2 could include performing additional assays in adipocytes, such as electrophoretic mobility shift assays to detect differential binding of alleles to nuclear proteins and transcription factors[53,84], ChIP-seq to identify which transcription factors bind to context-dependent regulatory elements or allelic differences in transcription factor binding[53,186], or CRISPR-Cas9 to delete or inactivate the regulatory region or create an alternate allele[53,187–189]. In Chapter 2, we used HOMER[190] to identify transcription factor binding motifs enriched in context-dependent regulatory elements, a computational method that could also be applied to other contexts such as sex-biased chromatin accessibility in liver. Finally, while individually functionally testing candidate regulatory mechanisms allows for accurate evaluation, assays such as massively parallel reporter assays would allow high-throughput testing of many candidate regulatory elements in a single experiment[53,191,192]. Together, these assays can be used to test additional predictions from our analyses to identify many more regulatory mechanisms of disease.

While cell models are useful for studying effects against a consistent genetic background in a controlled environment, cell models have some disadvantages. The consistent genetic background that simplifies many studies hinders the ability to identify interactions between genetic variation and environment. For example, in Chapter 2 we identified allelic imbalance in adipocyte chromatin accessibility. One limitation of allelic imbalance in our adipocyte cell model is that we could only test for allelic imbalance at heterozygous sites within the one individual with SGBS from whom the cells were derived. Allelic imbalance testing in a larger sample size would likely identify additional significant imbalances because more heterozygous sites are available to test, more sequencing reads exist at any given site, and imbalances can be validated across individuals. One approach to overcome the disadvantages of using a single cell model would be to use multiple cell lines in a model such as induced pluripotent stem cells derived from multiple individuals, which would allow diverse genetic backgrounds to be tested[68,117,118]. Another approach to study different cell types against diverse genetic backgrounds is to perform single nucleus sequencing strategies on tissue from multiple samples. Single nucleus ATAC-seq and RNA-seq can be performed tissue to resolve issues with heterogeneity and study the cell type-specific regulatory landscape[193–195]. Another disadvantage is that aspects of cell models are not biologically relevant. For example, I used SGBS adipocytes because they are mostly diploid, however, they grow in cell culture because they were derived from an individual with a disease state that causes adipocyte overgrowth[22]. Additionally, growing cells in culture can introduce changes due to the artificial environment. For these reasons, it is important to build resources such as our consensus map of adipose tissue chromatin accessibility, which can help identify regions that are more likely to be biologically relevant. We mitigated the limitation of using an adipocyte cell model with one genetic background by identifying consensus adipose chromatin accessibility from 11 individuals.

In addition to studies of adipose and adipocytes in a disease-relevant context, I studied regulatory elements in human liver tissue that are biased between males and females. Sex is a relevant context to cardiometabolic disease that contributes to differences in disease risk and response to treatment[24–26]. In Chapter 4, we profiled chromatin accessibility in 139 human liver samples and identified 774 autosomal sex-biased regions (LFC > 0, FDR < 5%) that demonstrated significant differences between males and females. The average LFC of these sex-biased regions is 0.5, suggesting that these are modest differences between males and females. When we applied a less stringent threshold (LFC > 0, FDR < 10%), selected to match a sex-biased chromatin accessibility study in peripheral blood mononuclear cells that had identified 577 sex-biased regions (0.69% of tested regions)[171], we identified a comparable number of 1300 autosomal sex-biased regions, which represents 0.75% of tested regions. We linked the sex-biased regions to genes using existing liver eQTL data[62] and to disease traits using the GWAS catalog[44]. Additional lines of evidence could be used to link sex-biased regulatory elements to genes in future work, including chromosome conformation capture profiles. Future work would also include functional testing of candidate regulatory elements using methods such as those discussed in Chapter 2[53]. While our liver samples have existing genotype and gene expression data[62] and represent a mix of sexes and ancestries that can capture additional genetic and environmental variation, we have limited phenotype data on the donor individuals, which could limit identifying associations between regulatory elements and traits. However, these liver chromatin accessibility profiles represent a valuable resource that can be used for future studies such as identifying genetic variants that alter chromatin accessibility through colocalization of eQTL and GWAS data[28].

Identifying molecular mechanisms at GWAS loci remains complex. The work presented in this dissertation contributes to the understanding of how genetic variation and cellular context contribute to cardiometabolic traits. I produced chromatin accessibility maps for a variety of tissues and adipocytes in multiple cellular contexts. I used these chromatin accessibility maps to predict candidate functional variants and regulatory mechanisms. At specific loci, we used these predictions to identify allelic differences in transcriptional activity. Furthermore, these chromatin accessibility profiles will be a useful resource for future work on identifying regulatory mechanisms of GWAS loci. Identifying genetic variants that alter gene expression to contribute to disease can identify drug targets and the direction of effect to increase or lower activity to treat disease. Although functional testing is needed, some of the candidate variants identified in these studies could identify individuals at higher risk of cardiometabolic disease or individuals who may respond better to specific treatments.

REFERENCES

1. Balakumar, P., Maung-U, K. & Jagadeesh, G. Prevalence and prevention of cardiovascular disease and diabetes mellitus. *Pharmacol. Res.* **113**, 600–609 (2016).

2. Virani, S. S. *et al.* Heart Disease and Stroke Statistics—2021 Update: A Report From the American Heart Association. *Circulation* **143**, (2021).

3. Murphy, S. L. Mortality in the United States, 2020. 8 (2021).

4. Lloyd-Jones, D. M. *et al.* Defining and Setting National Goals for Cardiovascular Health Promotion and Disease Reduction: The American Heart Association's Strategic Impact Goal Through 2020 and Beyond. *Circulation* **121**, 586–613 (2010).

5. Després, J.-P. & Lemieux, I. Abdominal obesity and metabolic syndrome. *Nature* **444**, 881–887 (2006).

6. Cameron, A. J. *et al.* The influence of hip circumference on the relationship between abdominal obesity and mortality. *International Journal of Epidemiology* **41**, 484–494 (2012).

7. Badimon, L. & Vilahur, G. LDL-cholesterol versus HDL-cholesterol in the atherosclerotic plaque: inflammatory resolution versus thrombotic chaos: Badimon & Vilahur. *Annals of the New York Academy of Sciences* **1254**, 18–32 (2012).

8. Lackey, D. E. & Olefsky, J. M. Regulation of metabolism by the innate immune system. *Nat Rev Endocrinol* **12**, 15–28 (2016).

9. Cao, H. Adipocytokines in obesity and metabolic disease. *J. Endocrinol.* **220**, T47-59 (2014).

10. Goossens, G. H. The Metabolic Phenotype in Obesity: Fat Mass, Body Fat Distribution, and Adipose Tissue Function. *Obes Facts* **10**, 207–215 (2017).

11. Lynes, M. D. & Tseng, Y.-H. Deciphering adipose tissue heterogeneity. *Ann. N. Y. Acad. Sci.* **1411**, 5–20 (2018).

12. Ghaben, A. L. & Scherer, P. E. Adipogenesis and metabolic health. *Nat Rev Mol Cell Biol* **20**, 242–258 (2019).

13. Shuster, A., Patlas, M., Pinthus, J. H. & Mourtzakis, M. The clinical importance of visceral adiposity: a critical review of methods for visceral adipose tissue analysis. *BJR* **85**, 1–10 (2012).

14. Suárez-Cuenca, J. A. *et al.* Enlarged adipocytes from subcutaneous vs. visceral adipose tissue differentially contribute to metabolic dysfunction and atherogenic risk of patients with obesity. *Sci Rep* **11**, 1831 (2021).

15. Laakso, M. *et al.* The Metabolic Syndrome in Men study: a resource for studies of metabolic and cardiovascular diseases. *J. Lipid Res.* **58**, 481–493 (2017).

16. Li, S. *et al.* Resveratrol inhibits lipogenesis of 3T3-L1 and SGBS cells by inhibition of insulin signaling and mitochondrial mass increase. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **1857**, 643–652 (2016).

17. Morrison, S. & McGee, S. L. 3T3-L1 adipocytes display phenotypic characteristics of multiple adipocyte lineages. *Adipocyte* **4**, 295–302 (2015).

18. Sadowski, H. B., Wheeler, T. T. & Young, D. A. Gene expression during 3T3-L1 adipocyte differentiation. Characterization of initial responses to the inducing agents and changes during commitment to differentiation. *Journal of Biological Chemistry* **267**, 4722–4731 (1992).

19. Zebisch, K., Voigt, V., Wabitsch, M. & Brandsch, M. Protocol for effective differentiation of 3T3-L1 cells to adipocytes. *Analytical Biochemistry* **425**, 88–90 (2012).

20. Schmidt, S. F. *et al.* Cross species comparison of C/EBPα and PPARγ profiles in mouse and human adipocytes reveals interdependent retention of binding sites. *BMC Genomics* **12**, 152 (2011).

21. Wabitsch, M. *et al.* Characterization of a human preadipocyte cell strain with high capacity for adipose differentiation. *International journal of obesity* **25**, 8–15 (2001).

22. Fischer-Posovszky, P., Newell, F. S., Wabitsch, M. & Tornqvist, H. E. Human SGBS cells - a unique tool for studies of human fat cell biology. *Obes Facts* **1**, 184–9 (2008).

23. Trefts, E., Gannon, M. & Wasserman, D. H. The liver. *Curr. Biol.* **27**, R1147–R1151 (2017).

24. Gerdts, E. & Regitz-Zagrosek, V. Sex differences in cardiometabolic disorders. *Nat Med* **25**, 1657–1666 (2019).

25. Krishnan, K. C., Mehrabian, M. & Lusis, A. J. Sex differences in metabolism and cardiometabolic disorders. 14 (2019).

26. Kerkhof, P. L. M. & Miller, Virginia. *Sex-specific analysis of cardiovascular function*. (2018).

27. Khramtsova, E. A., Davis, L. K. & Stranger, B. E. The role of sex in the genomics of human complex traits. *Nat Rev Genet* **20**, 173–190 (2019).

28. Currin, K. W. *et al.* Genetic effects on liver chromatin accessibility identify disease regulatory variants. *Am J Hum Genet* **108**, 1169–1189 (2021).

29. Pandey, A. *et al.* Association of Intensive Lifestyle Intervention, Fitness, and Body Mass Index With Risk of Heart Failure in Overweight or Obese Adults With Type 2 Diabetes Mellitus: An Analysis From the Look AHEAD Trial. *Circulation* **141**, 1295–1306 (2020).

30. Van Buren, D. J. & Tibbs, T. L. Lifestyle Interventions to Reduce Diabetes and Cardiovascular Disease Risk Among Children. *Curr Diab Rep* **14**, 557 (2014).

31. Zhang, X. *et al.* Effect of lifestyle interventions on cardiovascular risk factors among adults without impaired glucose tolerance or diabetes: A systematic review and meta-analysis. *PLoS ONE* **12**, e0176436 (2017).

32. Almgren, P. *et al.* Heritability and familiality of type 2 diabetes and related quantitative traits in the Botnia Study. *Diabetologia* **54**, 2811–2819 (2011).

33. Flannick, J. & Florez, J. C. Type 2 diabetes: genetic data sharing to advance complex disease research. *Nat Rev Genet* **17**, 535–549 (2016).

34. Khera, A. V. & Kathiresan, S. Genetics of coronary artery disease: discovery, biology and clinical translation. *Nat. Rev. Genet.* **18**, 331–344 (2017).

35. Zdravkovic, S. *et al.* Heritability of death from coronary heart disease: a 36-year follow-up of 20 966 Swedish twins. *J. Intern. Med.* **252**, 247–254 (2002).

36. Vinkhuyzen, A. A. E., Wray, N. R., Yang, J., Goddard, M. E. & Visscher, P. M. Estimation and Partition of Heritability in Human Populations Using Whole-Genome Analysis Methods. *Annu. Rev. Genet.* **47**, 75–95 (2013).

37. Rose, K. M., Newman, B., Mayer-Davis, E. J. & Selby, J. V. Genetic and Behavioral Determinants of Waist-Hip Ratio and Waist Circumference in Women Twins. *Obesity Research* **6**, 383–392 (1998).

38. Liu, H. *et al.* Heritability and Genome-Wide Association Study of Plasma Cholesterol in Chinese Adult Twins. *Front. Endocrinol.* **9**, 677 (2018).

39. Souren, N. Y. *et al.* Anthropometry, carbohydrate and lipid metabolism in the East Flanders Prospective Twin Survey: heritabilities. *Diabetologia* **50**, 2107–2116 (2007).

40. Poveda, A. *et al.* The heritable basis of gene–environment interactions in cardiometabolic traits. *Diabetologia* **60**, 442–452 (2017).

41. Sulc, J. *et al.* Quantification of the overall contribution of gene-environment interaction for obesity-related traits. *Nat Commun* **11**, 1385 (2020).

42. Dubois, L. *et al.* Genetic and Environmental Contributions to Weight, Height, and BMI from Birth to 19 Years of Age: An International Study of Over 12,000 Twin Pairs. *PLoS ONE* **7**, e30153 (2012).

43. Atanasovska, B., Kumar, V., Fu, J., Wijmenga, C. & Hofker, M. H. GWAS as a Driver of Gene Discovery in Cardiometabolic Diseases. *Trends in Endocrinology & Metabolism* **26**, 722–732 (2015).

44. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).

45. Mahajan, A. *et al.* Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513 (2018).

46. Suzuki, K. *et al.* Identification of 28 new susceptibility loci for type 2 diabetes in the Japanese population. *Nat Genet* **51**, 379–386 (2019).

47. Spracklen, C. N. *et al.* Identification of type 2 diabetes loci in 433,540 East Asian individuals. *Nature* **582**, 240–245 (2020).

48. Wen, W. *et al.* Genome-wide association studies in East Asians identify new loci for waist-hip ratio and waist circumference. *Sci Rep* **6**, 17958 (2016).

49. MAGIC *et al.* Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nat Genet* **42**, 949–960 (2010).

50. Klarin, D. *et al.* Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. *Nat. Genet.* **50**, 1514–1523 (2018).

51. Kanoni, S. *et al. Implicating genes, pleiotropy and sexual dimorphism at blood lipid loci through multi-ancestry meta-analysis*. http://medrxiv.org/lookup/doi/10.1101/2021.12.15.21267852 (2021) doi:10.1101/2021.12.15.21267852.

52. Edwards, S. L., Beesley, J., French, J. D. & Dunning, A. M. Beyond GWASs: Illuminating the Dark Road from Association to Function. *The American Journal of Human Genetics* **93**, 779–797 (2013).

53. Cannon, M. E. & Mohlke, K. L. Deciphering the Emerging Complexities of Molecular Mechanisms at GWAS Loci. *Am. J. Hum. Genet.* **103**, 637–653 (2018).

54. Benjamin, E. J. *et al.* Heart Disease and Stroke Statistics—2018 Update: A Report From the American Heart Association. *Circulation* **137**, (2018).

55. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).

56. Weidemüller, P., Kholmatov, M., Petsalaki, E. & Zaugg, J. B. Transcription factors: Bridge between cell signaling and gene regulation. *Proteomics* **21**, 2000034 (2021).

57. Pascual-Ahuir, A., Fita-Torró, J. & Proft, M. Capturing and Understanding the Dynamics and Heterogeneity of Gene Expression in the Living Cell. *IJMS* **21**, 8278 (2020).

58. Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–394 (2012).

59. Farmer, S. R. Transcriptional control of adipocyte formation. *Cell Metabolism* **4**, 263–273 (2006).

60. Clodfelter, K. H. *et al.* Sex-Dependent Liver Gene Expression Is Extensive and Largely Dependent upon Signal Transducer and Activator of Transcription 5b (STAT5b): STAT5b-Dependent Activation of Male Genes and Repression of Female Genes Revealed by Microarray Analysis. *Molecular Endocrinology* **20**, 1333–1351 (2006).

61. Raulerson, C. K. *et al.* Adipose Tissue Gene Expression Associations Reveal Hundreds of Candidate Genes for Cardiometabolic Traits. *The American Journal of Human Genetics* **105**, 773–787 (2019).

62. Etheridge, A. S. *et al.* A New Liver Expression Quantitative Trait Locus Map From 1,183 Individuals Provides Evidence for Novel Expression Quantitative Trait Loci of Drug Response, Metabolic, and Sex-Biased Phenotypes. *Clin. Pharmacol. Ther.* **107**, 1383–1393 (2020).

63. Nicolae, D. L. *et al.* Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. *PLoS Genet* **6**, e1000888 (2010).

64. Grundberg, E. *et al.* Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat. Genet.* **44**, 1084–1089 (2012).

65. Civelek, M. *et al.* Genetic Regulation of Adipose Gene Expression and Cardio-Metabolic Traits. *Am. J. Hum. Genet.* **100**, 428–443 (2017).

66. Scott, L. J. *et al.* The genetic regulatory signature of type 2 diabetes in human skeletal muscle. *Nat Commun* **7**, 11764 (2016).

67. Strunz, T. *et al.* A mega-analysis of expression quantitative trait loci (eQTL) provides insight into the regulatory architecture of gene expression variation in liver. *Sci Rep* **8**, 5865 (2018).

68. Alasoo, K. *et al.* Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nat. Genet.* (2018) doi:10.1038/s41588-018-0046-7.

69. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–8 (2013).

70. Schmidt, D. *et al.* ChIP-seq: Using high-throughput sequencing to discover protein–DNA interactions. *Methods* **48**, 240–248 (2009).

71. Galhardo, M. *et al.* ChIP-seq profiling of the active chromatin marker H3K4me3 and PPARγ, CEBPα and LXR target genes in human SGBS adipocytes. *Genomics Data* **2**, 230–236 (2014).

72. Nielsen, R. & Mandrup, S. Genome-Wide Profiling of Transcription Factor Binding and Epigenetic Marks in Adipocytes by ChIP-seq. in *Methods in Enzymology* vol. 537 261–279 (Elsevier, 2014).

73. Nielsen, R. *et al.* Genome-wide profiling of PPARγ:RXR and RNA polymerase II occupancy reveals temporal activation of distinct metabolic pathways and changes in RXR dimer composition during adipogenesis. *Genes Dev.* **22**, 2953–2967 (2008).

74. Jou, J. *et al.* The ENCODE Portal as an Epigenomics Resource. *Curr Protoc Bioinformatics* **68**, e89 (2019).

75. Ramaker, R. C. *et al.* A genome-wide interactome of DNA-associated proteins in the human liver. *Genome Res.* **27**, 1950–1960 (2017).

76. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

77. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

78. Cannon, M. E. & Mohlke, K. L. Deciphering the Emerging Complexities of Molecular Mechanisms at GWAS Loci. *The American Journal of Human Genetics* **103**, 637–653 (2018).

79. GTEx Consortium *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

80. Risca, V. I. & Greenleaf, W. J. Unraveling the 3D genome: genomics tools for multiscale exploration. *Trends Genet.* **31**, 357–372 (2015).

81. Ward, L. D. & Kellis, M. Interpreting noncoding genetic variation in complex traits and human disease. *Nat Biotechnol* **30**, 1095–1106 (2012).

82. Musunuru, K. *et al.* From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466**, 714–719 (2010).

83. Perrin, H. J. *et al.* Chromatin accessibility and gene expression during adipocyte differentiation identify context-dependent effects at cardiometabolic GWAS loci. *PLoS Genet* **17**, e1009865 (2021).

84. Cannon, M. E. *et al.* Open Chromatin Profiling in Adipose Tissue Marks Genomic Regions with Functional Roles in Cardiometabolic Traits. *G3* **9**, 2521–2533 (2019).

85. Shungin, D. *et al.* New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187–196 (2015).

86. Allum, F. *et al.* Characterization of functional methylomes by next-generation capture sequencing identifies novel disease-associated variants. *Nat Commun* **6**, 7211 (2015).

87. Siersbaek, R. *et al.* Extensive chromatin remodelling and establishment of transcription factor 'hotspots' during early adipogenesis: Chromatin remodelling during adipogenesis. *The EMBO Journal* **30**, 1459–1472 (2011).

88. Corces, M. R. *et al.* An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14**, 959–962 (2017).

89. Tong, Q. Function of GATA Transcription Factors in Preadipocyte-Adipocyte Transition. *Science* **290**, 134–138 (2000).

90. Zhang, W. *et al.* The TEA domain family transcription factor TEAD4 represses murine adipogenesis by recruiting the cofactors VGLL4 and CtBP2 into a transcriptional complex. *Journal of Biological Chemistry* **293**, 17119–17134 (2018).

91. Seo, J. B. *et al.* Activated Liver X Receptors Stimulate Adipocyte Differentiation through Induction of Peroxisome Proliferator-Activated Receptor ⫽ Expression. **24**, 15 (2004).

92. Lee, R. A., Harris, C. A. & Wang, J.-C. Glucocorticoid Receptor and Adipocyte Biology. *Nuclear Receptor Research* **5**, (2018).

93. Moore, J. E., Pratt, H. E., Purcaro, M. J. & Weng, Z. A curated benchmark of enhancer-gene interactions for evaluating enhancer-target gene prediction methods. *Genome Biol* **21**, 17 (2020).

94. Lenz, M., Arts, I. C. W., Peeters, R. L. M., de Kok, T. M. & Ertaylan, G. Adipose tissue in health and disease through the lens of its building blocks. *Sci Rep* **10**, 10433 (2020).

95. Ambele, M. A., Dessels, C., Durandt, C. & Pepper, M. S. Genome-wide analysis of gene expression during adipogenesis in human adipose-derived stromal cells reveals novel

patterns of gene expression during adipocyte differentiation. *Stem Cell Res* **16**, 725–734 (2016).

96. Hu, E., Liang, P. & Spiegelman, B. M. AdipoQ Is a Novel Adipose-specific Gene Dysregulated in Obesity. *Journal of Biological Chemistry* **271**, 10697–10703 (1996).

97. Pan, D. Z. *et al.* Integration of human adipocyte chromosomal interactions with adipose gene expression prioritizes obesity-related genes from GWAS. *Nat Commun* **9**, 1512 (2018).

98. Garske, K. M. *et al.* Reverse gene-environment interaction approach to identify variants influencing body-mass index in humans. *Nat Metab* **1**, 630–642 (2019).

99. Zivotić, I. *et al.* CDKN2B gene expression is affected by 9p21.3 rs10757278 in CAD patients, six months after the MI. *Clinical Biochemistry* **73**, 70–76 (2019).

100. Hannou, S. A., Wouters, K., Paumelle, R. & Staels, B. Functional genomics of the CDKN2A/B locus in cardiovascular and metabolic disease: what have we learned from GWASs? *Trends in Endocrinology & Metabolism* **26**, 176–184 (2015).

101. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**, 1228–1235 (2015).

102. Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).

103. Lee, Y.-Y. *et al.* Association between risk factors of metabolic syndrome with lung function. *Eur J Clin Nutr* **74**, 811–817 (2020).

104. Chen, Y.-Y. *et al.* Body Fat Percentage in Relation to Lung Function in Individuals with Normal Weight Obesity. *Sci Rep* **9**, 3066 (2019).

105. Yi, Y. H. *et al.* Metabolic syndrome as a risk factor for high intraocular pressure: the Korea National Health and Nutrition Examination Survey 2008&ndash;2010. *DMSO* **Volume 12**, 131–137 (2019).

106. Ahn, M. W., Lee, J. W., Shin, J. H. & Lee, J. S. Relationship between intraocular pressure and parameters of obesity in ocular hypertension. *Int J Ophthalmol* **13**, 794–800 (2020).

107. Wu, Y. *et al.* A meta-analysis of genome-wide association studies for adiponectin levels in East Asians identifies a novel locus near WDR11-FGFR2. *Human Molecular Genetics* **23**, 1108–1119 (2014).

108. Spracklen, C. N. *et al.* Adiponectin GWAS loci harboring extensive allelic heterogeneity exhibit distinct molecular consequences. *PLoS Genet* **16**, e1009019 (2020).

109. Strawbridge, R. J. *et al.* Genome-Wide Association Identifies Nine Common Variants Associated With Fasting Proinsulin Levels and Provides New Insights Into the Pathophysiology of Type 2 Diabetes. **60**, 2624–2634 (2011).

110. Tangirala, R. K. *et al.* Identification of macrophage liver X receptors as inhibitors of atherosclerosis. *Proceedings of the National Academy of Sciences* **99**, 11896–11901 (2002).

111. Wu, J. H. Y. *et al.* Genome-Wide Association Study Identifies Novel Loci Associated With Concentrations of Four Plasma Phospholipid Fatty Acids in the De Novo Lipogenesis Pathway: Results From the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium. *Circ Cardiovasc Genet* **6**, 171–183 (2013).

112. Paton, C. M. & Ntambi, J. M. Biochemical and physiological function of stearoyl-CoA desaturase. *American Journal of Physiology-Endocrinology and Metabolism* **297**, E28–E37 (2009).

113. Xue, A. Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *NATURE COMMUNICATIONS* **9**, 2941 (2018).

114. Dahlman, I. *et al.* Numerous Genes in Loci Associated With Body Fat Distribution Are Linked to Adipose Function. *Diabetes* **65**, 433–437 (2016).

115. Loft, A. *et al.* Browning of human adipocytes requires KLF11 and reprogramming of PPARγ superenhancers. *Genes Dev.* **29**, 7–22 (2015).

116. Schmidt, S. F. *et al.* Acute TNF-induced repression of cell identity genes is mediated by NFκB-directed redistribution of cofactors from super-enhancers. *Genome Res.* **25**, 1281–94 (2015).

117. Banovich, N. E. *et al.* Impact of regulatory variation across human iPSCs and differentiated cells. *Genome Res.* **28**, 122–131 (2018).

118. Calderon, D. *et al.* Landscape of stimulation-responsive chromatin across diverse human immune cells. *Nat. Genet.* **51**, 1494–1505 (2019).

119. Tadjuidje, E. & Hegde, R. S. The Eyes Absent proteins in development and disease. *Cell. Mol. Life Sci.* **70**, 1897–1913 (2013).

120. Fujiwara, S., Baek, S., Varticovski, L., Kim, S. & Hager, G. L. High Quality ATAC-Seq Data Recovered from Cryopreserved Breast Cell Lines and Tissue. *Sci Rep* **9**, 516 (2019).

121. Scharer, C. D. *et al.* ATAC-seq on biobanked specimens defines a unique chromatin accessibility structure in naïve SLE B cells. *Sci Rep* **6**, 27030 (2016).

122. Cannon, M. E. *et al.* Trans-ancestry Fine Mapping and Molecular Assays Identify Regulatory Variants at the ANGPTL8 HDL-C GWAS Locus. *G3 (Bethesda)* **7**, 3217–3227 (2017).

123. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).

124. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

125. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–9 (2012).

126. Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32**, D493-6 (2004).

127. Quinlan, A. R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr Protoc Bioinformatics* **47**, 11.12.1-34 (2014).

128. Varshney, A. *et al.* Genetic regulatory signatures underlying islet gene expression and type 2 diabetes. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 2301–2306 (2017).

129. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).

130. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).

131. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

132. Risso, D., Schwartz, K., Sherlock, G. & Dudoit, S. GC-Content Normalization for RNA-Seq Data. *BMC Bioinformatics* **12**, 480 (2011).

133. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014).

134. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–89 (2010).

135. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).

136. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338 (2019).

137. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

138. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).

139. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res* **4**, 1521 (2015).

140. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).

141. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–9 (2000).

142. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

143. Loh, P.-R., Palamara, P. F. & Price, A. L. Fast and accurate long-range phasing in a UK Biobank cohort. *Nat. Genet.* **48**, 811–816 (2016).

144. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).

145. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

146. Geijn, B. van de, McVicker, G., Gilad, Y. & Pritchard, J. K. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods* **12**, 1061–3 (2015).

147. Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E. & Lappalainen, T. Tools and best practices for data processing in allelic expression analysis. *Genome Biol.* **16**, 195 (2015).

148. Hinrichs, A. S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34**, D590-598 (2006).

149. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).

150. Pulit, S. L. *et al.* Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. *Hum Mol Genet* **28**, 166–174 (2019).

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

152. van der Harst, P. & Verweij, N. Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. *Circ. Res.* **122**, 433–443 (2018).

153. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).

154. International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).

155. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat Genet* **50**, 621–629 (2018).

156. Kim, J. I. *et al.* Lipid-Overloaded Enlarged Adipocytes Provoke Insulin Resistance Independent of Inflammation. *Mol. Cell. Biol.* **35**, 1686–1699 (2015).

157. Shaw, B. *et al.* Individual Saturated and Monounsaturated Fatty Acids Trigger Distinct Transcriptional Networks in Differentiated 3T3-L1 Preadipocytes. *J Nutrigenet Nutrigenomics* **6**, 1–15 (2013).

158. Lo, K. A. *et al.* Analysis of in vitro insulin-resistance models and their physiological relevance to in vivo diet-induced adipose insulin resistance. *Cell Rep* **5**, 259–270 (2013).

159. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* **13**, 613–626 (2012).

160. Do, M.-S. *et al.* Inflammatory Gene Expression Patterns Revealed by DNA Microarray Analysis in TNF-α-treated SGBS Human Adipocytes. *Yonsei Med J* **47**, 729 (2006).

161. Livak, K. J. & Schmittgen, T. D. Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2−ΔΔCT Method. *Methods* **25**, 402–408 (2001).

162. Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC) *et al.* Sex-dimorphic genetic effects and novel loci for fasting glucose and insulin variability. *Nat Commun* **12**, 24 (2021).

163. Lumish, H. S., O’Reilly, M. & Reilly, M. P. Sex Differences in Genomic Drivers of Adipose Distribution and Related Cardiometabolic Disorders: Opportunities for Precision Medicine. *ATVB* **40**, 45–60 (2020).

164. Rask-Andersen, M., Karlsson, T., Ek, W. E. & Johansson, Å. Genome-wide association study of body fat distribution identifies adiposity loci and sex-specific genetic effects. *Nat Commun* **10**, 339 (2019).

165. Karaderi, T., Drong, A. W. & Lindgren, C. M. Insights into the Genetic Susceptibility to Type 2 Diabetes from Genome-Wide Association Studies of Obesity-Related Traits. *Curr Diab Rep* **15**, 83 (2015).

166. Zillikens, M. C. *et al.* Sex-specific genetic effects influence variation in body composition. *Diabetologia* **51**, 2233–2241 (2008).

167. Çalışkan, M. *et al.* Genetic and Epigenetic Fine Mapping of Complex Trait Associated Loci in the Human Liver. *Am. J. Hum. Genet.* **105**, 89–107 (2019).

168. Soldin, O. P. & Mattison, D. R. Sex Differences in Pharmacokinetics and Pharmacodynamics: *Clinical Pharmacokinetics* **48**, 143–157 (2009).

169. Tamargo, J. *et al.* Gender differences in the effects of cardiovascular drugs. *European Heart Journal - Cardiovascular Pharmacotherapy* **3**, 163–182 (2017).

170. Lopes-Ramos, C. M. Sex Differences in Gene Expression and Regulatory Networks across 29 Human Tissues. *OPEN ACCESS* 17.

171. Kukurba, K. R. *et al.* Impact of the X Chromosome and sex on regulatory variation. 10.

172. Sugathan, A. & Waxman, D. J. Genome-Wide Analysis of Chromatin States Reveals Distinct Mechanisms of Sex-Dependent Gene Regulation in Male and Female Mouse Liver. *Molecular and Cellular Biology* **33**, 17 (2013).

173. Jermendy, G. *et al.* Effect of genetic and environmental influences on cardiometabolic risk factors: a twin study. *Cardiovasc Diabetol* **10**, 96 (2011).

174. Jun, G. *et al.* Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am. J. Hum. Genet.* **91**, 839–48 (2012).

175. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)* **57**, 289–300 (1995).

176. Márquez, E. J. *et al.* Sexual-dimorphism in human immune system aging. *Nat Commun* **11**, 751 (2020).

177. Kanai, M. *et al.* Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat Genet* **50**, 390–400 (2018).

178. Finney, A. C. *et al.* EphA2 Expression Regulates Inflammation and Fibroproliferative Remodeling in Atherosclerosis. *Circulation* **136**, 566–582 (2017).

179. Alisi, A. *et al.* Mirnome analysis reveals novel molecular determinants in the pathogenesis of diet-induced nonalcoholic fatty liver disease. *Lab Invest* **91**, 283–293 (2011).

180. Khan, Md. W., Priyadarshini, M., Cordoba-Chacon, J., Becker, T. C. & Layden, B. T. Hepatic hexokinase domain containing 1 (HKDC1) improves whole body glucose tolerance and insulin sensitivity in pregnant mice. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* **1865**, 678–687 (2019).

181. Hayes, M. G. *et al.* Identification of *HKDC1* and *BACE2* as Genes Influencing Glycemic Traits During Pregnancy Through Genome-Wide Association Studies. *Diabetes* **62**, 3282–3291 (2013).

182. Currin, K. W. *et al.* Genetic effects on liver chromatin accessibility identify disease regulatory variants. *The American Journal of Human Genetics* **108**, 1169–1189 (2021).

183. Jung, I. *et al.* A compendium of promoter-centered long-range chromatin interactions in the human genome. *Nat Genet* **51**, 1442–1449 (2019).

184. Aizarani, N. *et al.* A human liver cell atlas reveals heterogeneity and epithelial progenitors. *Nature* **572**, 199–204 (2019).

185. Arnold, M., Raffler, J., Pfeufer, A., Suhre, K. & Kastenmüller, G. SNiPA: an interactive, genetic variant-centered annotation browser. *Bioinformatics* **31**, 1334–1336 (2015).

186. Roman, T. S. *et al.* Multiple Hepatic Regulatory Variants at the GALNT2 GWAS Locus Associated with High-Density Lipoprotein Cholesterol. *The American Journal of Human Genetics* **97**, 801–815 (2015).

187. Pulecio, J., Verma, N., Mejía-Ramírez, E., Huangfu, D. & Raya, A. CRISPR/Cas9-Based Engineering of the Epigenome. *Cell Stem Cell* **21**, 431–447 (2017).

188. Zhang, F., Wen, Y. & Guo, X. CRISPR/Cas9 for genome editing: progress, implications and challenges. *Human Molecular Genetics* **23**, R40–R46 (2014).

189. Simeonov, D. R. *et al.* Discovery of stimulation-responsive immune enhancers with CRISPR activation. *Nature* **549**, 111–115 (2017).

190. Heinz, S. *et al.* Effect of natural genetic variation on enhancer selection and function. *Nature* **503**, 487–492 (2013).

191. Klein, J. C. *et al.* A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nat Methods* **17**, 1083–1091 (2020).

192. Starita, L. M. *et al.* Variant Interpretation: Functional Assays to the Rescue. *The American Journal of Human Genetics* **101**, 315–325 (2017).

193. Rai, V. *et al.* Single-cell ATAC-Seq in human pancreatic islets and deep learning upscaling of rare cells reveals cell-specific type 2 diabetes regulatory signatures. *Mol Metab* **32**, 109–121 (2020).

194. Ruf-Zamojski, F. *et al.* Single nucleus multi-omics regulatory landscape of the murine pituitary. *Nat Commun* **12**, 2677 (2021).

195. Ziffra, R. S. *et al.* Single-cell epigenomics reveals mechanisms of human cortical development. *Nature* **598**, 205–213 (2021).