CHROMATIN ACCESSIBILITY CHANGES AND GENOMIC INTEGRATION IDENTIFY GENETIC REGULATORY MECHANISMS

Kevin Williams Currin

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 Bioinformatics and Computational Biology program in the School of Medicine.

Chapel Hill
2020

Approved by:

Karen L. Mohlke

Terrence S. Furey

Yun Li

Praveen Sethupathy

Jason Stein

ABSTRACT

Kevin Williams Currin: Chromatin accessibility changes and genomic integration identify genetic regulatory mechanisms
(Under the direction of Karen L. Mohlke)

Cardiovascular disease, type 2 diabetes, and related cardiometabolic traits are responsible for substantial mortality and economic costs globally. Cardiometabolic traits are common complex phenotypes with both genetic and environmental components and are influenced by several tissues, including adipose, liver, skeletal muscle, and pancreas. Genome-wide association studies (GWAS) have identified thousands of loci associated with cardiometabolic traits, but most of these loci are in noncoding regions of the genome and their molecular functions are not well annotated. Variants at GWAS loci are often found within transcriptional regulatory elements and/or are associated with gene expression in trait-relevant tissues, suggesting that GWAS variants frequently alter gene regulation. Transcriptional regulatory elements vary by genotype, tissue, and cellular state, but regulatory element annotation in many of these disease-relevant contexts is lacking. A more complete annotation of regulatory elements may uncover mechanisms for GWAS loci for cardiometabolic traits. To this end, I profiled chromatin accessibility, a marker of regulatory elements, in adipose tissue, liver tissue, and multiple stages of adipocyte differentiation. The accessible chromatin landscape in adipose tissue was underannotated and our profiles helped identify GWAS variants that may alter adipose gene regulation. I identified accessible chromatin regions that vary by genotype in liver tissue, providing suggestive evidence that these variants alter regulatory element activity. The accessible chromatin regions that differ between stages of adipocyte differentiation suggest specific cellular states in which GWAS variants may alter gene regulation. I integrated accessible chromatin regions with multiple genomic data types to predict functional variants, disrupted TF binding motifs, and target genes at cardiometabolic GWAS loci. Variants at several loci showed allelic differences in transcriptional reporter and protein binding assays, providing further evidence of regulatory function. My findings contribute to the understanding of which variants, regulatory elements, and genes influence cardiometabolic traits. These predicted functional variants, regulatory elements, and target genes are strong candidates for testing in functional assays and may help guide therapeutic strategies for cardiometabolic diseases.

ACKNOWLEDGEMENTS

I am extremely grateful to the many people who have helped me reach this milestone. I would first like to thank my PhD mentor, Karen Mohlke. Thank you for helping me grow as a scientist, considering my many (and often tangential) scientific ideas, allowing me to pursue professional development activities outside of the lab, encouraging me to have the best poster or slides possible (even when I was fed up with formatting), the countless hours helping me with formatting, and for being an excellent and supportive mentor overall. Thank you for vastly improving my presentation skills by frequently encouraging me to present to diverse audiences. Thank you for always being understanding and accommodating during the several tough life events that I experienced during graduate school. I have learned so much from you about how to refine research ideas, conduct research, write, and present. You are an excellent scientist and mentor.

I would next like to thank my dissertation committee: Terry Furey, Praveen Sethupathy, Yun Li, and Jason Stein. First, thank you all for your continual encouragement and for answering my countless emails. I have learned a lot about science and mentorship from each of you. Terry, you taught me a ton about genomics and computational analyses beginning back when you were an instructor for one of my undergraduate classes until now when I send you constant emails about papers and analyses. Praveen, you were an excellent and supportive mentor during my PREP year, where you taught me so much about miRNAs, genomics, and computational analyses. Thank you also for your very thoughtful and helpful career discussions at that critical stage of my career. Yun, thank you for the countless times you helped me with statistical questions and for suggesting helpful analyses. Jason, thank you for helping with ATAC-seq interpretation, computational analyses in general, and paper feedback. Thank you all for serving on my committee and for helping me achieve this awesome goal.

I would next like to thank my undergraduate research mentor Alain Laederach and members of the Laederach lab. Alain, thank you for introducing me to bioinformatics research and for teaching me about RNA secondary structure. Meredith, thank you for teaching me bash programming and advising me on one of my projects. Matt, thank you for helping with computational analyses. I truly enjoyed working with all of you.

I would next like to thank Maren Cannon and Hannah Perrin, with whom I most closely worked during my graduate research. Maren, the work presented in CHAPTER 2 in this dissertation would not have been possible without your contributions to ATAC-seq data generation, data analysis, and functional experiments. Thank you for sharing this project with me and helping me to grow as a scientist. Thank you also for your contributions to generating and analyzing SGBS ATAC-seq data, which helped make the work in CHAPTER 4 possible. Thank you for the countless hours you spent helping get the work in CHAPTER 2 accepted for publication, which turned out to be a particularly challenging process. I really appreciate your help with this even after you graduated. Hannah, thank you for your numerous contributions to all three projects presented in this dissertation. Thank you for generating SGBS ATAC-seq libraries and performing functional experiments for the work in CHAPTER 2, for providing valuable interpretation and suggestions for the work presented in CHAPTER 3, and for generating ATAC-seq and RNA-seq libraries, performing numerous computational analyses, and performing functional experiments for the work presented in CHAPTER 4. The work presented in CHAPTER 4 would not have been possible without your contributions. Thank you for the countless hours you spent optimizing ATAC-seq library generation, assessing data quality, and brainstorming project ideas. Thank you for helping me grow as a scientist, particularly through the hard times when it felt like all of our project ideas were failing. I would not be the scientist I am today without both of your help. Thank you.

I would next like to thank the Mohlke lab members and affiliates. All of you have provided valuable help that helped me achieve my goal of obtaining a PhD. Raulie, thank you for helping me grow as a computational scientist, allowing me to contribute to your adipose eQTL project, and for providing data for the work presented in CHAPTER 4. Thank you for honoring Riza’s working status, even though her proximity must have made it really difficult. Cassie, thank you for your help with statistical analyses and for allowing me to contribute to your adiponectin project. Thank you also for beverage recommendations and for helping me out during a difficult personal time. Jim, thank you for your helpful analysis suggestions and for setting up my standing desk, which I should have used more often. Rani, thank you for all of your valuable work on all projects presented in this dissertation, including but not limited to functional experiments and help with ATAC-seq and RNA-seq library generation. Thank you for regularly checking in to see if I needed anything. Ying, thank you for providing GWAS and eQTL data for the work presented in CHAPTER 2 and for helpful feedback during lab meetings. Sarah, thank you for your help with RNA-seq and genotype analyses. Thank you also for all of the formatting help on my presentations and for helping me find presentation rooms and set up projectors. Alaine, thank you for contributing eQTL data for the work in CHAPTER 3, helping with locusZoom plots, and for answering my statistical questions. Kris, Thank you for performing functional experiments presented in CHAPTER 2, helping me with genotype and GWAS questions, and for suggesting helpful analyses. Shelley, thank you for suggesting helpful analyses and for helping format this dissertation. Laura, thank you for helping with genotype, imputation, and GWAS analyses. John, thank you for helping me with statistical analyses and for your lab meeting presentations that taught me about mediation analyses. Apoorva, thank you for your presentations that taught me about various types of functional studies. Gautam, thank you for the literature review you performed for CHAPTER 3 and for suggesting analyses. Sophie, thank you for your presentations that taught me about single nucleus ATAC-seq and RNA-seq and for providing helpful feedback during lab meetings. Tori, thank you for answering my questions about CRISPR and for your presentations that taught me about CRISPR and other functional experiments. Tamara, thank you for your presentations that taught me about functional analyses and for helpful feedback during lab meetings. Emma, thank you for your presentations that taught me about GWAS meta-analyses. Martin, thank you for developing the AA-ALIGNER pipeline that we used in CHAPTER 2. Kenneth, thank you for generating ATAC-seq and RNA-seq libraries for the work in CHAPTER 4. Thanks to everyone in the lab for all of your help and support, including navigating large conferences. Thank all of you for celebrating graduate school milestones with me, including passing exams and publishing papers. Thank all of you for your thoughtful gifts for my dissertation defense.

I would next like to thank scientific colleagues outside of the lab. Bryan Quach, thank you for your help with various bioinformatic analyses and for giving helpful advice on job opportunities. Dan Liang, thank you for helping with various ATAC-seq analyses. Thank you to many others in the UNC bioinformatics community, including but not limited to Jeremy Simon, Mike Love, Spencer Nystrom, and Austin Hepperla, for helping with ATAC-seq troubleshooting, statistical analyses, and data interpretation. Thank you to members of the Parker, Collins, and Innocenti labs, including but not limited to Ricardo Albanus, Peter Orchard, Vivek Rai, Arushi Varshney, Stephen Parker, Mike Erdos, Narisu Narisu, Lori Bonnycastle, Francis Collins, Amy Etheridge, and Federico Innocenti, for all of your help with the work presented in CHAPTER 3, including sample coordination, ATAC-seq and RNA-seq library generation, and data analysis.

I would next like to thank friends who have helped me along my academic journey. Ken Walsh, thank you for tutoring me in elementary and middle school, giving feedback on college applications, and for always holding me to a high standard. Thank you for introducing me to hiking and giving me my first set of trekking poles, which I still use. Alan Chase, thank you for encouraging me to go to graduate school and for always being there when I needed someone to talk to. Thank you to my high school VI teacher Gale Waters and braillist Carol Smith; the two of you really helped me do well in high school and subsequently get into a good college. Thank you to my orientation and mobility teacher Dion Ousley; you helped me become a confident traveler. Thank you to the SAS Accessibility Team, including but not limited to Ed Summers, Brice Smith, Sean Mealin, Jesse Sookne, Julianna Langston, Lisa Morton, Greg Kraus, and Tyler Williamson, for allowing me to work with you all during my summer internship. I enjoyed working with all of you and truly appreciate the career support and friendship you all have given me. Thank you to my middle school science teacher Mrs. Register and my high school science teachers Ms. Capps, Ms. Top, and Mrs. Johnson; you all inspired me to become a scientist. Thank you to my undergraduate genetics professors Blaire Steinwand, Joseph Kieber, Jeff Sekelsky, Greg Copenhaver, and Corbin Jones for inspiring my love of genetics. Thank you to John Cornett, Cara Marlow, Cathy Cornett, and Jackie Boyden for all the help navigating all of the graduate school paperwork, sorting out paychecks, scheduling rooms for presentations, setting up seminars, handling reimbursements, and being there to help answer any questions or provide support.

I would next like to thank my family for all of there support. Thank you to my grandparents Peggy and Warren West for driving me to school, advocating for me in IEP meetings, and always making sure I had enough to eat. Thank you to my grandparents Perry and Martha Currin for always supporting me, making sure I had enough to eat, and attending my school performances and graduations. Thank you to my brother Derrick Currin for supporting me, driving me to school, and holding me to a high standard. Thank you to my mom Joan and Wayne, aunt Patsy, cousins Adrian and Preston, aunt Tammy, cousin Jennifer, and everyone else. You all have helped me so many ways throughout my life and I am extremely grateful.

Finally, I would like to thank my wife Laralee. You have been extremely supportive through graduate school and so many other parts of life. Thank you for making sure that I take breaks from work, eat good food, and go outside, even when things are really busy. Thank you for standing by me during the hard times we have faced in the past few years. I couldn’t have asked for a better person to spend my life with. Thank you for encouraging me to achieve my goals and to become the best person I can be. I love you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AI: allelic imbalance

ATAC-seq: assay for transposase-accessible chromatin followed by sequencing

BH: Benjamini-Hochberg

BMI: body mass index

Bp: base pair

caQTL: chromatin accessibility quantitative trait loci

ChIP-seq: chromatin immunoprecipitation followed by sequencing

CVD: cardiovascular disease

EMSA: electrophoretic mobility shift assay

ENCODE: Encyclopedia of DNA Elements

eQTL: expression quantitative trait loci

FDR: false discovery rate

GWAS: genome-wide association study

HDL: high density lipoprotein

kb: kilobase

LD: linkage disequilibrium

LDL: low density lipoprotein

MACS2: Model-based Analysis of ChIP-seq version 2

Mb: megabase

METSIM: METabolic Syndrome in Men study

NIH: National Institutes of Health

PCA: principal component analysis

RASQUAL: Robust Allele-Specific Quantitation and Quality Control

RNA-seq: ribonucleic acid sequencing

SGBS: Simpson Golabi-Behmel Syndrome

T2D: type 2 diabetes

TC: total cholesterol

TF: transcription factor

TG: triglycerides

TSS: transcription start site

WHRadjBMI: waist-hip ratio adjusted for body mass index

CHAPTER 1: INTRODUCTION

Overview of cardiometabolic traits

Cardiovascular disease (CVD), type 2 diabetes (T2D), and related risk factors, termed cardiometabolic traits, account for substantial health and economic costs worldwide. CVD is the leading cause of death in the United States and worldwide, and was responsible for 17.8 million deaths globally in 2017[1,2]. An estimated 485.6 million people have CVD globally and the direct and indirect annual economic costs of CVD in the United States were estimated to be $351.3 billion from 2014-2015[1]. T2D was the seventh leading cause of death in the United States as of 2016[1] and was responsible for 1.5 million deaths worldwide in 2012[2]. In the United States, an estimated 26 million adults have T2D and 91.8 million have prediabetes[1]. Several cardiometabolic traits are risk factors for CVD and T2D, including obesity, insulin resistance, hypertension, high triglycerides, and high LDL cholesterol[1,3,4], and T2D itself is a risk factor for CVD[5].

Several tissues influence cardiometabolic traits, including adipose, liver, skeletal muscle, and pancreas. Adipose tissue influences cardiometabolic traits through its roles in lipid storage and hormone secretion[6,7]. Decreased lipid storage capacity in subcutaneous adipose tissue leads to increased lipid storage and increased insulin resistance in skeletal muscle, visceral adipose, and liver[6]. Adipose tissue secretes numerous hormones including leptin, which regulates energy intake and several other metabolic processes, and adiponectin, which is negatively associated with T2D and may protect against insulin resistance[7]. The liver regulates glucose, lipid, and cholesterol availability through multiple mechanisms[8]. The liver stores excess glucose as glycogen in response to increased insulin levels and secretes glucose through gluconeogenesis to provide energy to other tissues when insulin levels decrease[8]. Insulin resistance in the liver leads to increased hepatic glucose secretion and increased hepatic lipid accumulation, which is associated with the development of CVD, T2D, and nonalcoholic fatty liver disease[8]. Insulin resistance in skeletal muscle, resulting in reduced glucose uptake and metabolism, is associated with T2D, obesity, hypertension, and other conditions[9]. Impaired insulin secretion by pancreatic beta cells contributes to the development of T2D alongside insulin resistance in other tissues[10].

Genetics of cardiometabolic traits

Cardiometabolic traits are complex and have both genetic and environmental components. Risk for developing CVD, T2D, hypertension, and other cardiometabolic conditions is associated with environmental factors such as nutrition, physical activity, smoking, and stress[1,3]. Both rare and common genetic variation contributes to risk for CVD, T2D, and related risk factors[1,11,12]. A twin study estimated the heritability of death from coronary heart disease to be 38% in women and 57% in men[13], and individuals with a sibling with CVD are at an increased risk of CVD even after correcting for other risk factors[14]. Single gene mutations have been identified that drive familial hypercholesterolemia, a rare condition resulting in increased LDL cholesterol and CVD risk[15–17]. The heritability of T2D varies with the age of onset and is estimated to be 31-72%[18,19]. Genome-wide association studies (GWAS) have been instrumental in identifying genetic variation influencing common traits[20]. GWAS have identified thousands of genetic associations with cardiometabolic traits, including 161 loci for coronary artery disease[21], 403 distinct genetic signals at 243 loci for T2D[22], 901 loci for blood pressure traits[23], 941 distinct signals at 536 loci for body mass index (a measure of obesity)[24], and 826 distinct signals at 386 loci for blood lipid traits[25]. Identified GWAS signals explain 18% of T2D risk[22], 5.7% of variation in systolic blood pressure[23], and 6.0% of variation in body mass index[24]. While GWAS are instrumental in identifying the genetic basis of common complex traits, they do not identify which genetic variants at a signal are functional and how these variants impact the trait[20,26]. Additional approaches are needed to identify the functional variant/s at a signal, the gene/s influenced by the variants, the tissue in which the variant acts, and the mechanism of action[20,26].

Identifying target genes and tissues at GWAS loci

There are multiple approaches to linking GWAS variants to genes, including coding variants, gene expression quantitative trait locus (eQTL) mapping, and chromatin conformation capture (3C) techniques[26]. Rare variants at some GWAS loci are found within coding regions and are predicted to alter protein function, but most GWAS loci do not have coding variants[26]. eQTL are identified by associating genetic variants with gene expression levels across individuals in a given tissue[27]. Target genes at GWAS loci can be predicted by identifying shared, or colocalized, eQTL and GWAS signals[27–29]. Multiple methods exist for identifying colocalized GWAS and eQTL signals, including linkage disequilibrium (LD) between GWAS and eQTL lead variants, examining the eQTL association after conditioning on the GWAS lead (termed conditional analysis), and methods that compare GWAS and eQTL summary statistics, such as COLOC and eCAVIAR[26,28–31]. Colocalization of GWAS and eQTL signals has been used to predict target genes at cardiometabolic trait GWAS loci in trait-relevant tissues, including genes at body fat distribution and lipid GWAS loci in adipose tissue[28], lipids in liver[29,32], and glucose and T2D in skeletal muscle[33].

Chromosomes form loops that form contacts between different genomic regions, including those far apart based on linear DNA sequence[34]. The 3C technique identifies a pair of genomic regions that physically interact with one another, and the high-throughput version, Hi-C, identifies pairs of interacting regions genome-wide[34]. Promoter capture Hi-C involves specifically selecting interactions with gene promoters, which is particularly use for linking distal regulatory elements to genes[34,35]. Chromatin interactions vary by cell and tissue type and have been used to link GWAS variants to gene promoters[35].

Many methods that predict target genes at GWAS loci, such as eQTL and 3C, only provide suggestive evidence that a variant alters a gene and additional experiments are needed to prove a causal relationship. Variant-gene links predicted by multiple methods are more likely functional and are strong candidates for downstream experiments. While eQTL suggest that a variant may alter gene expression, additional experiments are needed to determine the mechanisms by which genetic variants alter gene expression.

Predicting functional variants and mechanisms using transcriptional regulatory elements

Individual studies and large-scale efforts, such as ENCODE[36] and the NIH Roadmap Epigenomics Project[37], have made great progress in mapping transcriptional regulatory elements of genes, such as promoters, enhancers, silencers, and insulators, but many cell and tissue types remain under-annotated. ENCODE has performed thousands of experiments to map multiple aspects of regulatory element activity in hundreds of cell and tissue types, including accessible chromatin regions, binding sites for transcription factors (TFs) and other proteins, DNA methylation, and histone modifications[36]. The Roadmap Epigenomics Project mapped chromatin accessibility, histone modifications, and DNA methylation in 111 cell and tissue types and integrated these data types to identify regulatory chromatin states, such as promoter, enhancer, transcribed, and repressed states[37]. Chromatin accessibility is a canonical feature of active and poised regulatory elements[36], and is thus broadly useful in mapping regulatory elements. However, the accessible chromatin landscapes of many human cardiometabolic-relevant tissues, such as adipose and liver, remain under-annotated. At the time of analysis of the data presented in CHAPTER 2, only three chromatin accessibility profiles existed from primary human adipose tissue: one in whole subcutaneous adipose tissue[36], one in whole omental adipose tissue[36], and one in adipocytes isolated from adipose tissue[38]. Additional accessible chromatin profiles have been mapped in adipocytes differentiated from cell models, including human multipotent adipose-derived stem cells[39] and the Simpson Golabi-Behmel Syndrome (SGBS) cell strain[40,41]. Liver chromatin accessibility has mainly been profiled in the HepG2 hepatoblastoma cell line, which is immortalized and aneuploid[36,42–44]. A limited number of primary tissue samples have been generated by ENCODE[36]. Mapping chromatin accessibility in liver, adipose, and other cardiometabolic-relevant tissues is needed to characterize transcriptional regulation in these tissues.

Several lines of evidence suggest that genetic variants, including those at GWAS loci, can alter regulatory element activity. First, activity of many regulatory elements is heritable[45]. Second, GWAS loci are over-represented in regulatory elements of tissues relevant to the GWAS trait[37], including body fat distribution loci in adipose tissue regulatory elements[46] and T2D loci in skeletal muscle[33]. Third, genetic variants associated with levels of chromatin accessibility, termed chromatin accessibility QTL (caQTL), have been identified in multiple tissues, a subset of which are colocalized with GWAS loci and eQTL[47–52], suggesting that genetic variants may mediate effects on gene expression and GWAS traits by altering chromatin accessibility. Consequently, mapping chromatin accessibility in cardiometabolic-relevant tissues in multiple individuals may uncover functional variants that alter chromatin accessibility and ultimately impact GWAS traits.

Regulatory element activity also differs between environmental contexts. Chromatin accessibility levels and regulatory element chromatin states vary between different tissue and cell types[36,37], including across cellular differentiation[53–55] and across different cell types of the same tissue[56,57]. Various stimuli can alter chromatin accessibility, including inflammation[58,59], dietary lipids[53], and hypoxia[60]. Context-dependent genetic effects on chromatin accessibility have also been identified across immune cell activation[49,58]. Consequently, mapping chromatin accessibility in a variety of disease-relevant cellular contexts may identify GWAS variants with context-specific effects on chromatin and GWAS traits.

Aims and overview

In this dissertation, I contribute to our understanding of how genetic variation impacts chromatin accessibility and disease. In CHAPTER 2 I describe chromatin accessibility profiles in three adipose tissue samples and in replicates from SGBS preadipocytes and adipocytes using ATAC-seq. I show that GWAS variants for body fat distribution, cholesterol, and other cardiometabolic traits are overrepresented in adipose chromatin accessibility. I use adipose chromatin accessibility to predict functional variants at colocalized GWAS and adipose tissue eQTL. In

CHAPTER 3 I present caQTL in liver tissue using genotypes and chromatin accessibility data from 20 individuals. I integrate caQTL with multiple genomic data types, such as eQTL and TF motifs, to predict functional variants, target genes, and mechanisms at GWAS loci. In CHAPTER 4 I describe differences in chromatin accessibility and gene expression between states of adipocyte differentiation. I use these data to predict GWAS variants that may act in different cellular states in adipose tissue. In CHAPTER 5 I summarize my findings, reflect on what I have learned, consider limitations, and discuss how my work fits into cardiometabolic research overall.

REFERENCES

1. Virani, S.S., Alonso, A., Benjamin, E.J., Bittencourt, M.S., Callaway, C.W., Carson, A.P., Chamberlain, A.M., Chang, A.R., Cheng, S., Delling, F.N., et al. (2020). Heart Disease and Stroke Statistics-2020 Update: A Report From the American Heart Association. Circulation 141, e139–e596.

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

3. Wu, Y., Ding, Y., Tanaka, Y., and Zhang, W. (2014). Risk factors contributing to type 2 diabetes and recent advances in the treatment and prevention. Int J Med Sci 11, 1185–1200.

4. Rochlani, Y., Pothineni, N.V., Kovelamudi, S., and Mehta, J.L. (2017). Metabolic syndrome: pathophysiology, management, and modulation by natural compounds. Ther Adv Cardiovasc Dis 11, 215–225.

5. Emerging Risk Factors Collaboration, Sarwar, N., Gao, P., Seshasai, S.R.K., Gobin, R., Kaptoge, S., Di Angelantonio, E., Ingelsson, E., Lawlor, D.A., Selvin, E., et al. (2010). Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. Lancet 375, 2215–2222.

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

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

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

9. Abdul-Ghani, M.A., and DeFronzo, R.A. (2010). Pathogenesis of insulin resistance in skeletal muscle. J. Biomed. Biotechnol. 2010, 476279.

10. Kahn, S.E. (2003). The relative contributions of insulin resistance and beta-cell dysfunction to the pathophysiology of Type 2 diabetes. Diabetologia 46, 3–19.

11. Kathiresan, S., and Srivastava, D. (2012). Genetics of human cardiovascular disease. Cell 148, 1242–1257.

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

13. Zdravkovic, S., Wienke, A., Pedersen, N.L., Marenberg, M.E., Yashin, A.I., and De Faire, U. (2002). Heritability of death from coronary heart disease: a 36-year follow-up of 20 966 Swedish twins. J. Intern. Med. 252, 247–254.

14. Murabito, J.M. (2005). Sibling Cardiovascular Disease as a Risk Factor for Cardiovascular Disease in Middle-aged Adults. JAMA 294, 3117.

15. Lehrman, M.A., Schneider, W.J., Südhof, T.C., Brown, M.S., Goldstein, J.L., and Russell, D.W. (1985). Mutation in LDL receptor: Alu-Alu recombination deletes exons encoding transmembrane and cytoplasmic domains. Science 227, 140–146.

16. Soria, L.F., Ludwig, E.H., Clarke, H.R., Vega, G.L., Grundy, S.M., and McCarthy, B.J. (1989). Association between a specific apolipoprotein B mutation and familial defective apolipoprotein B-100. Proc. Natl. Acad. Sci. U.S.A. 86, 587–591.

17. Abifadel, M., Varret, M., Rabès, J.-P., Allard, D., Ouguerram, K., Devillers, M., Cruaud, C., Benjannet, S., Wickham, L., Erlich, D., et al. (2003). Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. Nat. Genet. 34, 154–156.

18. Almgren, P., Lehtovirta, M., Isomaa, B., Sarelin, L., Taskinen, M.R., Lyssenko, V., Tuomi, T., Groop, L., and Botnia Study Group (2011). Heritability and familiality of type 2 diabetes and related quantitative traits in the Botnia Study. Diabetologia 54, 2811–2819.

19. Willemsen, G., Ward, K.J., Bell, C.G., Christensen, K., Bowden, J., Dalgård, C., Harris, J.R., Kaprio, J., Lyle, R., Magnusson, P.K.E., et al. (2015). The Concordance and Heritability of Type 2 Diabetes in 34,166 Twin Pairs From International Twin Registers: The Discordant Twin (DISCOTWIN) Consortium. Twin Res Hum Genet 18, 762–771.

20. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. Am. J. Hum. Genet. 101, 5–22.

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

22. Mahajan, A., Taliun, D., Thurner, M., Robertson, N.R., Torres, J.M., Rayner, N.W., Payne, A.J., Steinthorsdottir, V., Scott, R.A., Grarup, N., et al. (2018). Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nat. Genet. 50, 1505–1513.

23. Evangelou, E., Warren, H.R., Mosen-Ansorena, D., Mifsud, B., Pazoki, R., Gao, H., Ntritsos, G., Dimou, N., Cabrera, C.P., Karaman, I., et al. (2018). Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. Nat. Genet. 50, 1412–1425.

24. Yengo, L., Sidorenko, J., Kemper, K.E., Zheng, Z., Wood, A.R., Weedon, M.N., Frayling, T.M., Hirschhorn, J., Yang, J., Visscher, P.M., et al. (2018). Meta-analysis of genome-wide association studies for height and body mass index in ∼700000 individuals of European ancestry. Hum. Mol. Genet. 27, 3641–3649.

25. Klarin, D., Damrauer, S.M., Cho, K., Sun, Y.V., Teslovich, T.M., Honerlaw, J., Gagnon, D.R., DuVall, S.L., Li, J., Peloso, G.M., et al. (2018). Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. Nat. Genet. 50, 1514–1523.

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

27. GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC):, Biospecimen Collection Source Site—RPCI, ELSI Study, Lead analysts:, NIH program management:, Biospecimen collection:, Pathology:, eQTL manuscript working group:, Battle, A., et al. (2017). Genetic effects on gene expression across human tissues. Nature 550, 204–213.

28. Raulerson, C.K., Ko, A., Kidd, J.C., Currin, K.W., Brotman, S.M., Cannon, M.E., Wu, Y., Spracklen, C.N., Jackson, A.U., Stringham, H.M., et al. (2019). Adipose Tissue Gene Expression Associations Reveal Hundreds of Candidate Genes for Cardiometabolic Traits. Am. J. Hum. Genet. 105, 773–787.

29. Etheridge, A.S., Gallins, P.J., Jima, D., Broadaway, K.A., Ratain, M.J., Schuetz, E., Schadt, E., Schroder, A., Molony, C., Zhou, Y., et al. (2019). 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.

30. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 10, e1004383.

31. Hormozdiari, F., van de Bunt, M., Segrè, A.V., Li, X., Joo, J.W.J., Bilow, M., Sul, J.H., Sankararaman, S., Pasaniuc, B., and Eskin, E. (2016). Colocalization of GWAS and eQTL Signals Detects Target Genes. Am. J. Hum. Genet. 99, 1245–1260.

32. Çalışkan, M., Manduchi, E., Rao, H.S., Segert, J.A., Beltrame, M.H., Trizzino, M., Park, Y., Baker, S.W., Chesi, A., Johnson, M.E., et al. (2019). Genetic and Epigenetic Fine Mapping of Complex Trait Associated Loci in the Human Liver. Am. J. Hum. Genet. 105, 89–107.

33. Scott, L.J., Erdos, M.R., Huyghe, J.R., Welch, R.P., Beck, A.T., Wolford, B.N., Chines, P.S., Didion, J.P., Narisu, N., Stringham, H.M., et al. (2016). The genetic regulatory signature of type 2 diabetes in human skeletal muscle. Nat Commun 7, 11764.

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

35. Jung, I., Schmitt, A., Diao, Y., Lee, A.J., Liu, T., Yang, D., Tan, C., Eom, J., Chan, M., Chee, S., et al. (2019). A compendium of promoter-centered long-range chromatin interactions in the human genome. Nat. Genet. 51, 1442–1449.

36. ENCODE Project Consortium, Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shoresh, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A., et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature 583, 699–710.

37. Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature 518, 317.

38. Allum, F., Shao, X., Guénard, F., Simon, M.-M., Busche, S., Caron, M., Lambourne, J., Lessard, J., Tandre, K., Hedman, Å.K., et al. (2015). Characterization of functional methylomes by next-generation capture sequencing identifies novel disease-associated variants. Nat Commun 6, 7211.

39. Loft, A., Forss, I., Siersbæk, M.S., Schmidt, S.F., Larsen, A.-S.B., Madsen, J.G.S., Pisani, D.F., Nielsen, R., Aagaard, M.M., Mathison, A., et al. (2015). Browning of human adipocytes requires KLF11 and reprogramming of PPARγ superenhancers. Genes Dev. 29, 7–22.

40. Schmidt, S.F., Larsen, B.D., Loft, A., Nielsen, R., Madsen, J.G.S., and Mandrup, S. (2015). Acute TNF-induced repression of cell identity genes is mediated by NFκB-directed redistribution of cofactors from super-enhancers. Genome Res. 25, 1281–1294.

41. Fischer-Posovszky, P., Newell, F.S., Wabitsch, M., and Tornqvist, H.E. (2008). Human SGBS cells - a unique tool for studies of human fat cell biology. Obes Facts 1, 184–189.

42. Zhou, B., Ho, S.S., Greer, S.U., Spies, N., Bell, J.M., Zhang, X., Zhu, X., Arthur, J.G., Byeon, S., Pattni, R., et al. (2019). Haplotype-resolved and integrated genome analysis of the cancer cell line HepG2. Nucleic Acids Res. 47, 3846–3861.

43. López-Terrada, D., Cheung, S.W., Finegold, M.J., and Knowles, B.B. (2009). Hep G2 is a hepatoblastoma-derived cell line. Hum. Pathol. 40, 1512–1515.

44. Aden, D.P., Fogel, A., Plotkin, S., Damjanov, I., and Knowles, B.B. (1979). Controlled synthesis of HBsAg in a differentiated human liver carcinoma-derived cell line. Nature 282, 615–616.

45. Kasowski, M., Kyriazopoulou-Panagiotopoulou, S., Grubert, F., Zaugg, J.B., Kundaje, A., Liu, Y., Boyle, A.P., Zhang, Q.C., Zakharia, F., Spacek, D.V., et al. (2013). Extensive variation in chromatin states across humans. Science 342, 750–752.

46. Shungin, D., Winkler, T.W., Croteau-Chonka, D.C., Ferreira, T., Locke, A.E., Mägi, R., Strawbridge, R.J., Pers, T.H., Fischer, K., Justice, A.E., et al. (2015). New genetic loci link adipose and insulin biology to body fat distribution. Nature 518, 187–196.

47. Bryois, J., Garrett, M.E., Song, L., Safi, A., Giusti-Rodriguez, P., Johnson, G.D., Shieh, A.W., Buil, A., Fullard, J.F., Roussos, P., et al. (2018). Evaluation of chromatin accessibility in prefrontal cortex of individuals with schizophrenia. Nat Commun 9, 3121.

48. Degner, J.F., Pai, A.A., Pique-Regi, R., Veyrieras, J.-B., Gaffney, D.J., Pickrell, J.K., Leon, S.D., Michelini, K., Lewellen, N., Crawford, G.E., et al. (2012). DNase?I sensitivity QTLs are a major determinant of human expression variation. Nature 482, 390–394.

49. Alasoo, K., Rodrigues, J., Mukhopadhyay, S., Knights, A.J., Mann, A.L., Kundu, K., Hale, and C., Dougan, G., and Gaffney, D.J. (2018). Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. Nat. Genet.

50. Kumasaka, N., Knights, A.J., and Gaffney, D.J. (2016). Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. Nat. Genet. 48, 206–213.

51. Gate, R.E., Cheng, C.S., Aiden, A.P., Siba, A., Tabaka, M., Lituiev, D., Machol, I., Gordon, M.G., Subramaniam, M., Shamim, M., et al. (2018). Genetic determinants of co-accessible chromatin regions in activated T cells across humans. Nature Genetics 50, 1140–1150.

52. Khetan, S., Kursawe, R., Youn, A., Lawlor, N., Jillette, A., Marquez, E.J., Ucar, D., and Stitzel, M.L. (2018). Type 2 Diabetes-Associated Genetic Variants Regulate Chromatin Accessibility in Human Islets. Diabetes 67, 2466–2477.

53. Garske, K.M., Pan, D.Z., Miao, Z., Bhagat, Y.V., Comenho, C., Robles, C.R., Benhammou, J.N., Alvarez, M., Ko, A., Ye, C.J., et al. (2019). Reverse gene-environment interaction approach to identify variants influencing body-mass index in humans. Nat Metab 1, 630–642.

54. Ramirez, R.N., El-Ali, N.C., Mager, M.A., Wyman, D., Conesa, A., and Mortazavi, A. (2017). Dynamic Gene Regulatory Networks of Human Myeloid Differentiation. Cell Syst 4, 416-429.e3.

55. Bertero, A., Fields, P.A., Ramani, V., Bonora, G., Yardimci, G.G., Reinecke, H., Pabon, L., Noble, W.S., Shendure, J., and Murry, C.E. (2019). Dynamics of genome reorganization during human cardiogenesis reveal an RBM20-dependent splicing factory. Nat Commun 10, 1538.

56. Ackermann, A.M., Wang, Z., Schug, J., Naji, A., and Kaestner, K.H. (2016). Integration of ATAC-seq and RNA-seq identifies human alpha cell and beta cell signature genes. Mol Metab 5, 233–244.

57. Fullard, J.F., Hauberg, M.E., Bendl, J., Egervari, G., Cirnaru, M.-D., Reach, S.M., Motl, J., Ehrlich, M.E., Hurd, Y.L., and Roussos, P. (2018). An atlas of chromatin accessibility in the adult human brain. Genome Res. 28, 1243–1252.

58. Calderon, D., Nguyen, M.L.T., Mezger, A., Kathiria, A., Müller, F., Nguyen, V., Lescano, N., Wu, B., Trombetta, J., Ribado, J.V., et al. (2019). Landscape of stimulation-responsive chromatin across diverse human immune cells. Nat. Genet. 51, 1494–1505.

59. Lo, K.A., Labadorf, A., Kennedy, N.J., Han, M.S., Yap, Y.S., Matthews, B., Xin, X., Sun, L., Davis, R.J., Lodish, H.F., et al. (2013). Analysis of in vitro insulin-resistance models and their physiological relevance to in vivo diet-induced adipose insulin resistance. Cell Rep 5, 259–270.

60. Wang, J., Wang, Y., Duan, Z., and Hu, W. (2020). Hypoxia-induced alterations of transcriptome and chromatin accessibility in HL-1 cells. IUBMB Life 72, 1737–1746.

CHAPTER 2: OPEN CHROMATIN PROFILING IN ADIPOSE TISSUE MARKS GENOMIC REGIONS WITH FUNCTIONAL ROLES IN CARDIOMETABOLIC TRAITS[1]

Introduction

Dysregulation of genes expressed in adipose tissue influences cardiometabolic traits and diseases. Subcutaneous adipose tissue serves as a buffering system for lipid energy balance, particularly fatty acids,[1-3] and may play a protective role in cardiometabolic risk.[4] Subcutaneous adipose expression quantitative trait loci (eQTL) studies have identified genes involved in central obesity and metabolic traits,[5-9] and specific cardiometabolic genome-wide association study (GWAS) loci have been shown to colocalize with subcutaneous adipose eQTLs.[9-13] In addition, a recent GWAS study of waist-hip ratio, a measure of central obesity, identified loci that were enriched both for putative regulatory elements in adipose nuclei and for genes expressed in subcutaneous adipose tissue,[12] many of which have been linked to adipose function.[14] Identification and characterization of adipose tissue regulatory regions and variants would improve understanding of biological processes and the mechanisms underlying cardiometabolic loci.

Adipose tissue is composed of many cell types, including adipocytes, preadipocytes, vascular cells, immune cells, and nerve cells.[15] Characterization of heterogeneous whole adipose tissue and its component cell types are both needed to fully delineate the role of adipose tissue in cardiometabolic disease. Human adipose tissue samples can be used to identify differences in chromatin accessibility due to genotype and link variants to cardiometabolic traits; however, samples may also differ due to site of tissue extraction, sample handling and storage conditions, and environmental contributions. Although cell models do not fully replicate cells within a complex tissue, their growth, storage, and environmental conditions can be controlled. Cells from the Simpson Golabi-Behmel Syndrome (SGBS) human preadipocyte cell strain are diploid, easy to grow in culture, can be

[1] The work in this chapter has been previously published. The citation is:
Cannon, M.E., Currin, K.W., Young, K.L., Perrin, H.J., Vadlamudi, S., Safi, A., Song, L., Wu, Y., Wabitsch, M., Laakso, M., et al. (2019). Open chromatin profiling in adipose tissue marks genomic regions with functional roles in cardiometabolic traits. G3: Genes, Genomes, Genetics 9, 2521–2533.

differentiated to mature adipocytes[16] and are exposed to less experimental variation than primary human preadipocytes due to genotype or sample collection differences.

Adipose tissue and adipocytes are poorly represented in chromatin accessibility datasets because the high lipid content makes experimental assays challenging. To date, for human adipose tissue or adipocytes, only three DNase-seq datasets[17,18] and three ATAC-seq datasets[19,20] are available. In addition to chromatin accessibility, chromatin immunoprecipitation (ChIP)-seq for histone marks have been characterized in adipose nuclei from subcutaneous adipose tissue and in differentiated adipocytes from mesenchymal stem cells (Roadmap Epigenomics Project), and these data were integrated to annotate genomic regions into chromatin states characteristic of regulatory functions such as promoters, enhancers, or insulators.[21] Regions of chromatin accessibility in many cell types are located preferentially in regulatory regions,[21,22] suggesting that chromatin accessibility maps can improve accuracy of predicting regulatory chromatin states in adipose cell types.

Chromatin accessibility data can be used to characterize candidate variants at noncoding GWAS loci. Allelic differences have been found in levels of accessible chromatin, transcription factor binding, and histone marks of chromatin state,[23-28] and these differences have provided a functional context for interpreting GWAS loci.[29-31] Identifying transcription factor motifs and footprints in accessible chromatin regions can be used to predict transcription factor binding sites.[32] Improved annotation of candidate regulatory variants and candidate transcription factors in adipose tissue could aid identification of molecular mechanisms at GWAS loci.

In this study, we performed ATAC-seq on frozen clinical subcutaneous adipose tissue needle biopsy samples and SGBS preadipocytes and adipocytes to identify regions of accessible chromatin for each sample type. We identified cardiometabolic GWAS loci and transcription factor binding motifs in ATAC-seq open chromatin regions and used the ATAC-seq annotations to characterize candidate variants at cardiometabolic GWAS loci with colocalized adipose tissue eQTL associations. Finally, through experimental analysis of allelic differences in regulatory functions, we report functional non-coding variants at two cardiometabolic GWAS loci.

Materials and Methods

METSIM study participants

Subcutaneous adipose tissue needle biopsies were obtained from METabolic Syndrome in Men (METSIM) participants as previously described.[9] We used three adipose tissue needle biopsy samples for ATAC-seq. The METSIM study includes 10,197 men, aged from 45 to 73 years, randomly selected from Kuopio, Eastern Finland,

and examined in 2005 – 2010.[33,34] 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. DNA samples were genotyped on the Illumina OmniExpress and HumanCoreExome arrays and imputed using the Haplotype Reference Consortium[35] as previously described.[9]

Sample processing and ATAC-seq library preparation

Human adipose tissue was flash frozen and stored at -80º until use. For adipose tissue samples 1 and 3, we generated libraries using nuclei isolation buffers that contained detergent (1% NP-40) or did not contain detergent. For tissue sample 2, we generated libraries using ~12 mg or ~36 mg of tissue and contained detergent. Replicates including detergent and less tissue in library preparation resulted in a greater number of peaks and higher peak similarity between individuals compared to no detergent. From these observations, we performed all subsequent analyses with the three detergent-treated replicates. Tissue was pulverized in liquid nitrogen using a Cell Crusher homogenizer (cellcrusher.com). The tissue powder was resuspended in nuclei isolation buffer (20 mM Tris-HCl, 50 mM EDTA, 60 mM KCl, 40% glycerol, 5 mM spermidine, 0.15 mM spermine, 0.1% mercaptoethanol, 1% NP-40). Tubes were rotated at 4º for 5 minutes. The solution was homogenized using a tight homogenizer (Wheaton) for 10 strokes and was centrifuged at 1500 x g for 10 minutes at 4º. Following removal of the lipid layer and supernatant, the pellet was resuspended in buffer (10 mM Tris-HCl, 10 mM NaCl, 3 mM $MgCl_2$) and centrifuged at 1200 x g for 10 minutes at 4º. The supernatant was removed and the pellet was used for the transposase reaction as previously described.[36] We used 2.5 ul Tn5 for adipose tissue libraries. Following library PCR amplification for adipose tissue, we removed primer dimers using Ampure Beads (Agencourt) with a 1:1.2 ratio of library to beads. Libraries were visualized and quantified using a TapeStation or Bioanalyzer and sequenced with 50-bp reads on an Illumina Hi-Seq 2500 at the Duke University Genome Sequencing shared resource facility (single-end sequencing).

SGBS cells[37] were generously provided by Dr. Martin Wabitsch (University of Ulm) and cultured as previously described.[38] To differentiate SGBS cells, SGBS preadipocytes were cultured in serum-containing medium until confluent, then rinsed in PBS and differentiated for four days in basal medium (DMEM:F12 + 3.3mM biotin + 1.7mM panthotenate) 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, differentiated SGBS cells were maintained in basal medium supplemented with 0.01 mg/mL transferrin, 20 nM insulin, 200 nM cortisol, 0.4 nM triiodothyronine. We generated profiles with 50,000 cells following the Omni-ATAC protocol[39].

We removed primer dimers using Zymo DNA Clean and Concentrator, visualized and quantified libraries using a TapeStation or Bioanalyzer, and sequenced with 50-bp reads on an Illumina Hi-Seq 4000 at the University of North Carolina High-Throughput Sequencing Facility (paired-end sequencing).

ATAC-seq alignment and peak calling

We obtained previously published adipose ATAC-seq datasets from subcutaneous adipose tissue (ENCODE ENCSR540BML),[20] tissue-derived adipocytes,[19] and GM12878 lymphoblasts.[36] The tissue-derived adipocyte ATAC-seq data was shared by the McGill Epigenomics Mapping Centre and is available from the European Genome-phenome Archive of the European Bioinformatics Institute (dataset EGAD00001001300).

To minimize mapping differences between read length and single-end vs. paired-end samples, we merged the mate pair fastq files and trimmed reads to 50 nucleotides for each paired-end ATAC-seq sample and aligned reads from all samples as single-end. We removed sequencing adapters from raw ATAC-seq sequence reads using Tagdust[40] with a false discovery rate of 0.1% and selected high quality reads with a Phred score of at least 20 for at least 90% of bases using the FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit). We aligned filtered reads to the hg19 human genome using bowtie2[41], penalizing ambiguous bases as mismatches. We removed any alignments with mapping quality less than 20, mitochondrial reads, or blacklisted regions[42,43] and shifted the resulting alignments by +4 on the + strand and -5 on the – strand so that the 5' base of each alignment corresponded to the center of the binding site of the Tn5 transposase[36,44]. For the METSIM adipose tissue samples, we verified sample identity using verifyBamID[45] using genotyped variants with at least 10 ATAC-seq reads in the sample with the lowest read depth (Tissue 2; 8,683 variants), minimum minor allele frequency of 0.01, and call rate of at least 0.5; we used the best-matched genotypes for each sample. For all samples, we called peaks using MACS2[46] with no background dataset, smoothing ATAC-seq signal over a 200 bp window centered on the Tn5 integration site, allowing no duplicates, and a false discovery rate (FDR)<5%; we refer to peaks called on reads from technical replicate samples (SGBS adipocytes, SGBS preadipocytes, tissue-derived adipocytes, and GM12878 lymphoblasts) as 'replicate peaks'.

Representative ATAC-seq peaks

For samples with technical replicates, we pooled reads across replicates and called peaks (MACS2, FDR<5%), and then defined the portion of these peaks that shared at least one base with a replicate peak in two or more replicates as 'representative peaks'. The METSIM adipose tissue samples are from different individuals and are not technical replicates. Due to a low number of samples, we used the union of peaks across individuals as representative peaks. Unless otherwise noted, we selected the top 50,000 representative peaks in each group for downstream analyses. For the groups with technical replicates and the single ENCODE adipose tissue sample, we selected the top 50,000 representative peaks with the most significant peak p-values. For METSIM adipose tissue, we ranked the peak p-values in each individual (with 1 being the strongest) and used the average of these ranks to select the top 50,000 representative peaks. This approach reduced the chance that outlier p-values from a single individual would bias peak rank.

ATAC-seq principal component analysis

We generated a total set of accessible chromatin regions by taking the top 50,000 peaks in each group of ATAC-seq samples. For each ATAC-seq sample, we counted the number of non-duplicated nuclear reads overlapping the total set of accessible chromatin regions using featureCounts.[47] We performed library size normalization and variance stabilization using the regularized log (rlog) function in DESeq2.[48] We performed principal component analysis (PCA) using a modified version of the DESeq2 plotPCA function.

Peak genomic distribution and overlap with Roadmap chromatin states

We determined the location of ATAC-seq peaks relative to genes from the GENCODE 24lift37 Basic Set. Using BEDTools,[42] we divided peaks into the following categories: TSS-proximal (5 kb upstream to 1 kb downstream of a GENCODE transcription start site), intragenic (within a gene body but not within TSS-proximal regions), downstream (within 5 kb downstream of a transcription termination site but not within any gene body), and distal (>5 kb from either end of any gene). We obtained chromatin states for an 18-state model based on ChIP-seq data for 98 cell and tissue types using 6 histone marks (H3K4me1, H3K4me3, H3K36me3, H3K27me3, H3K9me3, and H3K27ac) from the Roadmap Epigenomics Consortium.[21] We generated the following combined states by merging states of similar genomic context: promoter (1_TssA, 2_TssFlnk, 3_TssFlnkU, 4_TssFlnkD, 14_TssBiv), transcribed (5_Tx, 6_TxWk), enhancer (7_EnhG1, 8_EnhG2, 9_EnhA1, 10_EnhA2, 11_EnhWk, 15_EnhBiv), and polycomb repressed (16_ReprPC, 17_ReprPCWk). Using BEDTools[42] we calculated the number of representative ATAC-seq peak bases that overlapped each chromatin state. We ranked the ATAC-seq peak overlap of each chromatin state in adipose nuclei (Roadmap epigenome ID E063) relative to all other cell types, where a rank of 1 corresponds to largest amount of overlap compared to all other cell types.

Enrichment of transcription factor motifs within ATAC-seq peaks

We tested for enrichment of 519 transcription factor binding motifs from the JASPAR core 2016 vertebrates database[49] within the top 50,000 representative peaks for adipose tissue and GM12878 lymphoblasts using Analysis of Motif Enrichment (AME)[50]. We used shuffled peak sequences with preserved dinucleotide content as background for the enrichment and the Fisher Exact Test to calculate enrichment significance. We classified motifs with an Expect value (E) less than $1x10^{-100}$ as significantly enriched.

Transcription factor motif scanning and footprinting within ATAC-seq peaks

To identify transcription factor motifs both disrupted and generated by GWAS variants, we constructed personalized reference genomes (hg19) with the –create_reference option in the AA-ALIGNER pipeline[51] using genotypes in the adipose tissue samples. We scanned the resulting haplotypes for 519 transcription factor binding motifs from the JASPAR core 2016 vertebrates database using FIMO.[49,52] If two motifs for the same factor existed at the exact same genomic coordinates and on the same strand on each haplotype, we used the motif with the highest motif score.

We performed transcription factor footprinting for 35 transcription factor motifs corresponding to 34 unique adipose-related transcription factors. The 34 transcription factors included 21 described as adipose core transcription factors[53], six dimer motifs that contained a core transcription factor, plus CEBPA, CEBPB, CEBPD, ZEB1, SPI1, SPIB, and CTCF. For the resulting motifs, we generated windows containing the genomic coordinates of the motif and 100 bp flanking both motif edges. We removed motif windows where fewer than 90% of bases could be uniquely mapped or that overlapped blacklisted regions.[42,43,54] We constructed matrices of the number of Tn5 transpositions across the remaining motif windows and predicted which motifs were likely bound using CENTIPEDE.[55] We used motif scores calculated by FIMO for CENTIPEDE priors and classified a motif with a CENTIPEDE posterior binding probability greater than 0.99 as bound and less than 0.5 as unbound.

Next, we determined which transcription factors exhibited an average decrease in ATAC-seq signal across their motifs relative to flanking regions, termed an aggregate footprint profile; we considered these footprints to be the most robust and consistent footprints across all motif sites. We calculated the average transposition probability at each window position separately for bound and the top 10,000 unbound sites to obtain aggregate bound and unbound profiles, calculated the transposition probability ratio (TPR) by dividing each position in the bound profiles by the corresponding position in the unbound profiles, and then calculated the average TPR across the motifs (mTPR) and the 100 bp flanking regions (fTPR). We considered transcription factor motifs to display an aggregate footprint profile if mTPR was less than fTPR.

Enrichment of GWAS variants in ATAC-seq peaks

We tested for enrichment of genetic variants in ATAC-seq peaks using GREGOR, which compares overlap of GWAS variants relative to control variants matched for number of LD proxies, allele frequency, and gene proximity.[56] We selected lead variants with a p-value less than $5x10^{-8}$ from 11 trait categories from the GWAS catalog (December 2016): type 2 diabetes, insulin, glucose, cardiovascular outcomes, blood pressure traits, low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), triglycerides, total cholesterol, body mass index (BMI), and waist-hip ratio adjusted for BMI (WHR). Loci that were associated with multiple traits were assigned to each trait. To remove multiple lead variants for the same association signal, we performed LD clumping using swiss (https://github.com/welchr/swiss) with the 1000G_2014-11_EUR LD reference; variants in moderate LD (r^2>0.2) and within 1 Mb of a variant with a more significant p-value were removed. We used GREGOR to test for enrichment of the resulting GWAS lead variants or their LD proxies (r^2 threshold of 0.8 within 1 Mb of the GWAS lead, 1000 Genomes Phase I) in ATAC-seq peaks relative to control variants. We tested for enrichment in the top 50,000 representative peaks for adipose tissue, SGBS adipocytes, SGBS preadipocytes, and GM12878 lymphoblasts. Enrichment was considered significant if the enrichment p-value was less than the Bonferroni-corrected threshold of $5x10^{-3}$ (0.05/11 trait groups). To compare enrichment magnitudes between regions and traits, we calculated an enrichment z-score:

$$\text{z-score}=\frac{\text{observed overlaps-expected overlaps}}{\text{standard deviation}}$$

The expected overlaps and standard deviation were estimated using GREGOR.[56] We visualized the enrichment results using the heatmap.2 function in the gplots R package.[57,58]

Overlap of GWAS-eQTL colocalized loci with ATAC-seq peaks

eQTL mapping in 770 subcutaneous adipose tissue samples and determination of GWAS-coincident eQTLs was described previously.[9,38] We identified overlap of ATAC-seq peaks with any variant in LD (r^2>0.8) with the GWAS lead variant at 110 loci (6,692 variants) using BEDTools.[42] LD was calculated using the 770 METSIM individuals included in the eQTL analysis.

Transcriptional reporter luciferase assays

SGBS preadipocyte, 3T3-L1 preadipocyte, SW872 liposarcoma, and THP-1 monocyte cells were maintained and transcriptional reporter luciferase assays were performed as previously described.[38,59] 3T3-L1 preadipocytes (ATCC, CL-173) were differentiated as described in the ATCC protocol. Amplified regions were inserted in pGL4.23 firefly luciferase reporter vectors (Promega) upstream of the minimal promoter and luciferase gene. We cloned two sizes of constructs for rs7187776 due to a restriction enzyme site in the middle of the larger construct; we tested both in luciferase assays. The long construct includes part of the 3' UTR of *TUFM* and part of the 5' UTR of *SH2B1*. Fragments containing potential enhancers are designated as 'forward' or 'reverse' based on their orientation with respect to the genome. Regions were designed to include the entire ATAC-seq peak overlapping the variant of interest. Three to five independent clones were cotransfected with *Renilla* luciferase vector in triplicate (SGBS, 3T3-L1 adipocytes) or duplicate (SW872, THP1, 3T3-L1 preadipocytes) wells using Lipofectamine 3000 (SGBS, THP-1, Life Technologies), Lipofectamine 2000 (3T3-L1 preadipocytes and adipocytes) or FUGENE 6 (SW872, Promega). Firefly luciferase activity of the clones containing the PCR fragments was normalized to *Renilla* luciferase readings to control for differences in transfection efficiency. We repeated all luciferase transcriptional reporter experiments on independent days and obtained consistent results. Data are reported as fold change in activity relative to an empty pGL4.23 vector. We used two-sided Student's t-tests to compare luciferase activity.

Electrophoretic mobility shift assays (EMSA)

For EMSA, we prepared nuclear cell extracts from SGBS preadipocyte and SW872 cells using the NE-PER nuclear and cytoplasmic extraction kit (Thermo Scientific) as previously described.[60] Double-stranded oligos were incubated with SGBS preadipocyte or SW872 nuclear extract or 100 ng purified PU.1 protein (Creative Biomart SPI1-172H) and DNA-protein complex visualization was carried out as previously described.[60] A positive control oligo contained the PU.1 motif from JASPAR and a negative control did not contain the motif. We repeated all EMSA experiments on independent days and obtained consistent results.

Allelic imbalance

We aligned reads for the adipose tissue samples to personalized genomes using the allele-aware aligner GSNAP allowing two mismatches, no indels, and treating ambiguous bases (encoded as N's) as mismatches.[61] We extracted unique alignments and filtered alignments to the mitochondrial genome and blacklisted regions.[43,54] Using

WASP,[62] we removed alignments that did not uniquely map to each allele at heterozygous sites. Allele count pileup files were generated at heterozygous sites with a minimum base quality Phred score of 30 to minimize the impact of sequencing errors using samtools. We removed heterozygous loci with aligned bases other than the two genotyped alleles and selected heterozygous sites with at least 10 total counts and at least 1 count per allele. To account for residual biases, we fit allele counts to a beta-binomial distribution with the probability of success (reference allele ratio) and dispersion estimated using maximum likelihood separately for each sample using the VGAM R package.[57,63] We performed two-tailed beta-binomial tests of allelic imbalance using VGAM.

To confirm allelic imbalance in PU.1 binding and chromatin accessibility at rs7187776 (genomic position chr16:28857645), we analyzed public genotype, SPI1 ChIP-seq, and DNase-seq data for the GM12891 cell line. We obtained genotypes for individual NA12891 from ftp://ftp-trace.ncbi.nih.gov/1000genomes/. We downloaded GM12891 SPI1 ChIP-seq alignments (ENCFF450BQJ, ENCFF152ZGE) and DNase-seq alignments (ENCFF070BAN) from ENCODE. Allele count pileup files were generated at heterozygous sites with a minimum base quality Phred score of 30 to minimize the impact of sequencing errors using samtools.

Results

Chromatin accessibility in frozen adipose tissue and SGBS preadipocytes and adipocytes

We generated ATAC-seq open chromatin profiles from three frozen subcutaneous adipose tissue needle biopsy samples, two replicates of SGBS preadipocytes, and three replicates of SGBS adipocytes. In the adipose tissue samples, we generated ~56-70 million non-duplicated nuclear reads and ~36-58 thousand peaks (FDR<5%, **Table 2.1**, Methods). We identified 68,571 representative adipose tissue peaks by taking the union of peaks across the three samples. We generated a comparable number of non-duplicated nuclear reads in the SGBS samples (~30-90 million), but identified many more peaks (122,924 and 164,252 representative peaks for SGBS preadipocytes and adipocytes respectively) (**Table 2.1**, Methods). The lower signal-to-noise of adipose tissue profiles compared to cultured, largely homogeneous SGBS cells is expected due to the heterogeneity of whole adipose tissue and stress resulting from sample freezing.

Using principal component analysis of ATAC-seq read counts within representative peaks, we identified that adipose tissue, SGBS preadipocyte, and SGBS adipocyte samples cluster into three distinct groups with strong within-group similarity (**Figure 2.1A**). The adipose tissue profiles were more similar to SGBS adipocyte profiles than to SGBS preadipocyte profiles (**Figure 2.1A**), suggesting the adipose tissue samples contain more adipocytes than preadipocytes.

We tested for enrichment of 519 transcription factor binding motifs from the JASPAR database in the top 50,000 representative adipose tissue ATAC-seq peaks using AME.[49,50] We identified 162 significantly enriched motifs ($E < 1x10^{-100}$), including 41 motifs enriched in adipose tissue but not lymphoblasts. The set of 41 contains motifs for transcription factors known to promote adipogenesis, such as CEBP family members, STAT family members, and PPARG.[64]

To evaluate the distribution of ATAC-seq peaks across samples, we examined the accessible chromatin landscape at *ADIPOQ*, which encodes adiponectin, a hormone secreted by adipocytes that is not expressed in preadipocytes.[65,66] Adipose tissue and SGBS adipocyte ATAC-seq peaks overlapped the transcription start site (TSS) and parts of previously described regulatory elements upstream and in intron 1 of *ADIPOQ* that showed increased transcriptional activity in reporter assays[67,68] (**Figure 2.1B**). Additionally, a strong ATAC-seq peak downstream of *ADIPOQ* was present in SGBS preadipocytes, suggesting this region may harbor preadipocyte-specific regulatory elements. These data demonstrate that reproducible ATAC-seq open chromatin profiles can be obtained from small amounts (12-36 mg, one-third to two-thirds of a needle biopsy) of frozen clinical subcutaneous adipose tissue samples and SGBS preadipocytes and adipocytes.

Comparison of adipose tissue, adipocyte, and preadipocyte open chromatin

We compared our adipose tissue and SGBS representative ATAC-seq peaks to existing ATAC-seq datasets from tissue-derived adipocytes,[19] ENCODE subcutaneous adipose tissue, and GM12878 lymphoblasts (outgroup) using three methods. First, principal component analysis of read counts within representative peaks shows that our adipose tissue profiles were most similar to ENCODE adipose tissue and tissue-derived adipocyte profiles (**Figure 2.1A**). These tissue-derived adipocyte and ENCODE adipose tissue profiles were also more similar to SGBS adipocytes than SGBS preadipocytes. Our adipose tissue and SGBS profiles were more similar to existing adipocyte profiles than to GM12878 profiles.

Second, we compared the distribution of ATAC-seq peaks to Roadmap Epigenomics Consortium chromatin states in adipose nuclei isolated from subcutaneous adipose tissue.[21] We used the top 50,000 representative peaks in each group of samples. For all ATAC-seq profiles, the majority of peaks were located in adipose nuclei promoter and enhancer states, with fewer peaks located in regions associated with closed chromatin (heterochromatin, polycomb states). Our adipose tissue peaks showed the strongest overlap (40% enhancer, 49% promoter, 89% combined) with adipose nuclei promoters and enhancers compared to all other ATAC-seq profiles (**Figure 2.1C**). With the exception of ENCODE adipose tissue, enhancer coverage was consistently higher for adipose tissue and adipocyte profiles compared to preadipocyte and GM12878 lymphoblast profiles, whereas promoter coverage was similar between all samples (**Figure 2.1C**). The ENCODE adipose tissue profile had more peak bases in regions near transcription start sites and fewer peak bases in distal regions compared to all other profiles which may reflect technical differences in sample processing.

Third, to characterize the epigenome distribution of ATAC-seq peaks across cell types, we determined the overlap of representative peaks from each ATAC-seq group with enhancer chromatin states from 98 Roadmap tissues and cell types including adipose nuclei.[21] Adipose tissue and tissue-derived adipocyte peaks showed the most overlap with adipose nuclei enhancers, and SGBS adipocytes showed the 4th most overlap with adipose nuclei enhancers compared to enhancers in other tissue and cell types. SGBS preadipocytes showed the most overlap with enhancers in fibroblast cell types, and adipose nuclei ranked 24th among all cell types. As expected, GM12878 lymphoblast peaks showed much less overlap with adipose nuclei enhancers, consistent with the cell type-specific nature of enhancers.[21] Across the three methods, our adipose tissue and SGBS ATAC-seq profiles showed strong similarity with existing adipocyte ATAC-seq profiles and with active regulatory element chromatin states in adipose nuclei.

Cardiometabolic GWAS loci in ATAC-seq peaks

To identify cardiometabolic traits that may be strongly affected by adipocyte regulatory elements, we tested for enrichment of GWAS variants for 11 cardiometabolic trait groups in the top 50,000 representative ATAC-seq peaks in adipose tissue, SGBS adipocytes, SGBS preadipocytes, and GM12878 lymphoblasts. Variants at loci for four trait groups (WHR, HDL-C, cardiovascular outcomes, and blood pressure traits) showed significant enrichment (P<5x10^{-3}) in adipose tissue, SGBS adipocyte, and SGBS preadipocyte peaks (**Figure 2.2**). WHR was the most strongly enriched trait in adipose tissue (z-score=8.66) and SGBS adipocyte (z-score=4.53) peaks, whereas blood pressure traits were most strongly enriched in SGBS preadipocyte peaks (z-score=4.29). Loci for insulin traits and WHR showed stronger enrichment in adipose tissue peaks compared to SGBS adipocyte or preadipocyte peaks, suggesting *in vivo* conditions and/or non-adipocyte cell types in adipose tissue may contribute to these traits. Loci for HDL-C, triglycerides, LDL-C, and total cholesterol were significantly enriched in SGBS adipocytes, consistent with the roles of adipocytes in lipid storage. In contrast, loci for none of the tested traits were enriched in GM12878 lymphoblast peaks. Our results suggest that genetic variation in adipose tissue and adipocyte accessible chromatin regions is frequently associated with several cardiometabolic traits and that the stronger enrichment of WHR and insulin trait loci in adipose tissue relative to adipocyte or preadipocyte peaks demonstrates the importance of profiling chromatin accessibility in tissue.

Functional evaluation of cardiometabolic GWAS variants overlapping ATAC-seq peaks

We next identified cardiometabolic GWAS variants that overlapped candidate regulatory elements defined by ATAC-seq peaks. We focused on ATAC-seq peaks at a subset of 110 cardiometabolic GWAS loci that were colocalized with gene expression quantitative trait loci (eQTLs) in subcutaneous adipose tissue;[9,38] these loci consisted of 6,692 variants (LD r^2>0.8 with lead GWAS variants). To strengthen annotation at these loci, we overlapped variants at these loci with all representative ATAC-seq peaks rather than the top 50,000 peaks. 147 variants at 59 loci overlapped an adipose tissue peak. The loci that had only one variant overlapping an adipose tissue ATAC-seq peak are shown in **Table 2.2**; these variants are strong candidates for functional activity at these loci. Of these 147 variants, 136 (93%) also overlapped an SGBS adipocyte peak and 116 (79%) overlapped both an SGBS adipocyte and preadipocyte peak. Variants that overlap peaks in adipose tissue and adipocytes or preadipocytes may be more likely to act through regulatory elements present in adipocytes rather than blood, immune, or other adipose tissue cell type regulatory elements. Of the 147 variants, 97 (66%) overlapped a transcription factor (TF) motif from JASPAR.[49] Using a stringent definition for transcription factor footprints (Methods), we identified aggregate footprint profiles for 12 of 35 tested TF motifs in adipose tissue and found that four variants overlapped a TF footprint. These candidate functional variants, target regulatory elements, and TFs provide a resource to investigate the mechanisms underlying cardiometabolic GWAS loci.

We tested variants at two loci for allelic differences in functional regulatory assays. The first, rs1534696, was identified as a candidate regulatory variant based on overlap with an ATAC-seq peak in adipose tissue and tissue-derived adipocytes, but was not a candidate based on SGBS adipocyte or preadipocyte ATAC-seq peaks or adipose promoter or enhancer Roadmap chromatin state (**Figure 2.3A**). rs1534696 is located in the second intron of *SNX10* (encoding sorting nexin 10), was associated with WHR (P=2x10^{-8}, ß=0.027, in women)[12] and exhibited a colocalized eQTL for *SNX10* (P=3.4x10^{-150}, ß=1.12) and *CBX3* (P=1.1x10^{-13}, ß=0.39) in adipose tissue.[9] We tested alleles of rs1534696 in a 250-bp region encompassing the ATAC-seq peak for transcriptional differences in luciferase reporter assays using four cell types (**Figure 2.3B**). In 3T3-L1 preadipocytes and adipocytes, the construct containing rs1534696-A showed higher transcriptional activity than rs1534696-C (*P*=0.01) in both orientations (**Figure 2.3B**). Similar trends were also observed in SW872 liposarcoma and SGBS preadipocyte cells; this direction of effect is consistent with the eQTL association of rs1534696-A with higher levels of *SNX10* and *CBX3*. In addition, rs1534696-A showed increased protein binding in EMSAs using nuclear extract from SGBS preadipocytes (**Figure 2.3C**). These data suggest that a transcriptional activator binds more strongly to rs1534696-A and increases transcriptional activity of *SNX10* and/or *CBX3*, contributing to the molecular mechanism at this GWAS locus (**Figure 2.3D**).

The second variant we tested overlapped an ATAC-seq peak in adipose tissue, SGBS preadipocytes, SGBS adipocytes, and tissue-derived adipocytes and a SPI1 (PU.1) ChIP-seq peak, motif and footprint (**Figure 2.4A**). In adipose tissue sample 1, we further observed an allelic imbalance in ATAC-seq reads (P=2.90x10^{-3}): 25 reads contained rs7187776-A and 3 reads contained rs7187776-G. rs7187776 is located near a long isoform of *SH2B1* (encoding SH2B adaptor protein 1) and is in strong LD (r2 > 0.8) with the lead variant associated with BMI (rs3888190, P=3.14x10^{-23}, ß=0.031).[13] This GWAS signal exhibited a colocalized eQTL for *SH2B1* (P=4.7x10^{-15}, ß=-0.39) and *ATXN2L* (P=2.5x10^{-11}, ß=-0.34) in adipose tissue.[9] rs7187776 is one of 124 candidate variants based on LD (r2>0.8) with the lead GWAS and eQTL variants, and one of five variants that overlapped ATAC-seq peaks at this locus. Using EMSA, we observed allele-specific binding of rs7187776-G to purified PU.1 protein and similar binding using nuclear extract from SW872 cells, consistent with the predicted motif (**Figure 2.4B**). We also tested alleles of rs7187776 in a 477-bp region encompassing the ATAC-seq peak and a smaller 186-bp region in transcriptional reporter assays (**Figure 2.4**). In THP-1 monocytes, the constructs containing rs7187776-A showed increased transcriptional activity compared to rs7187776-G (**Figure 2.4C**). In SGBS preadipocyte, SW872 liposarcoma, 3T3-L1 preadipocyte, and 3T3L-1 adipocyte cells, we observed extremely strong transcriptional activity (>200-fold compared to background) but no allelic differences; differences may have been masked by the massive >200-fold transcription-enhancing effect of this region. rs7187776-G is associated with decreased expression levels of *SH2B1* and *ATXN2L*, suggesting that PU.1 or another ETS family member may act as a transcriptional repressor at this locus. We observed fewer ATAC-seq reads corresponding to more PU.1 binding, a direction that has been observed less often than increased ATAC-seq reads corresponding to increased transcription factor binding.[23] We observed the same pattern in GM12891 SPI1 ChIP-seq and DNase-seq data from ENCODE; 2

ChIP-seq reads contained rs7187776-A and 11 reads contained rs7187776-G, whereas 11 DNase-seq reads contained rs7187776-A and 1 read contained rs7187776-G. Multiple ETS family members, including PU.1, can act as transcriptional repressors, including by recruiting histone deacetylases and DNA methyltransferases, resulting in closed chromatin,[69-72] consistent with rs7187776-G showing fewer ATAC-seq reads. These data suggest that rs7187776-G increases binding of an ETS family member, and may contribute to the molecular mechanism at the *ATP2A1-SH2B1* BMI GWAS locus (**Figure 2.4D**).

Allelic imbalance in ATAC-seq reads

We looked for other examples of allelic imbalance in ATAC-seq reads at heterozygous positions that may indicate altered chromatin accessibility. Only 387 sites showed nominal allelic imbalance (beta-binomial P<0.05) in at least one sample, 6 of which overlapped variants at GWAS-eQTL loci. However, only 40 of 6,692 total GWAS-eQTL variants were heterozygous in at least one adipose tissue sample and were covered by enough ATAC-seq reads for allelic imbalance analysis, suggesting that higher read depth and larger sample sizes that increase the chance of heterozygosity at more eQTL and GWAS loci may enable identification of more disease-associated loci that could mediate their effects on disease through chromatin accessibility.

Discussion

In this study, we generated ATAC-seq open chromatin profiles from three frozen clinical adipose samples and replicate preparations of SGBS preadipocytes and adipocytes. We identified differences between adipose tissue, preadipocyte, and mature adipocyte open chromatin profiles, including cell-type-specific peaks at selectively expressed promoters. Adipose tissue, SGBS adipocyte, and SGBS preadipocyte open chromatin profiles largely overlapped Roadmap adipose nuclei chromatin states. Transcription factor motifs and footprints in ATAC-seq peaks overlapped GWAS variants, and GWAS variants for several traits were enriched in ATAC-seq peaks. Finally, we used the ATAC-seq profiles to annotate potential regulatory variants at GWAS-eQTL colocalized loci and provided experimental evidence of allelic differences in regulatory activity for variants at the *SNX10* and *ATP2A1-SH2B1* GWAS loci. Taken together, these data are among the deepest characterization of chromatin accessibility in adipose tissue, adipocytes, and preadipocytes to date.

Important differences exist between adipose tissue, preadipocyte, and mature adipocyte ATAC-seq profiles. Explanations for these differences include cell-type composition/heterogeneity, the differentiation state of adipocytes, the cultured nature of SGBS cells, and technical differences of ATAC-seq data (e.g., sequencing depth).

At the TSS for *ADIPOQ*, we observed adipose tissue and SGBS adipocyte ATAC-seq peaks, and downstream of *ADIPOQ,* we observed ATAC-seq peaks specific to SGBS preadipocytes. The accessibility pattern of *ADIPOQ* is consistent with its role in adipocyte differentiation [73-75] and a previous finding that the *ADIPOQ* promoter is inaccessible until differentiation[76]. Among 98 Roadmap tissue and cell types, SGBS preadipocyte ATAC-seq profiles were more similar to fibroblast-like cells and cell lines than to adipose nuclei, and SGBS adipocytes were more similar to adipose nuclei, reflecting differences likely due to the fibroblast-like nature of preadipocytes. Differences between our adipose tissue ATAC-seq profiles and the ENCODE adipose tissue data may be due to differences in biopsy location, freezing method, storage conditions, or library preparation.

Adipose ATAC-seq profiles provide insight into the mechanisms of cardiometabolic GWAS loci. For example, we found that GWAS variants for WHR— but not BMI—are enriched in adipose ATAC-seq peaks. This enrichment is consistent with recent findings that WHR loci are enriched in adipose transcriptional regulatory elements[12] and that BMI GWAS loci are enriched in pathways involved in central nervous system biology.[13] We also identified enrichment of other cardiometabolic traits, including insulin traits, lipids, and cardiovascular outcomes, highlighting the relevance of adipose regulatory elements for these traits. Identifying the transcription factor(s) bound to a regulatory variant is a challenging part of defining the molecular mechanisms underlying cardiometabolic GWAS loci. While transcription factor footprints better predict that a transcription factor is bound at a locus compared to motif occurrence alone,[55] neither footprints nor motifs identify the bound transcription factor with 100% accuracy, particularly when multiple transcription factors share similar binding motifs. We successfully generated transcription factor footprints for 12 transcription factor motifs, which can be used to identify GWAS variants that may alter transcription factor binding. However, additional experiments are needed to confirm the identity of transcription factors bound at loci containing these footprints.

We described two GWAS loci for which ATAC-seq peaks helped prioritize candidate variants. At the *SNX10* WHR locus, we identified a potentially functional variant, rs1534696, which is not located in a predicted regulatory region based on existing chromatin state data. rs1534696 overlaps an ATAC-seq peak in adipose tissue and showed allelic differences in transcriptional reporter and protein-binding assays. Interestingly, we observed allelic differences in protein binding in SGBS preadipocytes, yet low transcriptional activity, similar to empty vector, in SGBS preadipocytes and 3T3L1 cells. One possibility is that a repressor binds in preadipocytes to prevent transcription and is then released to activate transcription in adipocytes; additional experiments are needed to determine the apparent differences between preadipocytes and adipocytes at this locus. At the *ATP2A1-SH2B1* BMI locus, we identified a PU.1 binding motif and footprint at rs7187776, as well as allelic imbalance in ATAC-seq reads, and confirmed the allelic differences in PU.1 binding *in vitro.* PU.1 is part of the ETS family of transcription factors, all of which have very similar DNA binding motifs,[77] so PU.1 may not be the specific TF binding at this locus, especially because PU.1 is expressed at very low levels in SGBS preadipocytes, SGBS adipocytes, and isolated mature adipocytes.[18,19] Interestingly, we observed significant allelic differences in transcriptional activity in THP-1 monocyte cells but not in preadipocyte or adipocyte cell types (**Figure 2.4**), suggesting that this variant might be important in non-adipocyte cells within adipose tissue. These data provide excellent examples of how to integrate GWAS, eQTL, and ATAC-seq data to identify functional variants at GWAS loci. Further experiments are needed to determine if these variants are the only functional variants at each locus, as we also observed allelic differences in protein binding for a second variant overlapping an ATAC-seq peak at the *SH2B1* locus and others have suggested different functional variants at this locus,[78,79] and which gene(s) are contributing to obesity risk.

In summary, we presented ATAC-seq open chromatin profiles for frozen adipose tissue and cultured preadipocytes and adipocytes. We showed the utility of open chromatin profiles in multiple tissue samples and across cell types within heterogeneous tissue. Together, these data add to the growing understanding of gene regulation in adipose and the complex genetic mechanisms of cardiometabolic traits and diseases.

Acknowledgements

I thank everyone who contributed to the work in this chapter: Maren Cannon, Hannah Perrin, and Alexias Safi for generating ATAC-seq libraries; Maren Cannon, Kristen Young, and Swarooparani Vadlamudi for performing functional assays; Lingyun Song and Ying Wu for helping with computational analyses; Martin Wabitsch for providing SGBS cells; Markku Laakso for his work on the METSIM study; and Gregory Crawford and Karen Mohlke for advising analyses.

We thank the participants in the METSIM study who donated adipose tissue samples. We acknowledge ENCODE and Michael Snyder’s lab (Stanford University) for ATAC-seq data and thank Dr. Elin Grundberg and the McGill Epigenomics Mapping Centre for access to the mature adipocyte ATAC-seq data. We acknowledge the NIH Roadmap Epigenomics Consortia for chromatin state data. We thank the personnel at the University of North Carolina High-Throughput Sequencing Facility and the Duke University Center for Genomic and Computational Biology for performing sequencing runs. We thank Mete Civelek (University of Virginia) for assistance with adipose eQTL data and Stephen Parker (University of Michigan), Terrence Furey, Jason Stein, and Bryan Quach for helpful discussion and bioinformatics expertise. This study was supported by NIH grants F31HL127984 (M.E.C.), R25GM055336 (K.W.C.), T32GM67553 (K.W.C.), T32HL069768 (H.J.P.), U01KD105561 (K.L.M.), R01DK093757 (K.L.M.), and KL2TR001109 (K.L.Y.); Academy of Finland grants 77299, 124243, 141226 (M.L.); Finnish Diabetes Foundation (M.L.); Finnish Heart Foundation (M.L.); and the Commission of the European Community grant HEALTH-F2- 2007-201681 (M.L.).

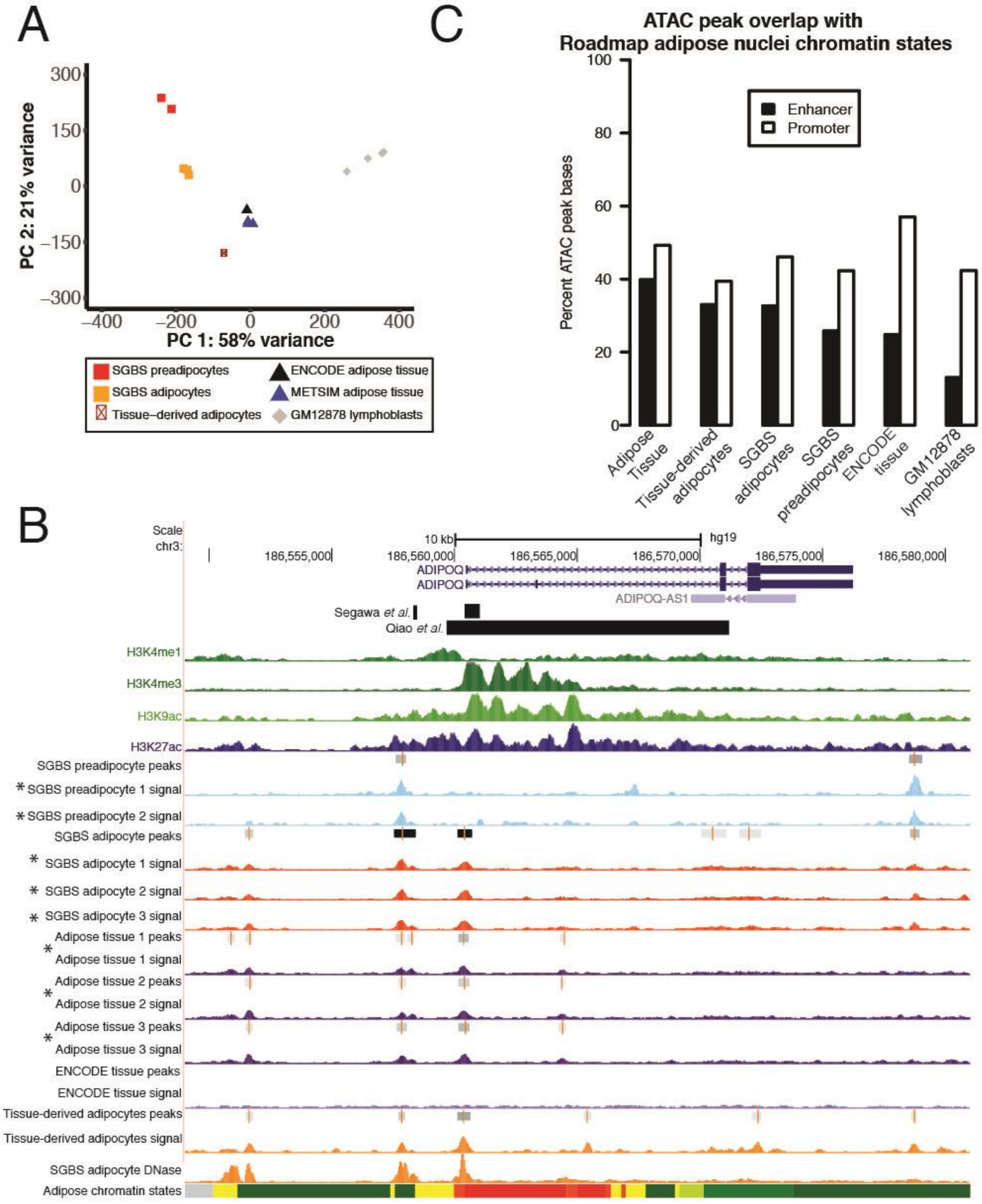


Figure 2.1. Comparison of ATAC-seq read profiles and peaks between samples and with Roadmap adipose nuclei chromatin states. (A) Principal components analysis (PCA) of ATAC-seq read counts within representative peaks. (B) UCSC genome browser image (hg19) showing the *ADIPOQ* gene regions. ChIP-seq for histone marks

from the Roadmap Epigenomics project adipose nuclei are shown at the top in green and blue. ATAC-seq signal tracks are shown in different colors by source: SGBS preadipocytes in light blue, SGBS adipocytes in red, adipose tissue in purple, ENCODE adipose tissue in light purple, and tissue-derived adipocytes in orange. DNase hypersensitivity signal tracks for SGBS adipocytes are also shown in orange. Asterisks represent ATAC-seq data generated in this manuscript. Peak regions are indicated by gray bars. The bottom track shows chromatin states from the Roadmap Epigenomics Project for adipose nuclei (yellow = enhancer; green = transcribed; orange/red = promoter; light green = genic enhancer; gray = repressed/polycomb; light red = bivalent/poised TSS; turquoise = heterochromatin). (C) Overlap of the top 50,000 ATAC-seq peaks with promoter and enhancer chromatin states identified in Roadmap adipose nuclei.

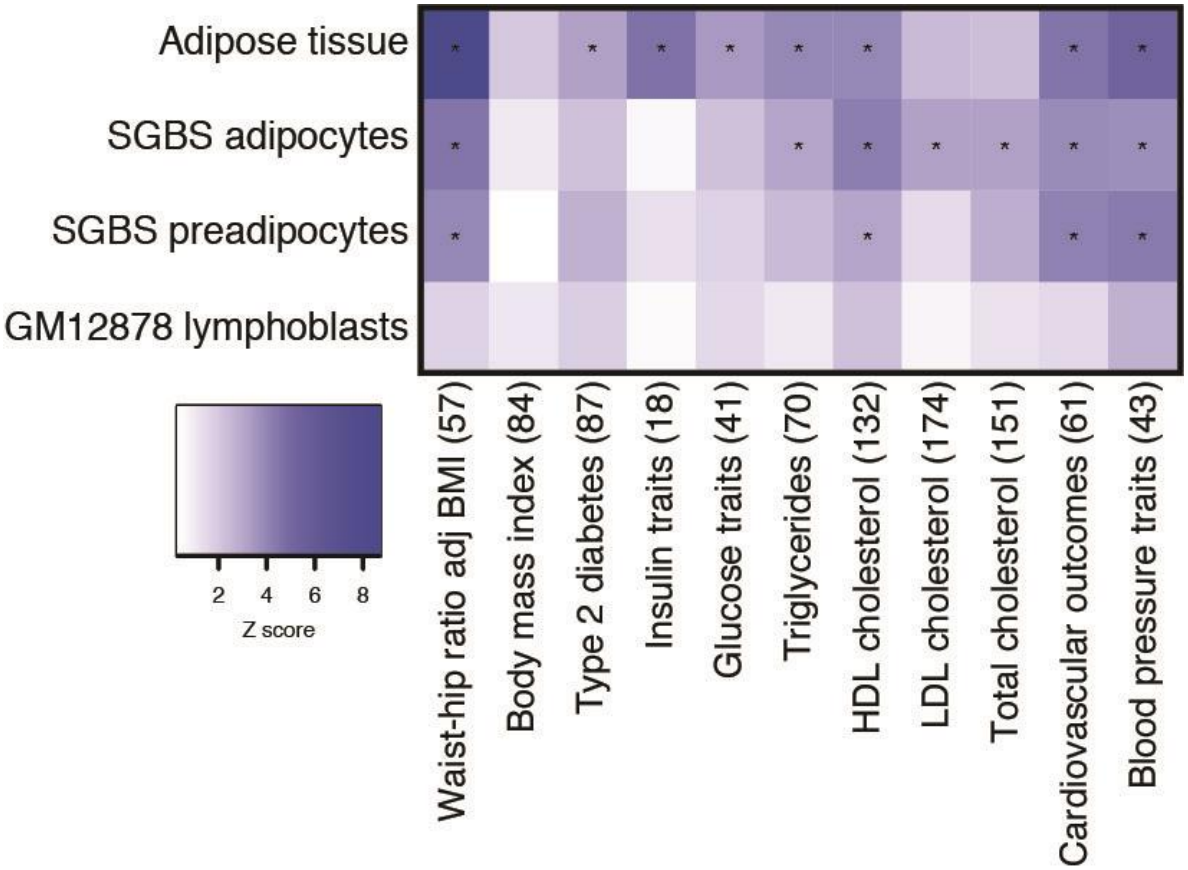


Figure 2.2. Cardiometabolic GWAS loci are enriched in ATAC-seq peaks. The heatmap shows enrichment of cardiometabolic GWAS loci (z-score) for the top 50,000 representative ATAC-seq peaks in adipose tissue, SGBS adipocytes, SGBS preadipocytes, and GM12878 lymphoblasts. Cells with a significant p-value (p<0.005) contain an asterisk.

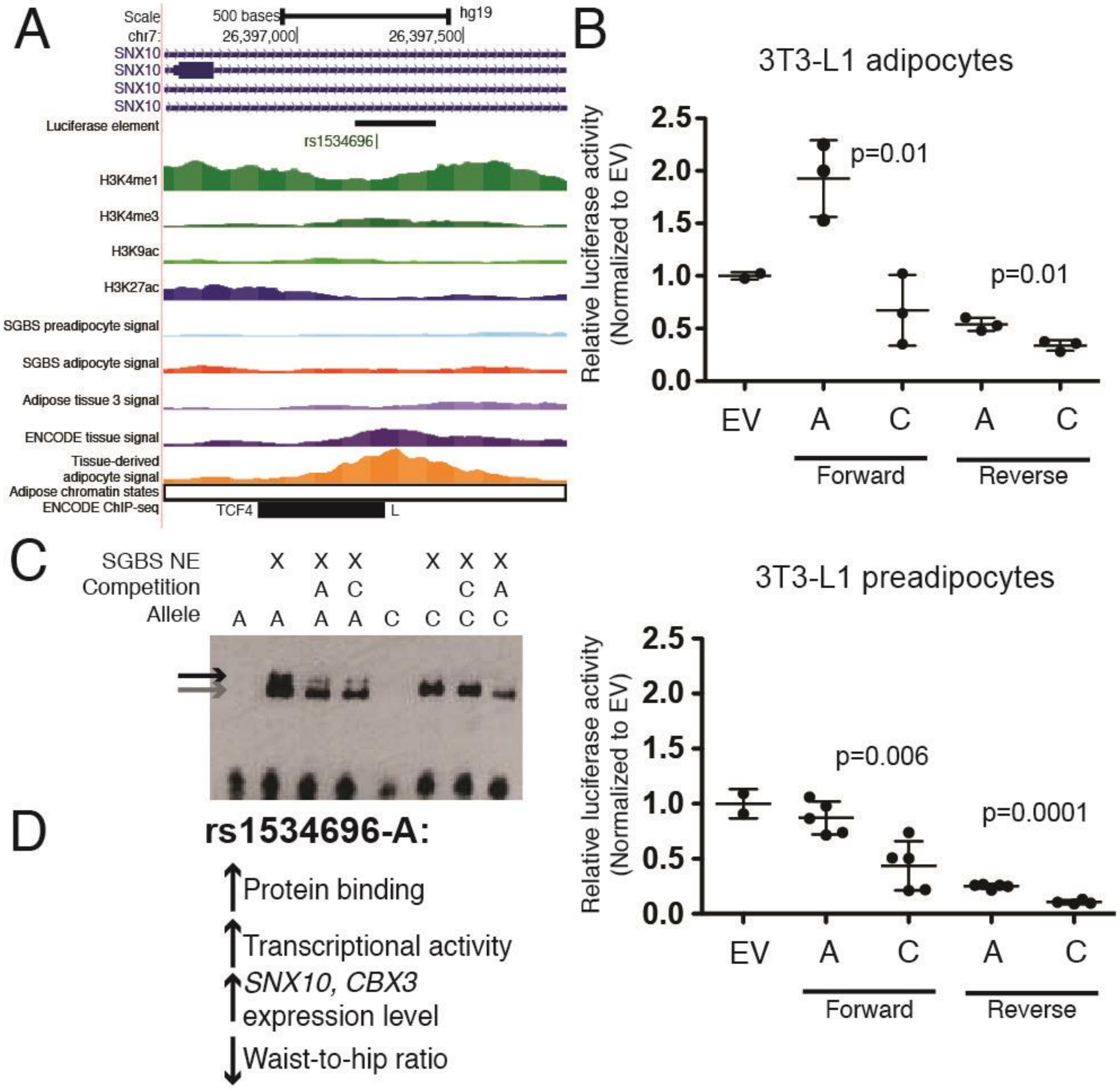


Figure 2.3. A variant at the *SNX10* WHR GWAS locus alters transcriptional activity and protein binding. (A) rs1534696 overlaps an ATAC-seq peak (adipose tissue 3 is shown in the figure; adipose tissue 1 shows stronger signal and peak) and is located in intron 2 of *SNX10*, transcribed left-to-right in the image, but is not located in a predicted regulatory region based on Roadmap chromatin states. TCF4 ENCODE ChIP-seq binding was observed in HepG2 cells. (B) The genomic region containing rs1534696-A shows increased transcriptional activity and allelic differences in transcriptional reporter luciferase assays in 3T3-L1 adipocytes and preadipocytes. The genomic region was cloned upstream of a minimal promoter and the luciferase gene. Dots represent the average of 2-3 technical replicates. Forward and reverse were designated with respect to the genome, so forward corresponds to left-to-right

in the image. P-values determined by Student’s t-test. EV, empty vector. (C) rs1534696-A shows increased protein binding in EMSA using SGBS preadipocyte nuclear extract. The black arrow shows allelic differences in protein binding. The gray arrow denotes non-specific binding observed for both rs1534696-A and rs1534696-G. (D) Summary of the direction of effect of rs1534696-A.

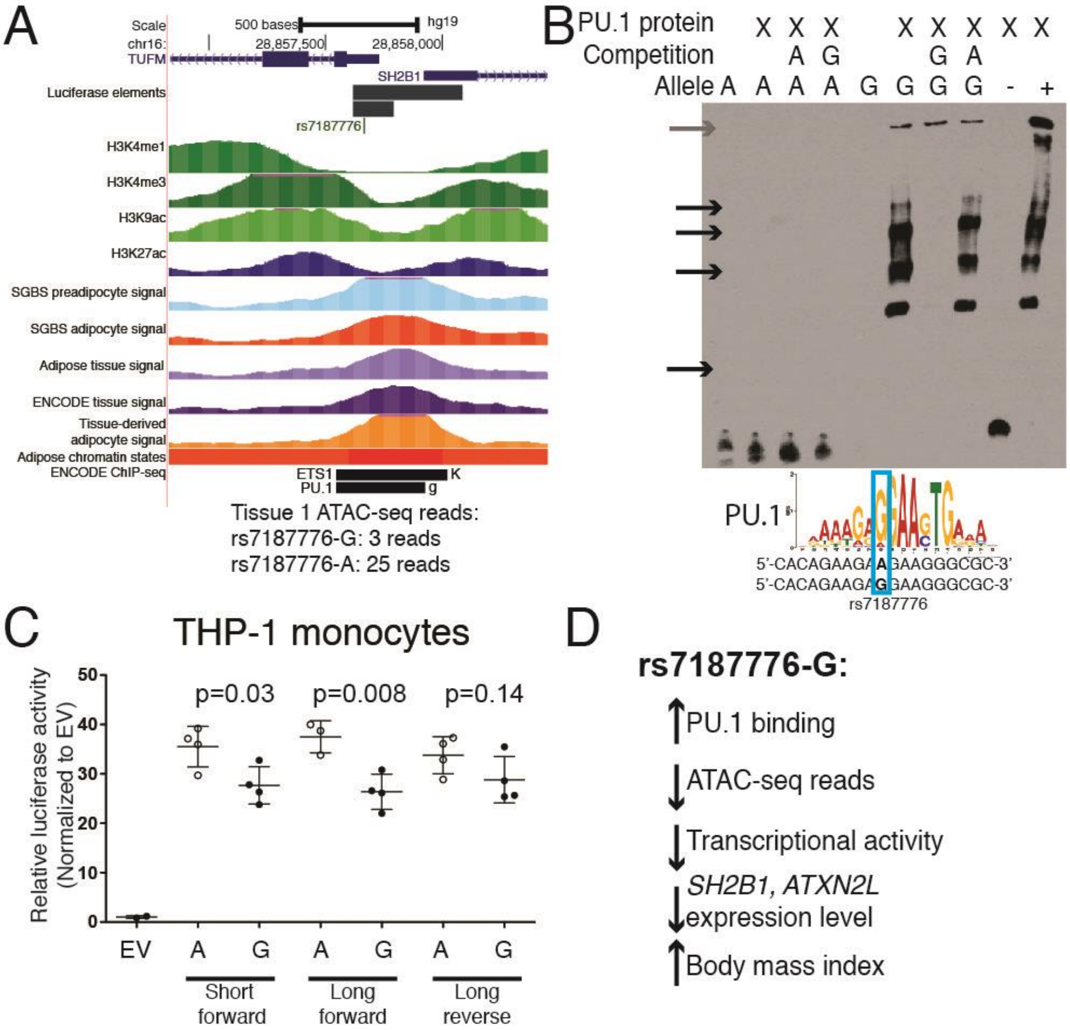


Figure 2.4. A variant at the *ATP2A1-SH2B1* BMI GWAS locus alters chromatin accessibility and PU.1 binding. (A) rs7187776 is located in the promoter of a long *SH2B1* isoform, transcribed left-to-right in the image; the 5'-UTR of *TUFM*, transcribed right-to-left in the image; and a region containing ATAC-seq peaks from multiple sources. ETS1 and PU.1 ENCODE ChIP-seq binding was observed in K562 and GM12891, respectively. Many additional transcription factor ChIP-seq peaks overlap this region in the ENCODE datasets. (B) A 19-nt probe containing rs7187776-G shows increased protein binding to purified PU.1 in EMSA, similar to a positive control probe containing the consensus PU.1 motif (+). A negative control probe (-) and a probe containing rs7187776-A showed no binding to PU.1. Black arrows indicate allele-specific protein binding, gray arrow indicates the well of the gel. Similar protein binding patterns and equal amounts of free DNA probe were observed using SW872 nuclear extract. PU.1 consensus motif from JASPAR 49. (C) The genomic region containing rs7187776-A shows increased transcriptional activity and allelic differences in THP-1 monocytes. The genomic region including part of the 3'

UTR of *TUFM* and part of the 5’ UTR of *SH2B1* was cloned upstream of a minimal promoter and the luciferase gene. Dots represent the average of two technical replicates. Forward and reverse designated with respect to the genome, so forward corresponds to left-to-right in the image. P-values determined by Student’s t-test. EV, empty vector. (D) Summary of the direction of effect of rs7187776-G.

Sample	Total reads	Aligned reads	Percent mitochondrial reads	Nuclear alignments	Remaining reads after duplicates removed	Number of peaks[b]
Tissue 1	129.5	87.4	8.5	80.0	70.6	58,550
Tissue 2	131.5	83.6	12.8	72.9	60.6	36,785
Tissue 3	119.3	70.5	11.9	62.2	57.1	49,962
Adipocytes 1[a]	382.6	275.9	2.1	268.6	90.4	184,455
Adipocytes 2[a]	245.1	172.9	1.9	168.7	84.1	172,247
Adipocytes 3[a]	253.7	181.0	1.5	177.2	87.5	191,141
Preadipocytes 1[a]	97.3	71.8	1.0	70.8	34.6	171,279
Preadipocytes 2[a]	75.1	54.1	1.1	53.3	30.5	139,911

Table 2.1. ATAC-seq alignment metrics of human adipose tissue and SGBS preadipocytes and adipocytes. Reads are reported in millions of reads. [a]Samples were sequenced using paired-end reads, but processed as single-end reads. [b]We identified 68,571 representative peaks across adipose tissue, 122,924 across SGBS preadipocytes, and 164,252 across SGBS adipocyte samples.

GWAS trait	GWAS locus	GWAS index variant	Colocalized eQTL gene(s)	eQTL index variant(s)	Variant in ATAC-seq peak	Total variants (r^2>.8) at locus	ATAC samples
Adiponectin	*GNL3*	rs2590838	*GNL3*, *NEK4*	rs35212380 rs7612511	rs1108842	21	1, 2, 3, Adipocytes, Preadipocytes
Coronary heart disease	*LIPA*	rs1412444	*LIPA*	rs1412445	rs1332328	8	3, Adipocytes, Preadipocytes
HDL cholesterol	*GSK3B*	rs6805251	*GSK3B*	rs334533	rs334558	61	1, 2, 3, Adipocytes, Preadipocytes
Intracranial aneurysm	*STARD13*	rs9315204	*KL*, *STARD13*	rs1998728 rs614691	rs1980781	22	1, 2, 3, Adipocytes, Preadipocytes
Serum metabolites	*NAT8*	rs13391552	*ALMS1*	rs6740766	rs4547554	180	1, 2, 3, Adipocytes
Proinsulin	*MADD*	rs10501320	*ACP2*, *FNBP4*	rs10501320 rs11039149	rs11039149	7	1, 2, 3, Adipocytes, Preadipocytes
Total cholesterol	*DOCK7-ANGPTL3*	rs2131925	*DOCK7*	rs631106	rs631106	237	1, Adipocytes, Preadipocytes
Triglycerides	*FADS1*	rs174548	*FADS1*	rs174555	rs174561	48	1, 2, 3, Adipocytes, Preadipocytes
Type 2 diabetes	*MPHOSPH9*	rs1727313	*C12orf65*, *CDK2AP1*, *SBNO1*	rs11057206 rs1616131 rs28583837	rs7485502	215	1, Adipocytes, Preadipocytes
WHRadjBMI	*SNX10*	rs1534696	*CBX3*, *SNX10*	rs1534696	rs1534696	1	1

Table 2.2. Selected variants at GWAS-eQTL colocalized loci that overlap ATAC-seq peaks. A subset of loci in which only one variant overlapped an ATAC-seq peak at a colocalized GWAS-eQTL locus in adipose tissue,[9] SGBS preadipocytes and/or SGBS adipocytes.

REFERENCES

1. Coelho, M., Oliveira, T., and Fernandes, R. (2013). Biochemistry of adipose tissue: an endocrine organ. Arch. Med. Sci. 9, 191-200.

2. Fernández-Veledo, S., Nieto-Vazquez, I., Vila-Bedmar, R., Garcia-Guerra, L., Alonso-Chamorro, M., and Lorenzo, M. (2009). Molecular mechanisms involved in obesity-associated insulin resistance: therapeutical approach. Arch. Physiol. Biochem. 115, 227-239.

3. Gustafson, B., Hedjazifar, S., Gogg, S., Hammarstedt, A., and Smith, U. (2015). Insulin resistance and impaired adipogenesis. Trends Endocrinol. Metab. 26, 193-200.

4. Porter, S.A., Massaro, J.M., Hoffmann, U., Vasan, R.S., O'Donnel, C.J., and Fox, C.S. (2009). Abdominal subcutaneous adipose tissue: a protective fat depot?. Diabetes Care 32, 1068-1075.

5. Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A.S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G.B., Gunnarsdottir, S. and Mouy, M. (2008). Genetics of gene expression and its effect on disease. Nature 452, 423-428.

6. Zhong, H., Beaulaurier, J., Lum, P.Y., Molony, C., Yang, X., Macneil, D.J., Weingarth, D.T., Zhang, B., Greenawalt, D., Dobrin, R., et al. (2010). Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes. PLoS Genet. 6, e1000932.

7. Greenawalt, D.M., Dobrin, R., Chudin, E., Hatoum, I.J., Suver, C., Beaulaurier, J., Zhang, B., Castro, V., Zhu, J., Sieberts, S.K., et al. (2011). A survey of the genetics of stomach, liver, and adipose gene expression from a morbidly obese cohort. Genome Res. 21, 1008-1016.

8. Nica, A.C., Parts, L., Glass, D., Nisbet, J., Barrett, A., Sekowska, M., Travers, M., Potter, S., Grundberg, E., Small, K., et al. (2011). The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. PLoS Genet. 7, e1002003.

9. Civelek, M., Wu, Y., Pan, C., Raulerson, C.K., Ko, A., He, A., Tilford, C., Saleem, N.K., Stancakova, A., Scott, L.J., et al. (2017). Genetic Regulation of Adipose Gene Expression and Cardio-Metabolic Traits. Am. J. Hum. Genet. 100, 428-443.

10. Willer, C.J., Schmidt, E.M., Sengupta, S., Peloso, G.M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M.L., Mora, S., et al. (2013). Discovery and refinement of loci associated with lipid levels. Nat. Genet. 45, 1274-1283.

11. DIAbetes Genetics Replication Meta-analysis Consortium, Asian Genetic Epidemiology Network Type 2 Diabetes (AGEN-T2D) Consortium, South Asian Type 2 Diabetes (SAT2D) Consortium, Mexican American Type 2 Diabetes (MAT2D) Consortium, Type 2 Diabetes Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples (T2D-GENES) Consortium. (2014). Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nat. Genet. 46, 234-244.

12. Shungin, D., Winkler, T.W., Croteau-Chonka, D.C., Ferreira, T., Locke, A.E., Magi, R., Strawbridge, R.J., Pers, T.H., Fischer, K., Justice, A.E., et al. (2015). New genetic loci link adipose and insulin biology to body fat distribution. Nature 518, 187-196.

13. Locke, A.E., Kahali, B., Berndt, S.I., Justice, A.E., Pers, T.H., Day, F.R., Powell, C., Vedantam, S., Buchkovich, M.L., Yang, J., et al. (2015). Genetic studies of body mass index yield new insights for obesity biology. Nature 518, 197-206.

14. Dahlman, I., Rydén, M., Brodin, D., Grallert, H., Strawbridge, R.J., and Arner, P. (2016). Numerous Genes in Loci Associated With Body Fat Distribution Are Linked to Adipose Function. Diabetes 65, 433-437.

15. Lynes, M.D., and Tseng, Y.H. (2017). Deciphering adipose tissue heterogeneity. Ann. N. Y. Acad. Sci.

16. Fischer-Posovszky, P., Newell, F.S., Wabitsch, M., and Tornqvist, H.E. (2008). Human SGBS cells - a unique tool for studies of human fat cell biology. Obes. Facts 1, 184-189.

17. Loft, A., Forss, I., Siersbæk, M.S., Schmidt, S.F., Larsen, A.-S.B., Madsen, J.G.S., Pisani, D.F., Nielsen, R., Aagaard, M.M., Mathison, A., et al. (2015). Browning of human adipocytes requires KLF11 and reprogramming of PPARgamma superenhancers. Genes Dev. 29, 7-22.

18. Schmidt, S.F., Larsen, B.D., Loft, A., Nielsen, R., Madsen, J.G.S., and Mandrup, S. (2015). Acute TNF-induced repression of cell identity genes is mediated by NFkappaB-directed redistribution of cofactors from super-enhancers. Genome Res. 25, 1281-1294.

19. Allum, F., Shao, X., Guénard, F., Simon, M.-M., Busche, S., Caron, M., Lambourne, J., Lessard, J., Tandre, K., Hedman, Å.K., et al. (2015). Characterization of functional methylomes by next-generation capture sequencing identifies novel disease-associated variants. Nat. Commun. 6, 7211.

20. ENCODE Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57-74.

21. Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature 518, 317-330.

22. Scott, L.J., Erdos, M.R., Huyghe, J.R., Welch, R.P., Beck, A.T., Wolford, B.N., Chines, P.S., Didion, J.P., Narisu, N., Stringham, H.M., et al. (2016). The genetic regulatory signature of type 2 diabetes in human skeletal muscle. Nat. Commun. 7, 11764.

23. Degner, J.F., Pai, A.A., Pique-Regi, R., Veyrieras, J.-B., Gaffney, D.J., Pickrell, J.K., Leon, S.D., Michelini, K., Lewellen, N., Crawford, G.E., et al. (2012). DNase I sensitivity QTLs are a major determinant of human expression variation. Nature 482, 390-394.

24. Kasowski, M., Kyriazopoulou-Panagiotopoulou, S., Grubert, F., Zaugg, J.B., Kundaje, A., Liu, Y., Boyle, A.P., Zhang, Q.C., Zakharia, F., Spacek, D.V., et al. (2013). Extensive variation in chromatin states across humans. Science 342, 750-752.

25. Kilpinen, H., Waszak, S.M., Gschwind, A.R., Raghav, S.K., Witwicki, R.M., Orioli, A., Migliavacca, E., Wiederkehr, M., Gutierrez-Arcelus, M., Panousis, N.I., et al. (2013). Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. Science 342, 744-747.

26. McVicker, G., van de Geijn, B., Degner, J.F., Cain, C.E., Banovich, N.E., Raj, A., Lewellen, N., Myrthil, M., Gilad, Y., and Pritchard, J.K. (2013). Identification of genetic variants that affect histone modifications in human cells. Science 342, 747-749.

27. Leung, D., Jung, I., Rajagopal, N., Schmitt, A., Selvaraj, S., Lee, A.Y., Yen, C.-A., Lin, S., Lin, Y., Qiu, Y., et al. (2015). Integrative analysis of haplotype-resolved epigenomes across human tissues. Nature 518, 350-354.

28. Kumasaka, N., Knights, A.J., and Gaffney, D.J. (2016). Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. Nat. Genet. 48, 206-213.

29. Gate, R.E., Cheng, C.S., Aiden, A.P., Siba, A., Tabaka, M., Lituiev, D., Machol, I., Gordon, M.G., Subramaniam, M., Shamim, M., et al. (2018). Genetic determinants of co-accessible chromatin regions in activated T cells across humans. Nat. Genet. 50, 1140-1150.

30. Astle, W.J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A.L., Mead, D., Bouman, H., Riveros-Mckay, F., Kostadima, M.A., et al. (2016). The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell 167, 1415-1429 e1419.

31. Roman, T.S., Marvelle, A.F., Fogarty, M.P., Vadlamudi, S., Gonzalez, A.J., Buchkovich, M.L., Huyghe, J.R., Fuchsberger, C., Jackson, A.U., Wu, Y., et al. (2015). Multiple Hepatic Regulatory Variants at the GALNT2 GWAS Locus Associated with High-Density Lipoprotein Cholesterol. Am. J. Hum. Genet. 97, 801-815.

32. Varshney, A., Scott, L.J., Welch, R.P., Erdos, M.R., Chines, P.S., Narisu, N., Albanus, R.D., Orchard, P., Wolford, B.N., Kursawe, R., et al. (2017). Genetic regulatory signatures underlying islet gene expression and type 2 diabetes. Proc. Natl. Acad. Sci. U. S. A. 114, 2301-2306.

33. Stancakova, A., Javorský, M., Kuulasmaa, T., Haffner, S.M., Kuusisto, J., and Laakso, M. (2009). Changes in insulin sensitivity and insulin release in relation to glycemia and glucose tolerance in 6,414 Finnish men. Diabetes 58, 1212-1221.

34. Laakso, M., Kuusisto, J., Stancakova, A., Kuulasmaa, T., Pajukanta, P., Lusis, A.J., Collins, F.S., Mohlke, K.L., and Boehnke, M. (2017). The Metabolic Syndrome in Men study: a resource for studies of metabolic and cardiovascular diseases. J. Lipid Res. 58, 481-493.

35. McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nat. Genet. 48, 1279-1283.

36. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods 10, 1213–1218.

37. Wabitsch, M., Brenner, R., Melzner, I., Braun, M., Möller, P., Heinze, E., Debatin, K., and Hauner, H. (2001). Characterization of a human preadipocyte cell strain with high capacity for adipose differentiation. International Journal of Obesity 25, 8.

38. Cannon, M.E., Duan, Q., Wu, Y., Zeynalzadeh, M., Xu, Z., Kangas, A.J., Soininen, P., Ala-Korpela, M., Civelek, M., Lusis, A.J., et al. (2017). Trans-ancestry Fine Mapping and Molecular Assays Identify Regulatory Variants at the ANGPTL8 HDL-C GWAS Locus. G3 (Bethesda) 7, 3217–3227.

39. Corces, M.R., Trevino, A.E., Hamilton, E.G., Greenside, P.G., Sinnott-Armstrong, N.A., Vesuna, S., Satpathy, A.T., Rubin, A.J., Montine, K.S., Wu, B., et al. (2017). An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. Nat. Methods 14, 959–962.

40. Lassmann, T., Hayashizaki, Y., and Daub, C.O. (2009). TagDust--a program to eliminate artifacts from next generation sequencing data. Bioinformatics 25, 2839–2840.

41. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359.

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

43. Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J. (2004). The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 32, D493-6.

44. Adey, A., Morrison, H.G., Asan, null, Xun, X., Kitzman, J.O., Turner, E.H., Stackhouse, B., MacKenzie, A.P., Caruccio, N.C., Zhang, X., et al. (2010). Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. Genome Biol 11, R119.

45. Jun, G., Flickinger, M., Hetrick, K.N., Romm, J.M., Doheny, K.F., Abecasis, G.R., Boehnke, M., and Kang, H.M. (2012). Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am. J. Hum. Genet. 91, 839–848.

46. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. 9, R137.

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

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

49. Mathelier, A., Fornes, O., Arenillas, D.J., Chen, C.-Y., Denay, G., Lee, J., Shi, W., Shyr, C., Tan, G., Worsley-Hunt, R., et al. (2016). JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res 44, D110-115.

50. McLeay, R.C., and Bailey, T.L. (2010). Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. BMC Bioinformatics 11, 165.

51. Buchkovich, M.L., Eklund, K., Duan, Q., Li, Y., Mohlke, K.L., and Furey, T.S. (2015). Removing reference mapping biases using limited or no genotype data identifies allelic differences in protein binding at disease-associated loci. BMC Med Genomics 8, 43.

52. Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics 27, 1017–1018.

53. Saint-André, V., Federation, A.J., Lin, C.Y., Abraham, B.J., Reddy, J., Lee, T.I., Bradner, J.E., and Young, R.A. (2016). Models of human core transcriptional regulatory circuitries. Genome Res 26, 385–396.

54. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and and, R.D. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079.

55. Pique-Regi, R., Degner, J.F., Pai, A.A., Gaffney, D.J., Gilad, Y., and Pritchard, J.K. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res 21, 447–455.

56. Schmidt, E.M., Zhang, J., Zhou, W., Chen, J., Mohlke, K.L., Chen, Y.E., and Willer, C.J. (2015). GREGOR: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. Bioinformatics 31, 2601–2606.

57. R Core Team (2015). R: A Language and Environment for Statistical Computing (Vienna, Austria: R Foundation for Statistical Computing).

58. Warnes, G.R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S. and Schwartz, M. (2016). gplots: Various R Programming Tools for Plotting Data. R package version 301.

59. Schick, U.M., Jain, D., Hodonsky, C.J., Morrison, J.V., Davis, J.P., Brown, L., Sofer, T., Conomos, M.P., Schurmann, C., McHugh, C.P., et al. (2016). Genome-wide Association Study of Platelet Count Identifies Ancestry-Specific Loci in Hispanic/Latino Americans. Am J Hum Genet 98, 229–242.

60. Kulzer, J.R., Stitzel, M.L., Morken, M.A., Huyghe, J.R., Fuchsberger, C., Kuusisto, J., Laakso, M., Boehnke, M., Collins, F.S., and Mohlke, K.L. (2014). A common functional regulatory variant at a type 2 diabetes locus upregulates ARAP1 expression in the pancreatic beta cell. Am J Hum Genet 94, 186–197.

61. Wu, T.D. and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26, 873-881.

62. Geijn, B. van de, McVicker, G., Gilad, Y., and Pritchard, J.K. (2015). WASP: allele-specific software for robust molecular quantitative trait locus discovery. Nat. Methods 12, 1061–1063.

63. Yee, T.W. (1996). Vector Generalized Linear and Additive Models With an Implementation in R. Journal of Royal Statistical Society, Series B 58, 481-493.

64. Sarjeant, K. and Stephens, J.M. (2012). Adipogenesis. Cold Spring Harb Perspect Biol 4, a008417.

65. Ambele, M.A., Dessels, C., Durandt, C., and Pepper, M.S. (2016). 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.

66. Körner, A., Wabitsch, M., Seidel, B., Fischer-Posovszky, P., Berthold, A., Stumvoll, M., Blüher, M., Kratzsch, J., and Kiess, W. (2005). Adiponectin expression in humans is dependent on differentiation of adipocytes and down-regulated by humoral serum components of high molecular weight. Biochem Biophys Res Commun 337, 540–550.

67. Segawa, K., Matsuda, M., Fukuhara, A., Morita, K., Okuno, Y., Komuro, R., and Shimomura, I. (2009). Identification of a novel distal enhancer in human adiponectin gene. J Endocrinol 200, 107–116.

68. Qiao, L., Maclean, P.S., Schaack, J., Orlicky, D.J., Darimont, C., Pagliassotti, M., Friedman, J.E., and Shao, J. (2005). C/EBPalpha regulates human adiponectin gene transcription through an intronic enhancer. Diabetes 54, 1744–1754.

69. Suzuki, M., Yamada, T., Kihara-Negishi, F., Sakurai, T., Hara, E., Tenen, D.G., Hozumi, N., and Oikawa, T. (2006). Site-specific DNA methylation by a complex of PU.1 and Dnmt3a/b. Oncogene 25, 2477–2488.

70. Suzuki, M., Yamada, T., Kihara-Negishi, F., Sakurai, T., and Oikawa, T. (2003). Direct association between PU.1 and MeCP2 that recruits mSin3A-HDAC complex for PU.1-mediated transcriptional repression. Oncogene 22, 8688–8698.

71. Kihara-Negishi, F., Yamamoto, H., Suzuki, M., Yamada, T., Sakurai, T., Tamura, T., and Oikawa, T. (2001). In vivo complex formation of PU.1 with HDAC1 associated with PU.1-mediated transcriptional repression. Oncogene 20, 6039–6047.

72. Yashiro, T., Kubo, M., Ogawa, H., Okumura, K., and Nishiyama, C. (2015). PU.1 Suppresses Th2 Cytokine Expression via Silencing of GATA3 Transcription in Dendritic Cells. PLoS One 10, e0137699.

73. Schäffler, A., Orsó, E., Palitzsch, K.D., Büchler, C., Drobnik, W., Fürst, A., Schölmerich, J., and Schmitz, G. (1999). The human apM-1, an adipocyte-specific gene linked to the family of TNF's and to genes expressed in activated T cells, is mapped to chromosome 1q21.3-q23, a susceptibility locus identified for familial combined hyperlipidaemia (FCH). Biochem Biophys Res Commun 260, 416–425.

74. Yamauchi, T., Kamon, J., Minokoshi, Y., Ito, Y., Waki, H., Uchida, S., Yamashita, S., Noda, M., Kita, S., Ueki, K., et al. (2002). Adiponectin stimulates glucose utilization and fatty-acid oxidation by activating AMP-activated protein kinase. Nat Med 8, 1288–1295.

75. Yokota, T., Meka, C.S.R., Medina, K.L., Igarashi, H., Comp, P.C., Takahashi, M., Nishida, M., Oritani, K., Miyagawa, J.-I., Funahashi, T., et al. (2002). Paracrine regulation of fat cell formation in bone marrow cultures via adiponectin and prostaglandins. J Clin Invest 109, 1303–1310.

76. Musri, M.M., Corominola, H., Casamitjana, R., Gomis, R., and Párrizas, M. (2006). Histone H3 lysine 4 dimethylation signals the transcriptional competence of the adiponectin promoter in preadipocytes. J Biol Chem 281, 17180–17188.

77. Wei, G.-H., Badis, G., Berger, M.F., Kivioja, T., Palin, K., Enge, M., Bonke, M., Jolma, A., Varjosalo, M., Gehrke, A.R., et al. (2010). Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. EMBO J 29, 2147–2160.

78. Giuranna, J., Volckmar, A.-L., Heinen, A., Peters, T., Schmidt, B., Spieker, A., Straub, H., Grallert, H., Müller, T.D., Antel, J., et al. (2018). The Effect of SH2B1 Variants on Expression of Leptin- and Insulin-Induced Pathways in Murine Hypothalamus. Obes Facts 11, 93–108.

79. Volckmar, A.-L., Bolze, F., Jarick, I., Knoll, N., Scherag, A., Reinehr, T., Illig, T., Grallert, H., Wichmann, H.-E., Wiegand, S., et al. (2012). Mutation screen in the GWAS derived obesity gene SH2B1 including functional analyses of detected variants. BMC Med Genomics 5, 65.

CHAPTER 3: GENETIC EFFECTS ON LIVER CHROMATIN ACCESSIBILITY IDENTIFY DISEASE REGULATORY VARIANTS

Introduction

Genome-wide association studies (GWAS) have identified thousands of loci associated with complex traits, but the causal variants, molecular mechanisms, target genes, and tissues of action for most loci have not been characterized. Gene expression quantitative trait loci (eQTL) studies have been instrumental in identifying plausible target genes and tissues for GWAS loci[1]. Chromatin conformation capture techniques, such as Hi-C, have identified variants at GWAS loci that physically interact with gene promoters[2]. However, additional approaches are needed to further pinpoint functional variants and to identify how these variants alter gene expression.

Variants at GWAS loci are enriched in transcriptional regulatory elements, which are typically marked by chromatin accessibility, in trait-relevant tissues[3]. Recent studies have identified chromatin accessibility QTL (caQTL), many of which overlap transcription factor (TF) binding sites and motifs[4–9]. A subset of caQTL are colocalized with eQTL and GWAS loci, suggesting that variants at these loci impact gene expression and GWAS traits by altering chromatin accessibility[4–9]. However, caQTL have been mapped in a limited set of human tissues. Mapping caQTL in additional tissues and cell types is valuable to characterize the transcriptional regulatory mechanisms for a larger set of GWAS loci.

Liver is involved in numerous processes, including lipid metabolism, glucose storage, drug metabolism, and immune response[10]. Several studies have mapped eQTL in liver tissue, and liver eQTL are colocalized with GWAS loci for lipid, drug response, and other traits[11–13]. Lipid GWAS loci are enriched in regulatory chromatin states, including enhancers and promoters, in HepG2 hepatocyte cells[14]. QTL for the active regulatory element histone marks H3K27ac and H3K4me3 have been identified in liver tissue, including a subset colocalized with liver eQTL and GWAS loci[12]. Chromatin accessibility marks active regions containing H3K4me3 and H3K27ac, as well as poised promoters and enhancers that often do not display these histone marks[15,16]. Consequently, mapping caQTL in liver tissue can help functionally characterize GWAS loci that act by altering gene expression in liver.

In this study, we jointly mapped genotypes, gene expression, and chromatin accessibility in liver tissue from 20 organ donors and identified caQTL in liver tissue. We predicted the impact of caQTL variants on TF binding and predicted caQTL target genes using four approaches. Finally, we used caQTL, TF binding motifs, and target gene links to predict mechanisms at GWAS loci for multiple traits.

Material and Methods

Liver tissue samples

Healthy human liver tissue was collected from deceased organ donors at St. Jude (Memphis, TN) as part of the National Institutes of Health Liver Tissue Cell Distribution System. Tissue was collected with the approval of institutional review boards (IRBs) and the University of North Carolina (Chapel Hill, NC) approved their use for this study as non-human subjects research.

Genotyping and imputation

We genotyped over 2.5 million variants using the Infinium Omni2.5Exome-8 BeadChip array v1.3 (Illumina, San Diego, CA, USA) at the NHGRI Genomics Core facility. Overall genotyping call rates ranged from 99.0-99.6%. We mapped the Illumina array probe sequences to the hg19 genome assembly[17] using novoalign (see web resources), excluding variants with ambiguous probe alignments and variants with 1000 Genomes (1000G) phase 3 minor allele frequency (MAF) >.01 within 7 bp of the 3' end of probes. No individuals were related at a 3rd-degree relationship threshold using KING v1.4[18]. Prior to testing for population stratification, we removed variants with minor allele count < 5 and that were found within regions of unusually high linkage disequilibrium (LD, see web resources) using VCFtools v0.1.14[19], and selected distinct (r^2<0.2) variants using PLINK v1.9[20]. We did not observe evidence of population stratification using principal component analysis (PCA) of 65,113 genotypes using PLINK v1.9[20].

Prior to genotype imputation, we combined the genotypes of the samples in this study with genotypes from 173 samples from a separate study genotyped on similar chips and removed variants that met the following criteria: allele frequency difference >20% with 1000G phase 3 Europeans, palindromic variants with MAF>.2, genotype missingness > 2.5%, and extreme deviation from Hardy-Weinberg Equilibrium ($p<1x10^{-4}$). Using the Michigan Imputation Server[21], we phased 1,825,454 variants using Eagle v2.3[22] and imputed missing genotypes using minimac3[21] with the Haplotype Reference Consortium (hrc.r1.1.2016) panel[23]. We retained variants with imputation r^2>.3 for downstream analyses.

RNA-seq library preparation, read alignment, and selection of expressed genes

We extracted and purified total RNA from frozen liver tissue using Trizol as previously described[24]. Paired-end, strand-specific, poly-A RNA sequencing (RNA-seq) was performed on an Illumina NovaSeq 6000 with 2x151 bp cycles. RNA-seq reads were trimmed using Trimmomatic[25] and aligned to the hg19 genome assembly[17] using STAR v2.53[26] with default parameters. Using verifyBamID v1.1.1[27], we found no evidence of library contamination or sample swaps. Expression levels of GENCODE v19[28] genes were quantified using QoRTs v1.2.42[29]. We classified genes as expressed if the median transcripts per million (TPM) across the 20 individuals was at least 1. We performed principal component analysis on gene counts normalized by library size and variance-stabilized using DESeq2[30].

ATAC-seq library preparation

Nuclei were isolated as previously described[31] with the following modifications. We pulverized 50-mg pieces of frozen human liver tissue in liquid nitrogen using a Cell Crusher (CellCrusher, Cork Island), homogenized the tissue powder in ice cold 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) using a 1-mL dounce for 40 strokes, and rotated for 5 minutes at 4^{o}C. We filtered the solution through a Miracloth (Calbiochem, San Diego, Ca USA), centrifuged at 1100g for 10 minutes at 4^{o}C, washed the pellet with 250-uL NIB containing 0.5% Triton-X, centrifuged at 500g for 5 minutes at 4^{o}C, and resuspended the pellet in 250-uL of resuspension buffer (10 mM Tris-HCl, 10 mM NaCl, 3 mM $MgCl_2$, pH 7.4). After counting isolated nuclei, we pelleted 50,000 nuclei at 500g for 5 minutes at 4^{o}C for each of three replicate ATAC-seq libraries per sample. Libraries were prepared using Nextera kits (Illumina) as previously described[32].

ATAC-seq read alignment and identification of consensus peaks

We trimmed ATAC-seq reads to a uniform length of 126 bp using cutadapt[33] and aligned reads as previously described[34]. Briefly, we trimmed sequencing adapters using CTA (see web resources) and aligned reads to the hg19 human genome[17] using BWA-MEM (see Web Resources). We selected properly paired autosomal alignments with high mapping quality (mapq>30) with samtools[35] and removed duplicate alignments using Picard (see web resources). We used ataqv[36] to generate ATAC-seq quality metrics and confirmed ATAC-seq libraries corresponded to the correct genotypes using verifyBamID[27].

To assess reproducibility of libraries from the same individual, we called narrow peaks separately for each library using MACS2[37] with parameters –nomodel -100 –extsize 200, then merged peaks across all individuals and replicates using BEDTools merge[38], and selected peaks present in at least 3 libraries. We counted the number of reads overlapping each peak using featureCounts[39] and performed library size normalization and variance-stabilization using DESeq2[30]. We computed pairwise Pearson correlations of normalized counts for all peaks and for the 10,000 most variable peaks between libraries and visualized the results using the heatmap.2 function in the gplots R package[40] (see Web Resources). Libraries from the same individual were highly correlated, so we merged the alignment .bam files across libraries for each individual using SAMtools[35].

To identify consensus peaks, we converted the merged .bam files for each individual to .bed files using BEDTools[38], called narrow peaks for each individual using MACS2[37] with parameters –nomodel –shift -100 –extsize 200 –keep-dup all, and removed peaks overlapping blacklisted regions[38,41]. We then merged peaks across individuals using BEDTools[38] and defined consensus peaks as merged peaks that shared at least 1 base with a peak present in samples from at least 3 individuals.

Overlap of consensus peaks with roadmap chromatin states

We computed overlap of ATAC-seq consensus peaks with chromatin states in adult liver tissue from the Roadmap Epigenomics Consortium[3]. We defined the following states: promoter (1_TssA, 2_TssFlnk, 3_TssFlnkU, 4_TssFlnkD, 14_TssBiv), transcribed (5_Tx, 6_TxWk), enhancer (7_EnhG1, 8_EnhG2, 9_EnhA1, 10_EnhA2, 11_EnhWk, 15_EnhBiv), polycomb (16_ReprPC, 17_ReprPCWk), heterochromatin (13_Het), ZNF repeats (12_ZNF/Rpts), and quiescent (18_Quies). For each consensus ATAC peak, we computed the fraction of bases that overlapped each chromatin state in liver tissue (Roadmap epigenome ID E066) using BEDTools coverage[38]. We assigned each peak to the chromatin state with which it shared the most bases, except for the quiescent state; we only assigned a peak to a quiescent state if all bases of a peak were found within a quiescent state. If a peak shared most, but not all, of its bases with a quiescent state, we assigned the peak to the state with the second highest coverage.

Selection of transcription factor motifs

We obtained transcription factor (TF) binding motifs from Cis-BP v1.02[42], selected all directly determined motifs per TF or the best inferred motif when a TF did not have a directly determined motif (TF_Information.txt dataset from Cis-BP), and restricted to motifs for TFs expressed in liver tissue from GTEx v8 (median transcripts per million ≥1). We performed clustering to remove redundant motifs using RSAT matrix-clustering[43] with parameters -hclust_method average -calc sum -metric_build_tree Ncor -lth w 5 -lth cor 0.8 -lth Ncor 0.8 -quick, resulting in 516 motif clusters. For each motif cluster, we defined the representative TF as the TF with the highest expression in liver tissue from GTEx v8 (measured in median TPM) and the representative motif as the motif assigned to the representative TF. If multiple motifs existed for the representative TF in a given cluster, we selected the motif with the highest information content. Although we often use the representative TF name to refer to motif clusters for convenience, any TF in the cluster may bind at a given locus. Therefore, we listed all expressed TFs in the cluster in supplemental tables. Some TFs were assigned as the representative TF for multiple clusters, potentially representing distinct binding profiles for the same TF. We retained all of these clusters unless otherwise noted.

Enrichment of TF motifs and ChIP-seq binding sites in ATAC peaks

We tested for enrichment of 286 non-redundant transcription factor (TF) motifs in consensus ATAC peaks using Analysis of Motif Enrichment (AME)[44] with parameters –control –shuffle-- –kmer 2 –scoring max –hit-lo-fraction 0.75. We classified motifs with E-value < $1x10^{-100}$ as significantly enriched. We derived the 286 motifs from the set of 516 non-redundant motifs (see “Selection of transcription factor motifs”) by selecting the motif with the highest information content per TF.

We downloaded liver tissue ChIP-seq peaks for 17 TFs[45] from the ENCODE portal[46] (sample accession ENCDO882MMZ) and defined binding sites as the summit of the ChIP-seq peaks. We computed the number of binding sites overlapping consensus ATAC-seq peaks for each TF using BEDTools intersect[38]. To determine if the number of binding sites overlapping ATAC peaks was more than expected given their genomic frequency, we permuted binding sites across the genome 1,000 times excluding blacklisted regions[41] using BEDTools shuffle[38] and computed the number of overlaps for each permutation. We calculated an enrichment p-value by determining the fraction of permuted overlaps that were equal to or greater than the observed number of overlaps.

Chromatin accessibility QTL identification

We identified chromatin accessibility quantitative trait loci (caQTL) using RASQUAL[5], which jointly tests for association of genotype with peak accessibility across individuals and allelic imbalance in read counts at heterozygous variants within the same individual. We selected ~4 million genetic variants with MAF > .1 in the 20 individuals and within 100 kb of consensus peak centers and restricted to variants present in 1000 genomes phase 3 Europeans. To quantify peak accessibility across samples, we extended alignments 100 bp from either end of the 5’-

most base using BEDTools[38] and counted the number of alignments overlapping each peak using featureCounts[39]. We used DESeq2 size factors[30] to adjust for library size and the gcCor.R script provided with RASQUAL[5] to adjust for GC bias. To identify global variation between samples that may confound caQTL detection, we performed PCA on peak counts adjusted for library size and variance-stabilized by DESeq2[30,40]. We ran RASQUAL using differing numbers of PCs as covariates ranging from 0 to 10 in increments of 1 and selected 2 PCs to maximize the number of peaks with a caQTL at false discovery rate (FDR) of 5%. We performed multiple testing correction using the two-step eigenMT-BH procedure[47]. First, we used eigenMT[48] with the 1000 genomes phase 3 European reference panel to adjust for the differing variant density around each peak, taking into account the LD between variants. Second, we selected the most significant eigenMT-adjusted p-value for each peak and calculated FDR using the Benjamini-Hochberg (BH) procedure[49]. We selected significant caQTL with FDR<5% and correlation r^2 between prior and posterior genotypes >0.8. We refer to peaks with a significant caQTL as caPeaks. We repeated the caQTL analysis using ~0.6 million variants within 1 kb of peak centers. Unless otherwise noted, all downstream analyses were performed using caQTL identified using variants within 1 kb of peak centers.

ATAC-seq allelic imbalance and comparison to caQTL effect sizes

To assess the robustness of the caQTL, we used an alternative method for calculating allelic imbalance (AI). We removed ATAC-seq reads exhibiting allelic mapping bias using the WASP mapping pipeline[50] and counted the number of ATAC-seq reads mapping to each allele at heterozygous variants using ASEReadCounter[51] with the option –min-base-quality 30. We removed variants that had aligned bases other than the two genotyped alleles and selected variants with >=10 total reads, >=3 reads per allele, and that were heterozygous in >=3 individuals. After pooling reads across individuals, each variant had a minimum of 30 total reads and 9 reads per allele. We fit allele counts to a beta-binomial distribution using the VGAM R package[40,52], tested for AI using a two-tailed beta-binomial test, and adjusted for multiple testing using the BH procedure.

To compare effect sizes of AI variants and caQTL signals, we selected caQTL that had at least one AI variant in strong LD (r^2>0.8, 1000G phase 3 Europeans) with the caQTL lead variant and that resided within the caPeak; LD was calculated using PLINK v1.9[20]. For each caQTL with a linked AI variant, we selected the AI variant with the strongest evidence of AI (smallest beta-binomial p-value). For both methods, we calculated an effect size by subtracting 0.5 from the estimated fraction of reads containing the alternate allele, which is the RASQUAL PI value for caQTL. An alternate allele fraction of 0.5 corresponds to an equal number of reads on each allele, which is an effect size of 0. We then computed the Pearson correlation between the absolute value of effect sizes between the caQTL and AI variants.

caQTL enrichment in chromatin states

To identify which regulatory elements preferentially contain caPeaks, we compared the number of caPeaks (FDR<5%) and non-caPeaks (eigenMT-adjusted p>0.5) assigned to various liver tissue chromatin states from Roadmap[3]. We tested if caQTL variants were enriched in liver tissue chromatin states relative to variants matched for MAF, number of LD proxies, and distance to nearest gene using the logistic regression model implemented in GARFIELD[53]. We defined caQTL variants as significantly enriched in a chromatin state if the p-value for the logistic regression beta was less than the Bonferroni-corrected threshold (for 7 chromatin states) of 7.1×10^{-3} and the odds ratio was greater than 1. We defined caQTL variants as significantly depleted in a chromatin state if $p<7.1 \times 10^{-3}$ and odds ratio<1.

Transcription factor motif disruption by caQTL variants

We selected 5,378 caQTL variants that resided within the caPeak using BEDTools intersect[38] and that were in strong LD ($r^2>0.8$, calculated with PLINK[20]) with the caQTL lead variant. To ensure that each motif occurrence was disrupted by only one variant, we removed 793 variants within 30 bp of another caQTL variant, resulting in 4,585 variants. For both alleles of each caQTL variant, we extracted the nucleotide sequence for the region containing the variant and the 30 nucleotides on either side of the variant using the BEDTools slop and getfasta tools[38]. We scanned these sequences for occurrences of 516 non-redundant TF motifs using Find Individual Motif Occurrences (FIMO)[54] with parameters --thresh 0.01 --max-stored-scores 1000000 --no-qvalue --skip-matched-sequence –text and only retained motif occurrences that overlapped caQTL variant positions. For each motif-variant pair, we selected the strongest motif match (smallest p-value) per allele and only retained motif occurrences that matched strongly to at least one allele ($p<1 \times 10^{-4}$). If different motifs for the same representative TF overlapped the same variant, we selected the motif with the strongest match.

Similar to a recent study[55], we quantified the difference in motif match between alleles of a variant using the log ratio of FIMO p-values. The FIMO p-value for a given motif occurrence is the probability of observing a motif occurrence with the same or greater score, which inherently accounts for differences in score distributions between different motifs. For a given variant-motif pair, we define motif disruption as $\log_{10}(p_{aw}) - \log_{10}(p_{as})$, where p_{aw} and p_{as} are the FIMO p-values for the alleles with the weaker and stronger motif match, respectively. As motif disruption is always positive, we classified a motif as disrupted if motif disruption was >1, corresponding to a 10-fold difference in the FIMO p-values between alleles.

We identified motifs whose disruption was associated with caQTL status using logistic regression. To generate a set of non-caQTL variants, we first selected peaks with no evidence of genetic regulation (caQTL eigenMT-adjusted p>0.5), that overlapped at least one variant tested in the caQTL analysis, and that were similar to caPeaks in GC content (±5%), peak width (±20%), and distance to nearest transcription start site (TSS) of a protein-coding gene in GENCODE[28] (±20%). We identified 10 non-caPeaks for >99% of the caPeaks used in the motif disruption analysis and defined non-caQTL variants as the 50,054 variants that were within non-caPeaks and were located more than 30 bp from the nearest variant. We tested these non-caQTL variants for TF motif disruption using the same procedure as for caQTL variants and restricted analysis to the 109 motifs with at least 20 disruptions by caQTL variants. For each representative TF, we selected the motif with the most disruptions by caQTL variants to ensure that we used only one motif per representative TF. We then regressed caQTL status (1=caQTL, 0=non-caQTL) against motif disruption status (1=disrupted, 0=not disrupted) for each motif-variant pair using logistic regression. We classified motif disruption as associated with caQTL status if the p-value for the logistic regression beta was less than the Bonferroni-corrected threshold (for 109 motifs) of 4.6×10^{-4}. Because residual differences may exist in peak GC content, width, and distance to nearest protein coding TSS, we performed logistic regression with and without these features as covariates and obtained the same set of significantly enriched motifs.

caPeak target gene identification

We used four methods to identify target genes for caPeaks: proximity to a gene's TSS, overlap of caPeaks with promoter-centered chromatin contacts, correlation of caPeaks with peaks at gene promoters or with gene expression, and colocalization of caQTL and eQTL. We excluded genes from the analysis if their Entrez ID did not map to exactly one Ensembl ID (eQTL data only) or if their symbol (common name) didn't map to exactly one Ensembl ID. When combining results across the four methods, we matched genes based on Ensembl ID.

TSS proximity: We classified a caPeak as TSS proximal if it was located within 2 kb upstream and 1 kb downstream of the TSS of any of the 13,782 expressed genes (median TPM>1) in our 20 liver samples using BEDTools closest[38].

Promoter-centered chromatin contacts: We obtained promoter-distal and promoter-promoter contacts mapped in liver tissue using promoter capture Hi-C from a recent study[2] (see web resources). Using described filtering criteria[2], we selected contacts with p-value<0.01 and interaction frequency >=5. We identified caPeaks overlapping distal ends of promoter-distal contacts or either end of promoter-promoter contacts using BEDTools intersect[38].

Correlation of caPeaks with promoter peaks and gene expression: We classified an ATAC-seq peak as the promoter peak for an expressed gene if it was the closest peak to the TSS of the gene and it was within 2 kb upstream and 1 kb downstream of the TSS[56]. A promoter peak may or may not be a caPeak. We identified promoter peaks for 10,074 of 13,782 expressed genes. For each gene with a promoter peak, we identified caPeaks for correlation that were within 1 Mb of the gene's TSS but that were not TSS proximal. For peak and gene counts, we performed library size normalization and variance-stabilization using DESeq2[30] and GC bias-correction using RASQUAL[5]. We additionally adjusted peak counts by the fraction of reads in peaks, which was strongly correlated with the first ATAC-seq PC, and gene counts by the percent of reads mapping to the most expressed gene and the percent of reads mapping to the top 10 most expressed genes (geneDiversityProfile_top1pct and geneDiversityProfile_top10pct metrics from QoRTs[29]), which were strongly correlated with RNA-seq PCs 1 and 2 respectively, using the limma removeBatchEffects function[57]. We then computed the Spearman correlation between (1) gene expression and caPeaks and (2) promoter peaks and caPeaks using the cor.test function in R[40]. We adjusted for multiple testing using the BH procedure[49] and classified correlations with FDR<5% as significant.

Colocalization of caQTL and eQTL: We obtained liver tissue expression quantitative trait loci (eQTL) for 15,668 genes (FDR<5%) from a meta-analysis of 1,183 individuals[11] and restricted to the 15,418 eQTL on autosomes. We calculated LD and haplotype phase between eQTL and caQTL lead variants using PLINK[20] v1.9 and classified signals as colocalized if these lead variants exhibited strong pairwise LD (r^2>0.8, 1000G phase 3 Europeans). To compare the direction of effect for colocalized caQTL and eQTL, we compared the sign of the caQTL effect size (RASQUAL pi statistic - 0.5) and the eQTL effect size (meta T statistic).

Colocalization of caQTL and GWAS signals

We downloaded the NHGRI-EBI GWAS catalog[58] on October 28, 2019, extracted only single variant associations, and converted variant genomic coordinates from GRCh38[17] to GRCh37 (hg19) using liftOver[59]. We extracted variants associated with 19 trait groups ($p<5x10^{-8}$) relevant to liver function and cardiometabolic diseases: liver enzymes, high-density lipoprotein cholesterol (HDL), low-density lipoprotein cholesterol (LDL), total cholesterol (TC), triglycerides (TG), cardiovascular disease (CVD), hypertension/blood pressure (HTBP), type 2 diabetes (T2D), insulin, glucose, glycated albumin, serum albumin, glycated hemoglobin (HbA1c), C-reactive protein (CRP), bilirubin, body mass index (BMI), waist-hip ratio adjusted for BMI (WHRadjBMI), liver injury, and non-alcoholic fatty liver disease (NAFLD). We extracted alleles for each variant from the dbSNP[60] build 151 common variant set (see web resources), restricting to biallelic variants. To select one variant per association signal, we performed LD clumping separately for each trait using swiss (see web resources); variants in strong LD (r^2>0.8, 1000G phase 3 Europeans) and within 1 Mb of a variant with a more significant p-value were removed. We calculated LD between lead caQTL and GWAS variants using PLINK[20] v1.9 and classified signals in high LD (r^2>0.8) as colocalized.

Transcriptional activity reporter assays

HepG2 hepatocyte cells were cultured in MEM-alpha supplemented with 10% FBS and 1 mM sodium pyruvate, THP-1 monocyte cells were cultured in RPMI-1640 supplemented with 10% FBS, and both cell types were maintained at 37°C with 5% CO_2. To test haplotypic differences in transcriptional activity, we designed PCR primers (5'-TATGTTGCACAGGCTGGTCT and 5'- GGCAATAACGCCCACCTC) to amplify a 666-bp DNA element (chr16:11,644,551 – 11,645,216) spanning the ATAC-seq peak and containing variants rs3784924, rs11644920, and rs57792815, and we generated PCR products using DNA from individuals homozygous for both haplotypes. We cloned the derived PCR products into luciferase reporter vector pGL4.23 (Promega) as described previously[61]. The day before transfection, we plated 120,000 HepG2 cells, and on the day of transfection, we plated 300,000 THP-1 cells. We co-transfected four to five sequence-verified constructs with phRL-TK Renilla reporter vector using lipofectamine 3000 (Life Technologies) following the manufacturer's protocol. To induce differentiation into macrophages, we added 100 nM 1α,25-Dihydroxyvitamin D_3 (Sigma) to the THP-1 cells at the time of transfection. To obtain activated macrophages, we added 100 ng/ml lipopolysaccharides (Sigma) to vitamin D_3-treated cells 24 hours after transfection and incubated cells for an additional 24 hours. Firefly luciferase activity was measured 48 hours post-transfection and normalized to Renilla activity to adjust for differences in transfection efficiency. Fold-changes in luciferase activity were calculated relative to an empty pGL4.23 vector, and statistical differences in activity were determined using two-tailed Student's t-tests.

Electrophoretic mobility shift assays (EMSA)

We designed and annealed 3 biotin-labeled and unlabeled 17-bp complementary oligonucleotide probes centered on each of variants rs3784924, rs11644920, and rs57792815. We conducted EMSAs using the LightShift

Chemiluminescent EMSA kit (Thermo Scientific) following the manufacturer's protocol. The binding reactions consisted of 6 μg HepG2 nuclear extract (NE-PER Kit, Thermo Fisher Scientific), 1 μg poly(dI-dC), 1x binding buffer, and 400 fmol biotinylated oligonucleotide as described previously[61]. To test the specificity of the protein complexes to each allele, we added 25-fold excess unlabeled probes, and for supershift assays, we added 6 ug of antibody to the binding reactions. We tested antibodies for ATF2, IRF1, IRF2, IRF3, IRF7, FOXA2, P300, SPI1, STAT4. Protein-DNA complexes were resolved by gel electrophoresis and transferred and detected by chemiluminescence as described previously[61].

Results

Joint profiling of gene expression and chromatin accessibility in human liver tissue

We obtained liver tissue from 20 deceased donors from the St. Jude liver bank and profiled gene expression using RNA-seq and chromatin accessibility using ATAC-seq[32] (**Figure 3.1A**). We identified 13,782 expressed genes. By generating triplicate ATAC-seq libraries, we obtained an average of 204 million high-quality autosomal ATAC-seq alignments (HQAA) per sample and used these reads to identify 223,265 consensus accessible chromatin regions (peaks) with median peak width of 617 base pairs (**Figure 3.1B**).

To predict the regulatory function of ATAC-seq peaks, we assigned peaks to liver tissue chromatin states from the Roadmap Epigenomics Project[3] and tested for enrichment of transcription factor (TF) binding sites and motifs in peaks. Among all 223,265 peaks, 34% were located in enhancers and 10% in promoters, and among the 50,000 most accessible peaks, 54% were located in enhancers and 38% in promoters (**Figure 3.1C**). We found 90 TF motifs enriched in peaks (E-value<$1x10^{-100}$), including motifs for HNF4G, FOXA family members (HNF3), CEBPB[62], the multifaceted protein CTCF[63], and KLF family members, which regulate numerous processes in liver[64]. Of 17 TFs with ChIP-seq data in liver tissue[45], binding sites for all TFs were significantly enriched (permutation p<$1x10^{-3}$) in ATAC peaks, and 11 TFs had over 90% of their binding sites within ATAC peaks, similar to previous findings[15]. Taken together, ATAC peaks marked previously annotated transcriptional regulatory elements and TF binding sites in liver tissue.

We next determined if genes with ATAC peaks at their transcription start site (TSS) were more likely to be expressed compared to genes without TSS peaks. A larger proportion of expressed genes had an ATAC peak directly overlapping the TSS (74%) compared to non-expressed genes (24%). Similarly, genes with a peak at the TSS tended to have higher expression than genes without a peak at the TSS (**Figure 3.1D**; Kolmogorov-Smirnov test, $p<2.2 \times 10^{-16}$). Together, the data provide high-quality gene expression and chromatin accessibility profiles in human liver tissue.

Identification of genetic variants associated with liver chromatin accessibility

We identified chromatin accessibility quantitative trait loci (caQTL) using RASQUAL[5] and two distance thresholds: variants within 100 kilobases (kb) and within 1 kb of peak centers (**Figure 3.2A**). Testing variants within 100 kb of peak centers, we identified a significant caQTL for 1,770 peaks (caPeaks), corresponding to 1,740 unique lead caQTL variants (**Figure 3.2A**). For a substantial portion of caPeaks, the lead caQTL variant was either within the caPeak (n=654, 37%) or was within 1 kb of the caPeak center (n=692, 39%). Testing variants within 1 kb of peak centers, we identified a significant caQTL for 3,123 peaks (**Figure 3.2A**). We used this set of 3,123 caQTL for all subsequent analyses unless noted otherwise.

We compared caQTL results from RASQUAL to simple allelic imbalance (AI). 1,912 (81%) caQTL exhibited nominal (beta-binomial $p<0.05$) and 1,112 (47%) exhibited genome-wide AI (FDR<5%), all with the same direction of effect as the caQTL. Lead caQTL variants and representative AI variants exhibiting nominal AI showed strongly correlated effect sizes (Pearson's R=0.75, **Figure 3.2B**). AI effect sizes tended to be larger than caQTL effect sizes (**Figure 3.2B**), possibly because AI was calculated using individual variants whereas caQTL were identified using entire peaks.

To determine the extent of shared genetic effects across different markers of transcriptional regulatory elements, we compared the 3,123 caQTL to 921 H3K27ac QTL from a recent report[12]. Of the 921 H3K27ac QTL peaks, 85 (9%) are within 1 kb of a caPeak and have a lead variant in strong LD ($r^2>0.8$) with the caQTL lead. The largely distinct results may be due to the small sample sizes, analysis differences, and different genetic effects on the two epigenetic marks.

To predict the regulatory function of caPeaks, we compared caPeaks to liver tissue chromatin states from the Roadmap Epigenomics Consortium[3]. Relative to non-caPeaks (eigenMT-adjusted $p>0.5$), caPeaks were more frequently located in enhancers (48.6% vs. 33.0%) and promoters (11.7% vs. 9.3%) (**Figure 3.2C**). caQTL variants were significantly enriched in enhancers (OR=2.9), promoters (OR=2.0), and transcribed regions (OR=1.8) and depleted in polycomb (OR=0.5) and heterochromatin (OR=0.6) states, which are associated with gene repression and presumably inaccessible chromatin (**Figure 3.2D**). Taken together, caQTL showed strong overlap with active transcriptional regulatory elements, with particularly strong enrichment in enhancers.

Disruption of transcription factor binding motifs by caQTL

One way genetic variants may alter chromatin accessibility is by disrupting TF binding sites[5,6,8]. Among 4,585 variants within a caPeak and in strong LD with the caQTL lead, 3,132 (68%) variants altered the binding affinity of a TF motif (**Figure 3.3A**). Of the 2,793 variant-containing caPeaks, 2,249 (81%) contained at least one variant predicted to disrupt a motif, and 602 of these contained 2 or more predicted motif-disrupting variants. Motifs for many TFs were disrupted by multiple caQTL variants, and 109 TF motifs were disrupted by 20 or more variants. Disruption of motifs for 29 of these 109 TFs was significantly associated with caQTL status (log OR>0, $p<4.6\text{x}10^{-4}$) (**Figure 3.3B**), including TFs from the HNF, FOXA, and CEBP families[62], CTCF, and ATF2. FOXA and CEBP factors can act as pioneer factors by binding to inaccessible chromatin and initiating the establishment of accessible chromatin[65] and ATF2 can alter chromatin structure to activate or repress transcription[66], suggesting that this approach identifies TFs that may influence chromatin accessibility.

To investigate how often TFs bind the more accessible allele, we compared alleles associated with higher chromatin accessibility to the motifs. Among 7,629 motifs for all TFs, the more accessible allele matched the motif better for 4,770 motifs (63%, binomial $p<4.1\text{x}10^{-107}$). Similarly, among 3,132 motifs corresponding to the highest expressed TF among motifs at each variant, the more accessible allele matched the motif better for 1,953 motifs (62%, binomial $p<8.0\text{x}10^{-44}$). When restricting analysis to 993 observations of the 29 TFs for which motif disruption is associated with caQTL status, the more accessible allele matched the motif better for 834 motifs (84%, binomial $p<5.1\text{x}10^{-111}$). TFs exhibited variation in the percent of motifs that matched better to the more accessible allele (**Figure 3.3C**). For 11 TFs, including HNF4A, ATF4, ERF, and FOXA2, over 90% of stronger motif matches corresponded to the more accessible allele, while for SPI1 only 56% of stronger motif matches corresponded to the more accessible allele. These results suggest that TFs typically, but not always, bind to the more accessible allele.

Identifying putative target genes for caPeaks

Connecting caPeaks to their target genes is challenging, particularly when the caPeaks are distal to transcription start sites (TSS's). Individual approaches for identifying target genes have limitations and may not always show a direct regulatory relationship between a caPeak and gene. To address these challenges, we used four approaches to connect caPeaks to genes (**Figure 3.4A**).

First, we identified caPeaks proximal (-2 kb/+1 kb) to TSSs of genes expressed in liver. Of 3,123 total caPeaks, 114 (4%) were proximal to the TSS of at least 1 gene. Among these 114 caPeaks, 15 were proximal to the TSS of two or three genes (**Figure 3.4A**). This approach identified 131 unique caPeak-gene connections.

Second, we used liver tissue promoter capture Hi-C[2] to identify caPeaks that physically interact with gene promoters. We identified 329 distal caPeaks (>15 kb from any promoter as defined in the Hi-C analysis) that interact with promoters for 451 genes, including a caPeak that interacts with the promoter of *SNX10* (**Figure 3.4B**). Among caPeaks that overlapped the promoter of one gene and interact with the promoter of another gene, we identified an additional 104 caPeaks that interact with promoters of 190 genes. Combining promoter-distal and promoter-promoter interactions, we identified 697 caPeak-gene connections (**Figure 3.4A**).

Third, we identified caPeak sizes that either correlated with expression level of nearby genes or with the size of ATAC peaks at promoters. More caPeaks were correlated with promoter ATAC peaks than with gene expression level; 120 caPeaks were significantly correlated (FDR<5%) with promoter ATAC peaks while only 2 caPeaks were correlated with gene expression (FDR<5%), resulting in 121 unique caPeaks because gene RP11-101E14.2 had both types of correlations (**Figure 3.4A**). When using the same p-value threshold for both analyses ($p<2.9x10^{-4}$), 5 additional caPeaks were correlated with gene expression. As an example at a regulatory element previously shown to regulate *SORT1*[67], caPeak9372 is positively correlated with a peak proximal to a *SORT1* TSS (peak9400, Spearman rho=0.76, $p<1.6x10^{-4}$; **Figures 3.4C-4D**) and nominally correlated with *SORT1* expression (Spearman rho=0.69, $p<1.2x10^{-3}$). The vast majority of peak-peak correlations (167 of 173, 97%) are positive, suggesting that higher caPeak accessibility is usually associated with higher accessibility of connected promoter peaks. Using either caPeak-promoter peak or caPeak-gene correlations, we identified 196 caPeak-gene connections (**Figure 3.4A**).

Finally, we identified caQTL for which the lead variant exhibited high LD ($r^2>0.8$) with an eQTL lead variant for 15,418 autosomal genes from a liver tissue eQTL meta-analysis of 1,183 individuals[11]. Of 3,119 unique caQTL lead variants, 414 (13%) were in strong LD with at least 1 eQTL lead variant, which is similar to the percentage reported in a previous caQTL study[6]. Among caQTL lead variants, 71 were in strong LD with more than one eQTL lead variant, suggesting some caPeaks may affect expression of multiple genes. In total, we identified 463 target genes for 415 caPeaks, representing 506 unique caPeak-gene connections (**Figure 3.4A**). For example, we identified a caQTL signal with the same variants as an eQTL signal for *SORT1* (**Figures 3.4E-4F**). At connected loci, the allele associated with higher chromatin accessibility was usually associated with higher gene expression (390 of 506 loci, 77%; **Figure 3.4G**), suggesting caPeaks frequently act as promoters or enhancers to gene expression. We obtained a similar result when restricting to caQTL variants associated with only one peak and colocalized with eQTL variants associated with only one gene (273 of 337 loci, 81%).

Together the four methods identified a total of 1,461 caPeak-gene connections, although the approaches showed low overlap. Only 69 caPeak-gene connections were predicted by two methods, and no connections by three methods, likely due to the low power of many of the approaches (**Figure 3.4H**). These 69 caPeak-gene associations consist of 67 unique caPeaks and 67 unique genes; two caPeaks had two target genes. The methods that showed the most overlap were eQTL and TSS proximity (32 connections) and eQTL and HiC chromatin contacts (22 connections) (**Figure 3.4H**). This integrated approach predicted a target gene for 861 of 3,123 caPeaks (28%), suggesting caPeaks frequently interact with genes.

Prediction of regulatory mechanisms at GWAS loci

To identify genetic variants that may influence disease by altering chromatin accessibility, we identified caQTL and GWAS signals that may be shared, based on strong LD (r^2>0.8) between lead caQTL and lead GWAS variants. Using GWAS variants for 19 traits relevant to liver function and cardiometabolic traits from the NHGRI-EBI GWAS catalog[58], we identified 110 potentially shared caQTL and GWAS signals, corresponding to 111 caPeaks, because one caQTL signal was associated with two caPeaks. We identified at least one colocalized caQTL for 15 of the 19 traits, and liver enzymes showed the highest percentage of potentially shared caQTL and GWAS signals (15 signals, 19%) (**Table 3.1**). For traits with at least 5 GWAS-caQTL shared signals, we identified a high percentage of shared signals (>5%) for C-reactive protein, total cholesterol, and LDL cholesterol, consistent with the involvement of liver in inflammation and lipid metabolism[10].

To identify plausible regulatory mechanisms at GWAS loci, we integrated our GWAS-colocalized caQTL with TF motif-disrupting variants and predicted caPeak target genes. Of the 111 caPeaks at potentially shared caQTL-GWAS signals, 85 harbored a TF motif-disrupting variant, 56 had a predicted target gene, and 45 had both types of data. The gene with a TSS closest to the GWAS lead variant was predicted to be a target gene for 25 of 56 caPeaks (45%).

We identified seven GWAS-caQTL shared signals with strong evidence of regulatory mechanisms. At these GWAS loci, the caPeak had a target gene identified by two approaches and harbored TF motif-disrupting variants (**Table 3.2**). We identified shared caQTL, eQTL, and GWAS signals and a correlated caPeak-promoter peak pair (**Table 2; Figures 3.4C-4F**) at the *SORT1* locus associated with LDL cholesterol for which the alternate allele (rs12740374-T) has been shown to create a CEBP binding site and increase hepatic *SORT1* expression[67]. At a less well characterized locus, the caQTL signal with lead variant rs13395911 associated with caPeak119621 is colocalized with GWAS signals for plasma liver enzyme levels in European[68] and Asian[69] individuals and an eQTL for *EFHD1*[11] (**Figures 3.5A-5C**). Increased accessibility corresponds to higher *EFHD1* expression level and higher liver enzyme levels. caPeak119621 physically interacts with the promoter of *EFHD1* in liver tissue promoter capture Hi-C data[2] (**Figure 3.5D**), further suggesting that caPeak119621 may affect *EFHD1* expression. The peak overlaps ChIP-seq peaks for 12 TFs in liver (**Figure 3.5E**), and rs13395911 disrupts motifs for eight TFs expressed in liver. The motif with the largest difference between rs13395911 alleles is for *FOXA2*, and the allele with higher chromatin accessibility matches the motif better (**Figure 3.5F**). These and other connections provide potential regulatory mechanisms linking variants to regulatory element, transcription factors and genes that may influence the GWAS traits.

Identification of a putative functional variant at the LITAF locus

Near the *LITAF* gene, which encodes lipopolysaccharide (LPS) induced TNF factor, we identified a caQTL signal for caPeak75869 and tested variants for allelic differences in transcriptional activity and protein binding. This caQTL signal is potentially shared with a GWAS signal for LDL cholesterol[70] and an eQTL signal for *LITAF*[11] (**Figures 3.6A-6B**). caPeak75869 loops to the promoter of *LITAF* in liver tissue promoter capture Hi-C[2] (**Figure 3.6C**). caPeak75869 contains the lead caQTL variant rs57792815 (caQTL $p<5.0x10^{-17}$) and two additional variants in strong LD with the caQTL lead, rs3784924 ($r^2=0.95$) and rs11644920 ($r^2=0.98$). The haplotype associated with higher accessibility consists of the rs57792815-T, rs3784924-A, and rs11644920-A alleles. We tested a 666-bp DNA construct spanning the three variants for haplotype differences in transcriptional activity using luciferase reporter assays, testing the construct in two orientations relative to a minimal promoter. Given that *LITAF* is involved in lipopolysaccharide (LPS)-stimulated immune response[71], we tested transcriptional activity in four cell types: HepG2 hepatocytes, THP-1 monocytes, THP-1 differentiated macrophages, and LPS-stimulated THP-1 macrophages. In all four cell types, the forward orientation construct containing the alleles associated with higher accessibility showed significantly higher transcriptional activity than the construct containing the other alleles, with the strongest differences observed in hepatocytes (fold change=2.49, $p=2x10^{-4}$) and LPS-stimulated macrophages (fold change=1.39, p=7x10^{-4}; **Figure 3.6D**). The same haplotype showed significantly higher transcriptional activity in the reverse orientation for hepatocytes (p=1x10^{-4}) and unstimulated macrophages (p=0.02) and a trend toward higher transcriptional activity in the other cell types. We next tested each of the three haplotype variants for allelic differences in protein binding using nuclear extract from HepG2 cells. Only rs11644920 showed allele-specific binding, with the T allele showing increased binding (**Figure 3.6E**). Although caPeak75869 contained motifs and liver ChIP-seq binding sites for numerous TFs (**Figure 3.6F**), we were unable to identify the protein that showed allelic differences in binding to rs11644920 through supershift assays. Together, these results suggest that altered transcription factor binding at rs11644920 and increased chromatin accessibility of the regulatory element marked by caPeak75869 may lead to increased transcriptional activity and higher *LITAF* expression.

Discussion

We profiled chromatin accessibility in 20 individuals and identified caQTL in human liver tissue. caQTL variants frequently disrupt TF binding motifs, and alleles that better match a motif more often have higher chromatin accessibility, consistent with TFs stabilizing chromatin in an accessible state. We identified 1,461 putative caPeak-gene links using four approaches, suggesting that caPeaks frequently regulate gene expression. We identified 110 caQTL at GWAS signals, including 56 with a predicted caPeak target gene, identifying regulatory mechanisms that may be responsible for trait variation. Among variants at a caQTL, eQTL, and LDL cholesterol GWAS signal near the *LITAF* gene, one variant showed allelic differences in transcriptional activity and *in vitro* TF binding. This study contributes to the epigenomic characterization of human liver tissue and will aid in functional characterization of GWAS loci that act in liver.

Combining caQTL, caPeak-gene links, and disrupted TF motifs helps identify mechanisms at GWAS loci. At the well-characterized *SORT1* GWAS locus for lipid and cardiovascular traits[67], we showed that the previously described functional variant rs12740374 is associated with chromatin accessibility and that the caPeak containing this variant is correlated with a peak at the *SORT1* promoter. We also identified plausible regulatory mechanisms at less well-characterized loci. At a GWAS signal for BMI[72] and LDL cholesterol[70], we identified a caQTL potentially shared with a *PRMT6* eQTL signal and observed that the caPeak overlapped the *PRMT6* TSS. *PRMT6* has been shown to regulate hepatic glucose metabolism in mice[73]. Our data suggest that a variant at this locus may increase chromatin accessibility and alter TF binding at the *PRMT6* TSS, leading to higher *PRMT6* expression and decreased LDL cholesterol. At a GWAS locus for plasma liver enzyme levels[6869], we predicted *EFHD1* as a target gene based on both caQTL-eQTL colocalization and a promoter capture Hi-C link. While *EFHD1* is expressed in liver tissue, the GTEx portal shows that expression is much higher in other tissues[1], and the gene's roles in liver have not been characterized[74]. Our data suggest that *EFHD1* may be a target gene at this locus and act through one of the many cell types in liver tissue. These and other results highlight the utility of caQTL to identify mechanisms at GWAS loci.

At the *LITAF* locus, we provided direct evidence that variant rs11644920 can alter transcriptional regulation. Here, the caQTL, liver eQTL, and LDL cholesterol GWAS signals are shared, and the variant, mechanism and cell type responsible for these associations were unknown. *LITAF* encodes a transcription factor that can mediate effects on inflammation[71], suggesting a potential role in hepatocytes and/or macrophages in an inflammatory environment. We showed that variants in the caPeak alter transcriptional reporter activity in hepatocytes, monocytes, macrophages and lipopolysaccharide-stimulated macrophages. In all cell types, the caPeak showed a similar magnitude of enhancer activity and alleles showed differences in transcriptional activity, suggesting that the variant may act in any or all of these cell types. We further provided evidence that rs11644920 alters protein binding, at least *in vitro*, although we failed to experimentally validate the specific TFs that discriminate between alleles. Further study is needed to provide direct evidence that these variants alter transcription of the *LITAF* gene and how altered levels of *LITAF* may affect cholesterol levels.

The maximum distance threshold between peaks and tested variants had a substantial impact on caQTL detection. Analyzing variants within a narrow region around a peak reduced the multiple testing burden for nearby variants, whereas testing variants in a broader region allowed identification of variants within one peak that may also influence another peak. A wide range of distance thresholds have been applied to caQTL discovery, including variants within 1 kb and 20 kb of peak centers[6], 50 kb from peak ends[4], and 1 Mb from peak ends[8]. We found many more significant results when using variants within 1 kb of peak centers compared to variants within 100 kb of peak centers, potentially due to reduced multiple testing burden and low power to detect long-range caQTL effects due to small sample size.

We used four approaches to provide suggestive evidence that a caPeak may regulate a specific gene. TSS proximity is useful to detect variation in promoter accessibility, although our results showed only 4% of caPeaks are TSS-proximal. Promoter capture Hi-C data[2] identifies distal regions that physically interact with promoters, indicating potential regulatory relationships. However, Hi-C data mapped in only one sample may miss chromatin contacts that differ by genotype or environmental exposure. The identification of caPeaks correlated with promoter peaks is based on the concept that co-regulated regions should show similar chromatin accessibility patterns. The same concept motivates correlating caPeak accessibility with gene expression, although gene expression is affected by many other factors, which may be why we identified more correlated caPeak-promoter peak pairs than caPeak-gene pairs, and both correlation approaches are limited by sample size. The LD-based method we used to predict shared caQTL and eQTL signals helps identify peaks and genes with a shared genetic basis, although this method is influenced by limited fine-mapping of the lead caQTL variant, use of an LD threshold, and choice of LD reference panel, and would be more comprehensive if we could analyze conditionally distinct eQTL signals[1]. While each of these approaches was useful to predict links between caPeaks and genes, additional experiments are needed to identify causal relationships.

The caQTL presented here are a resource for studying liver regulatory elements and will help identify mechanisms at GWAS loci for multiple traits that act through liver. The 56 caQTL at GWAS loci with predicted target genes are strong candidates for future functional studies. While caQTL can pinpoint functional regulatory variants, the modest sample size and analyses restricted to common variants limit fine-mapping potential and highlight the importance of considering LD proxies. The promising regulatory mechanisms identified here motivate identification of liver caQTL in larger sample sizes.

Acknowledgments

I thank everyone who contributed to the work in this chapter: Michael Erdos and Jacqueline Idol for generating ATAC-seq and RNA-seq libraries; Narisu Narisu, Tingfen Yan, Vivek Rai, Peter Orchard, Ricardo Albanus, Hannah Perrin, and Alaine Broadaway for helping with data analysis; Swarooparani Vadlamudi for performing functional assays; Amy Etheridge, Federico Innocenti, Erin Schuetz, John Didion, and Amarjit Chaudhry for study design and sample procurement; Amy Etheridge and Federico Innocenti for providing eQTL data; Lori Bonnycastle, Peter Orchard, Hannah Perrin, and Karen Mohlke for figure generation; and Laura Scott, Stephen Parker, Francis Collins, and Karen Mohlke for advising analyses.

This study was supported by NIH grants R01DK072193, U01DK105561, R01DK093757, and 1-ZIA-HG000024. Individuals were supported by NIH grants F31HL146121 (K.W.C.), T32GM067553 (K.W.C.), T32HL069768 (H.J.P.), T32HL129982 (K.A.B.), and R25GM055336 (K.W.C). Liver samples were obtained from the Liver Tissue Cell Distribution System supported by NIH contract N01DK70004/HHSN267200700004C.

We thank Dan Liang, Jason Stein and Terry Furey for helpful discussions on computational analyses and Gautam Pandey for literature reviews and biological interpretation of loci.

We also thank the Roadmap Epigenomics Project for liver tissue chromatin states, the ENCODE Consortium for liver transcription factor ChIP-seq data, Cis-BP for TF binding motifs, and GTEX for liver tissue gene expression data. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx Portal.

Web Resources

Regions of unusually high linkage disequilibrium:

https://genome.sph.umich.edu/wiki/Regions_of_high_linkage_disequilibrium_(LD)

Novoalign: http://www.novocraft.com/products/novoalign)

BWA: https://github.com/lh3/bwa

CTA: https://github.com/ParkerLab/cta

Picard: https://github.com/broadinstitute/picard

Gplots R package: https://rdrr.io/cran/gplots/

dbSNP build 151 common variants:

ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606_b151_GRCh37p13/VCF/00-common_all.vcf.gz

swiss: https://github.com/statgen/swiss

Promoter capture Hi-C data download (liver code is LI11): http://kobic.kr/3div/download

Promoter capture Hi-C promoter baits: https://junglab.wixsite.com/home/db-link

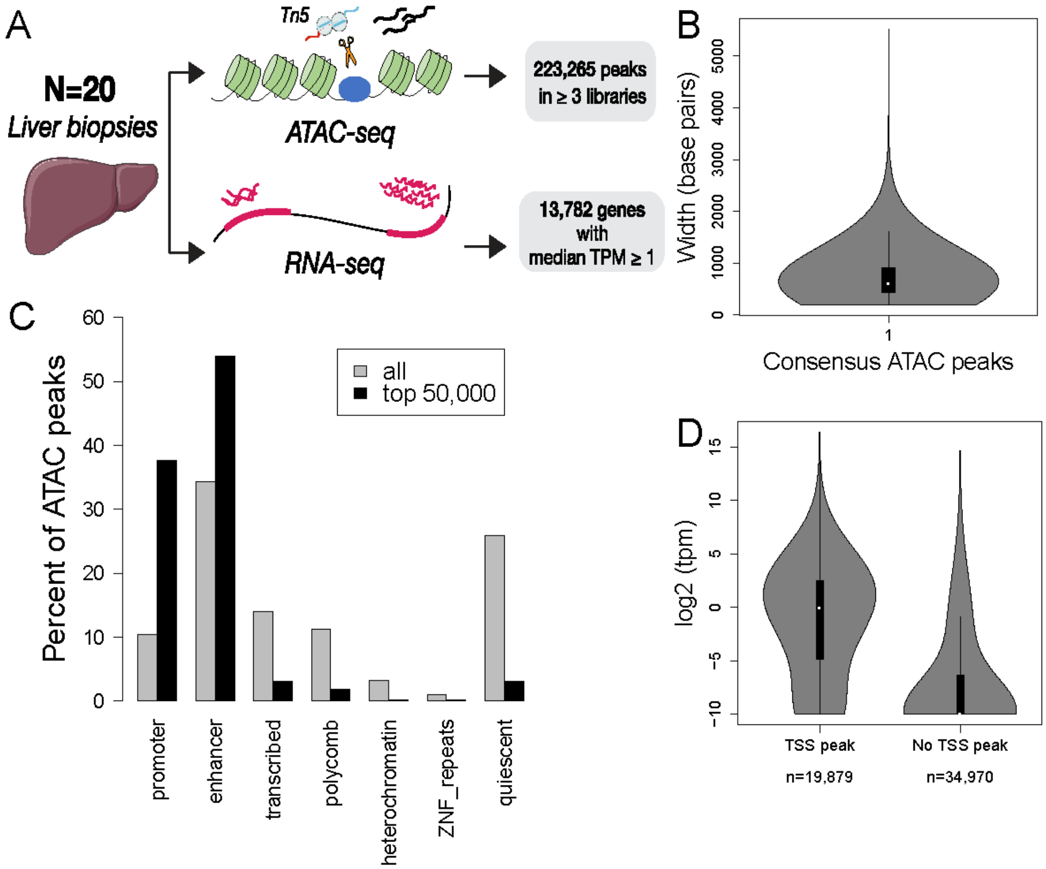
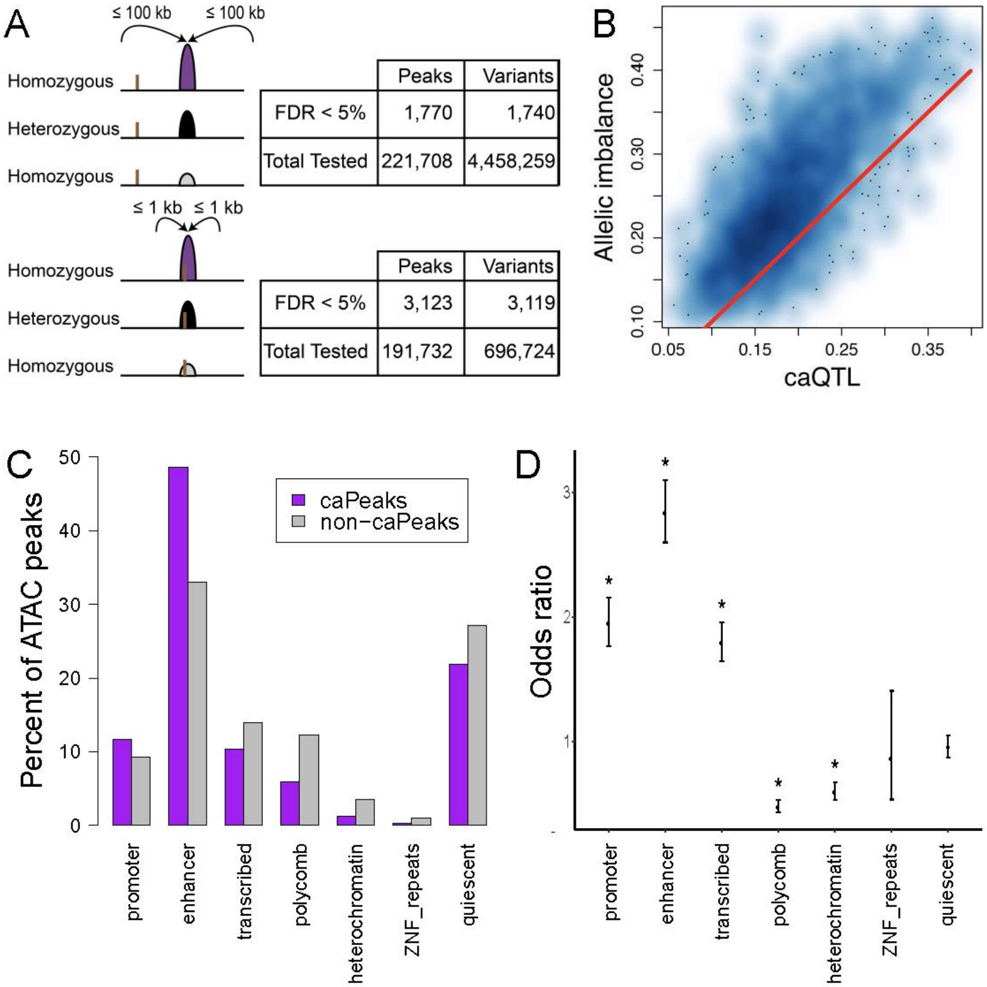


Figure 3.1. Joint profiling of gene expression and chromatin accessibility in human liver tissue. (A) RNA-seq and ATAC-seq was performed in liver samples from 20 donors. (B) Distribution of consensus ATAC peak widths in base pairs. (C) Percent of consensus ATAC peaks by chromatin state in liver tissue from the Roadmap Epigenomics Project. All peaks, gray; 50,000 most accessible consensus peaks, black; quiescent represents unannotated regions. (D) Comparison of the distribution of expression between genes with and without an ATAC peak overlapping the transcription start site (TSS).



	Peaks	Variants
FDR < 5%	1,770	1,740
Total Tested	221,708	4,458,259

	Peaks	Variants
FDR < 5%	3,123	3,119
Total Tested	191,732	696,724

Figure 3.2. Identification and characterization of caQTL. (A) caQTL identified using variants within 100 kb or 1 kb of peak centers. (B) Comparison of effect sizes between caQTL and simple allelic imbalance (Pearson's R=0.75). (C) Comparison of the number of caPeaks and non-caPeaks assigned to each chromatin state in liver tissue from the Roadmap Epigenomics Project. caPeaks, purple; non-caPeaks, gray; quiescent represents unannotated regions. (D) Enrichment of caQTL variants in liver chromatin states. Error bars represent 95% confidence intervals. * indicates significant enrichment (p<0.0071).

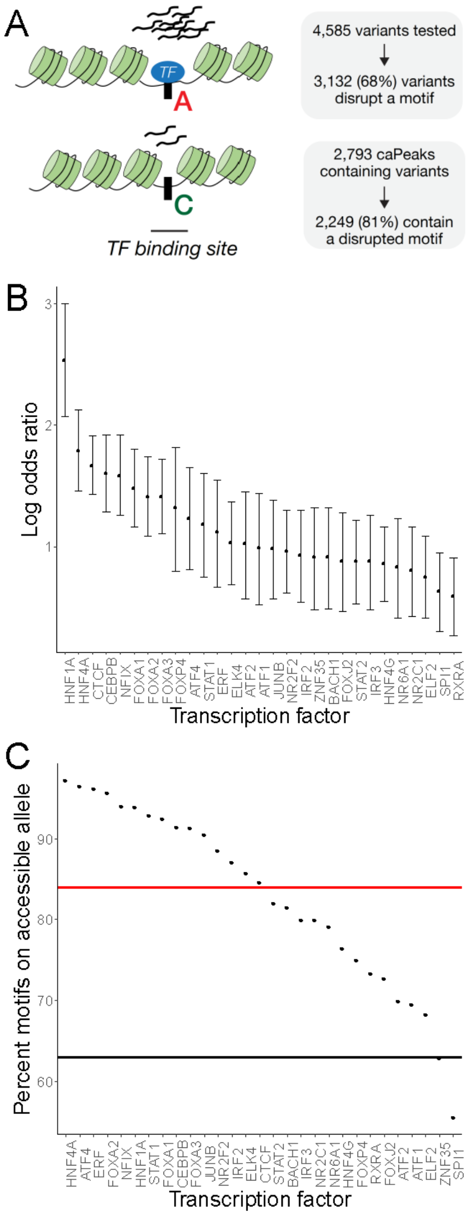
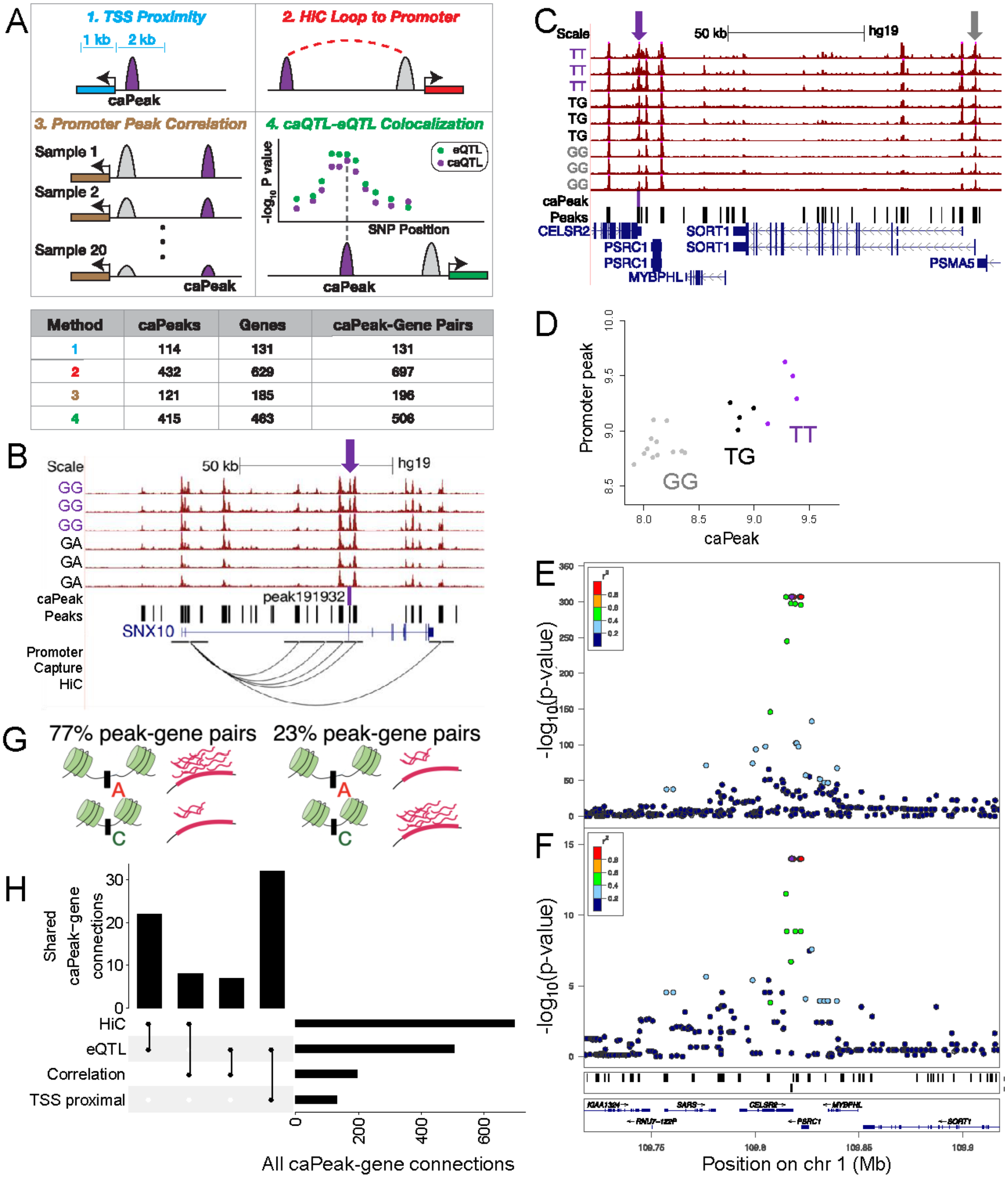


Figure 3.3. Disruption of TF binding motifs by caQTL variants. (A) Allele affinities for TF binding and chromatin accessibility for variants within caPeaks and in strong LD with the caQTL lead variant (r^2>0.8). (B) Association of caQTL status with motif disruption status. Only TFs with at least 20 motifs disrupted by caQTL variants were included, and only significant associations ($p<4.6x10^{-4}$) are shown. (C) Percent of disrupted motifs for which the allele with higher chromatin accessibility matched the motif better. Percents are shown for the 29 TFs that had at least 20 motifs disrupted by caQTL variants. Black line, percent for all disrupted motifs across all tested TFs; red line, average percent across the 29 TFs.



Method	caPeaks	Genes	caPeak-Gene Pairs
1	114	131	131
2	432	629	697
3	121	185	196
4	415	463	508

Figure 3.4. Prediction of target genes for caPeaks using four approaches. (A) Illustrations of four approaches to predict caPeak target genes. (B) Hi-C chromatin contact shown as an arc between caPeak191932 and the *SNX10* promoter. Selected ATAC-seq signal tracks are shown for each caQTL genotype of rs12534816. More accessible homozygotes, purple; heterozygotes, black. (C) Genome browser image showing the correlation across rs12740374 genotypes of caPeak9372 and a peak at a *SORT1* promoter. (D) The same peak correlation with points representing normalized peak counts of individual samples colored by rs12740374 genotype. (E) *SORT1* eQTL associations at the signal shared with the caQTL for caPeak9372 and (F) caQTL associations with caPeak9372. In both plots, the caQTL lead variant within 1 kb of the peak center is indicated by a purple diamond and LD is based on 1000G phase 3 Europeans. (G) Comparison of directions of effect among all shared caQTL and eQTL signals. The A allele represents the more accessible allele than C, and more red marks indicate higher gene expression. (H) UpSet plot comparing the number of shared and unique caPeak-gene links identified by the four approaches.

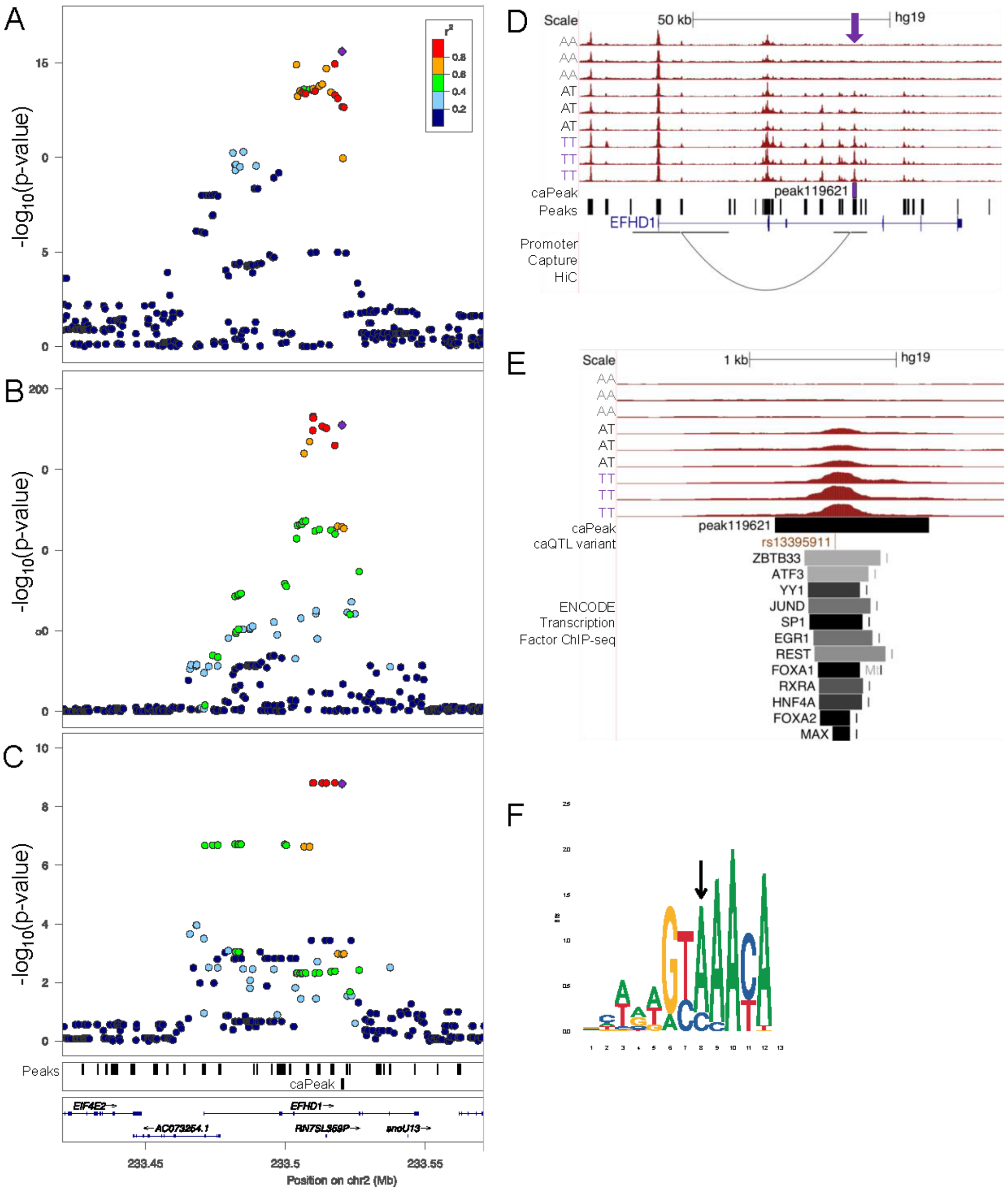

A
15
0
5
0
-log$_{10}$(p-value)
r^2
0.8
0.8
0.4
0.2
B
200
0
0
o0
0
-log$_{10}$(p-value)
C
10
8
6
4
2
0
-log$_{10}$(p-value)
Peaks
caPeak
EIF4E2
EFHD1
AC073254.1
RN7SL359P
snoU13
233.45
233.5
233.55
Position on chr2 (Mb)
D
Scale
50 kb
hg19
AA
AA
AA
AT
AT
AT
TT
TT
TT
caPeak
peak119621
Peaks
EFHD1
Promoter
Capture
HiC
E
Scale
1 kb
hg19
AA
AA
AA
AT
AT
AT
TT
TT
TT
caPeak
peak119621
caQTL variant
rs13395911
ENCODE
Transcription
Factor ChIP-seq
ZBTB33
ATF3
YY1
JUND
SP1
EGR1
REST
FOXA1
RXRA
HNF4A
FOXA2
MAX
F
2.5
2.0
1.5
1.0
0.5
0.0
Bits
1
2
3
4
5
6
7
8
9
10
11
12
13

Figure 3.5. A plausible regulatory mechanism at the *EFHD1* locus for plasma liver enzyme levels. (A) Variant association with plasma levels of the liver enzyme alanine transaminase in Japanese individuals, (B) eQTL association for *EFHD1*, and (C) caQTL associations for caPeak119621. For all three plots, the caQTL lead variant within 1 kb of the peak center is indicated by a purple diamond and LD is based on 1000G phase 3 East Asians (A) or Europeans (B and C). (D) Hi-C chromatin contact shown as an arc between caPeak119621 and the *EFHD1* promoter. Selected ATAC-seq signal tracks are shown for each rs13395911 genotype. More accessible homozygotes, purple; heterozygotes, black; less accessible homozygote, gray. (E) Transcription factor ChIP-seq peaks in liver tissue from ENCODE that overlap caPeak119621. (F) Sequence logo plot for the *FOXA2* motif s disrupted by caQTL variant rs13395911 (arrow). The motif match is shown on the negative strand, and variant alleles in D and E are shown on the positive strand.

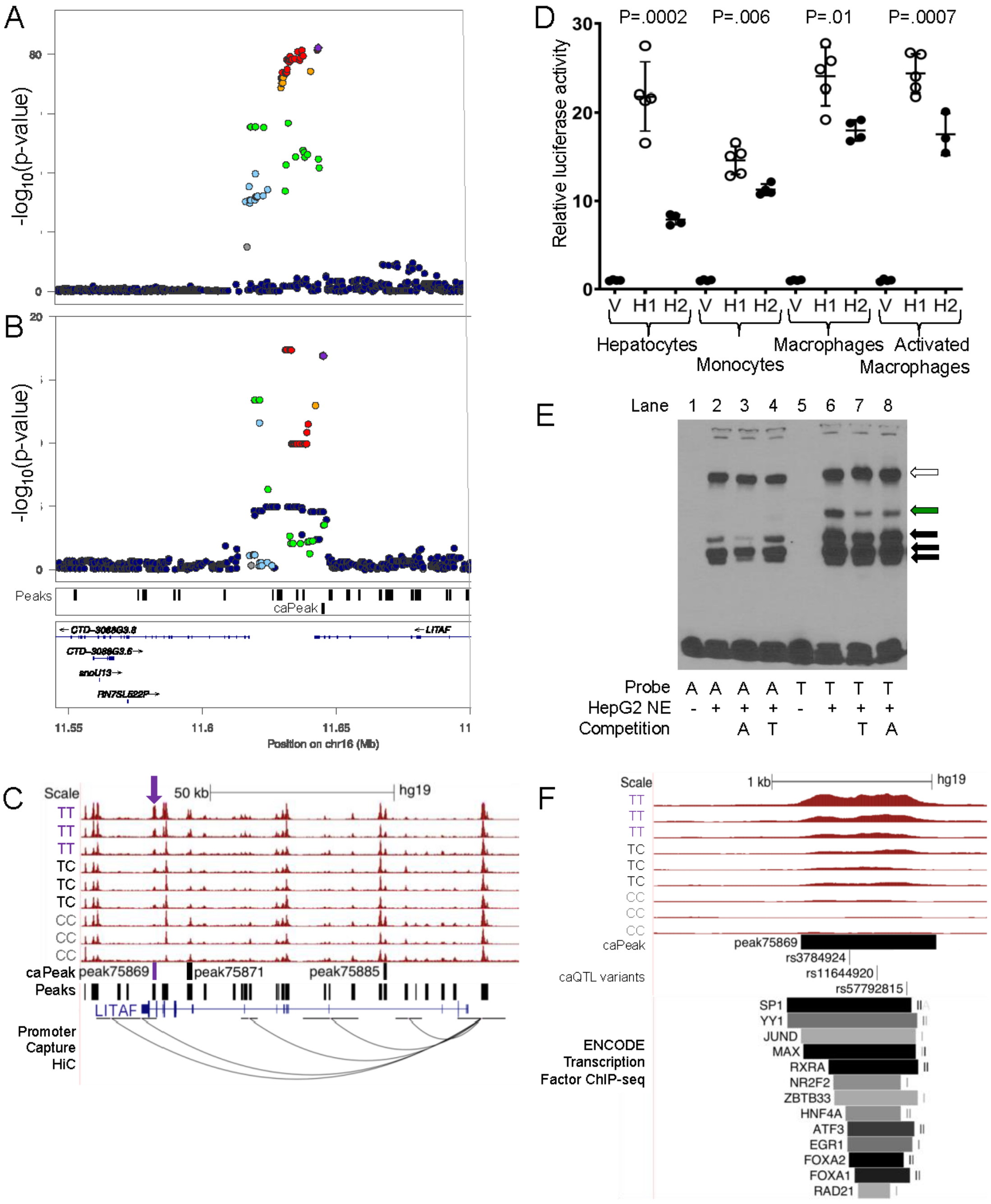

A
80
0
-log10(p-value)
B
20
0
-log10(p-value)
Peaks
caPeak
← CTD-3088G3.8
← LITAF
CTD-3088G3.6 →
snoU13 →
RN7SL522P →
11.55
11.6
11.65
11
Position on chr16 (Mb)
C
Scale
50 kb
hg19
TT
TT
TT
TC
TC
TC
CC
CC
CC
caPeak
peak75869
peak75871
peak75885
Peaks
LITAF
Promoter Capture HiC
D
P=.0002
P=.006
P=.01
P=.0007
30
20
10
0
Relative luciferase activity
V
H1
H2
Hepatocytes
Monocytes
Macrophages
Activated Macrophages
E
Lane
1
2
3
4
5
6
7
8
Probe
A A A A T T T T
HepG2 NE
- + + + - + + +
Competition
A T T A
F
Scale
1 kb
hg19
TT
TT
TT
TC
TC
TC
CC
CC
CC
caPeak
peak75869
caQTL variants
rs3784924
rs11644920
rs57792815
ENCODE Transcription Factor ChIP-seq
SP1
YY1
JUND
MAX
RXRA
NR2F2
ZBTB33
HNF4A
ATF3
EGR1
FOXA2
FOXA1
RAD21

Figure 3.6. Identification of a putative functional variant at the *LITAF* locus for LDL cholesterol. (A) eQTL association for *LITAF* and (B) caQTL associations for caPeak75869 at an LDL cholesterol GWAS signal. In both plots, the caQTL lead variant within 1 kb of the peak center is indicated by a purple diamond, and LD is based on 1000G phase 3 Europeans. (C) Hi-C chromatin contact between caPeak75869 and the *LITAF* promoter. Selected ATAC signal tracks are shown for each rs57792815 genotype. More accessible homozygotes, purple; heterozygotes, black; less accessible homozygotes, gray. (D) Transcriptional activity of a 666-bp DNA element spanning caPeak75869 and containing rs3784924, rs11644920, and rs57792815 in HepG2 hepatocytes, THP-1 monocytes, THP-1 differentiated macrophages, and LPS-stimulated THP-1 macrophages. The DNA element was tested in the forward orientation relative to the genome. V, empty vector; H1, haplotype 1 of more accessible alleles; H2, haplotype 2 of less accessible alleles. (E) EMSA using HepG2 nuclear extract shows allelic differences in protein binding for rs11644920. Green arrow, band represents T-allele-specific binding; black arrows, T-allele-preferential binding; white arrow, non-specific binding. Competition probes were unlabeled and in 25-fold excess. (F) TF ChIP-seq peaks in liver tissue from ENCODE that overlap caPeak75869.

Trait	Number of GWAS signals [a]	Number of shared caQTL-GWAS signals [b]	Percent of shared caQTL-GWAS signals [c]
Liver enzymes	77	15	19.5
C-reactive protein	81	5	6.2
Total cholesterol	292	18	6.2
LDL cholesterol	240	14	5.8
Glucose	54	3	5.6
Insulin	18	1	5.6
Bilirubin	20	1	5.0
HDL cholesterol	314	14	4.5
Triglycerides	279	11	3.9
Cardiovascular disease	454	13	2.9
Body mass index	986	24	2.4
Blood pressure	1,540	37	2.4
WHRadjBMI	209	4	1.9
Type 2 diabetes	268	5	1.9
HbA1c	66	1	1.5
Glycated albumin	2	0	0.0
Liver injury	17	0	0.0
NAFLD	9	0	0.0
Serum albumin	15	0	0.0

Table 3.1. Shared GWAS-caQTL signals by trait. [a]Counted as lead GWAS variants not in high LD (r^2<0.8) with another. [b]Shared if the caQTL lead variant was in strong LD (r^2>0.8) with the GWAS lead. [c]Percent of all GWAS signals that are shared with a caQTL. LDL, low-density lipoprotein; HDL, high-density lipoprotein; WHRadjBMI, waist-hip ratio adjusted for BMI; NAFLD, non-alcoholic fatty liver disease.

caQTL variant	caPeak	GWAS variant	GWAS trait	LD r^2 [a]	Gene	Methods [b]	caQTL, eQTL directions [c]
rs12740374	peak9372	rs12740374	LDL cholesterol	1.00	*SORT1*	eQTL, Corr	D, D
rs17276527	peak13768	rs4077194	HDL cholesterol	1.00	*RALGPS2*	eQTL, HiC	D, D
rs13395911	peak119621	rs13395911	ALT	1.00	*EFHD1*	eQTL, HiC	I, I
rs2232015	peak9185	rs1730859	LDL cholesterol	0.97	*PRMT6*	TSS, eQTL	D, D
rs2037517	peak71475	rs832890	Pulse pressure	0.90	*PLEKHO2*	eQTL, HiC	D, D
rs12677006	peak205272	rs1906672	Sys. blood pressure	0.89	*DDHD2*	eQTL, HiC	I, I
rs57792815	peak75869	rs34318965	LDL cholesterol	0.81	*LITAF*	eQTL, HiC	I, I

Table 3.2. Selected caQTL at GWAS loci. Loci are shown for shared caQTL-GWAS signals if the caPeak was linked to a target gene by two methods and if the caPeak harbored motif-disrupting variants. LD r^2 between the caQTL and GWAS lead variants. [b] Methods that linked the caPeak to a gene. Corr, correlation between caPeak and promoter peak accessibility. [c] Direction of chromatin accessibility and gene expression relative to the allele associated with an increase in the GWAS trait, where “I” indicates increased and “D” indicates decreased accessibility or expression. ALT, alanine aminotransferase levels.

REFERENCES

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

2. Jung, I., Schmitt, A., Diao, Y., Lee, A.J., Liu, T., Yang, D., Tan, C., Eom, J., Chan, M., Chee, S., et al. (2019). A compendium of promoter-centered long-range chromatin interactions in the human genome. Nat. Genet. *51*, 1442–1449.

3. Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317.

4. Alasoo, K., Rodrigues, J., Mukhopadhyay, S., Knights, A.J., Mann, A.L., Kundu, K., Hale, and C., Dougan, G., and Gaffney, D.J. (2018). Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. Nat Genet.

5. Kumasaka, N., Knights, A.J., and Gaffney, D.J. (2016). Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. Nat Genet *48*, 206–213.

6. Degner, J.F., Pai, A.A., Pique-Regi, R., Veyrieras, J.-B., Gaffney, D.J., Pickrell, J.K., Leon, S.D., Michelini, K., Lewellen, N., Crawford, G.E., et al. (2012). DNase?I sensitivity QTLs are a major determinant of human expression variation. Nature *482*, 390–394.

7. Bryois, J., Garrett, M.E., Song, L., Safi, A., Giusti-Rodriguez, P., Johnson, G.D., Shieh, A.W., Buil, A., Fullard, J.F., Roussos, P., et al. (2018). Evaluation of chromatin accessibility in prefrontal cortex of individuals with schizophrenia. Nat. Commun. *9*, 3121.

8. Gate, R.E., Cheng, C.S., Aiden, A.P., Siba, A., Tabaka, M., Lituiev, D., Machol, I., Gordon, M.G., Subramaniam, M., Shamim, M., et al. (2018). Genetic determinants of co-accessible chromatin regions in activated T cells across humans. Nat. Genet. *50*, 1140–1150.

9. Khetan, S., Kursawe, R., Youn, A., Lawlor, N., Jillette, A., Marquez, E.J., Ucar, D., and Stitzel, M.L. (2018). Type 2 Diabetes-Associated Genetic Variants Regulate Chromatin Accessibility in Human Islets. Diabetes *67*, 2466–2477.

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

11. Etheridge, A.S., Gallins, P.J., Jima, D., Broadaway, K.A., Ratain, M.J., Schuetz, E., Schadt, E., Schroder, A., Molony, C., Zhou, Y., et al. (2020). 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.

12. Çalışkan, M., Manduchi, E., Rao, H.S., Segert, J.A., Beltrame, M.H., Trizzino, M., Park, Y., Baker, S.W., Chesi, A., Johnson, M.E., et al. (2019). Genetic and Epigenetic Fine Mapping of Complex Trait Associated Loci in the Human Liver. Am. J. Hum. Genet. *105*, 89–107.

13. Strunz, T., Grassmann, F., Gayán, J., Nahkuri, S., Souza-Costa, D., Maugeais, C., Fauser, S., Nogoceke, E., and Weber, B.H.F. (2018). 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.

14. Willer, C.J., Schmidt, E.M., Sengupta, S., Peloso, G.M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M.L., Mora, S., et al. (2013). Discovery and refinement of loci associated with lipid levels. Nat. Genet. *45*, 1274–1283.

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

16. Klemm, S.L., Shipony, Z., and Greenleaf, W.J. (2019). Chromatin accessibility and the regulatory epigenome. Nat. Rev. Genet. *20*, 207–220.

17. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

18. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867–2873.

19. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinforma. Oxf. Engl. *27*, 2156–2158.

20. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

21. Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods. Nat. Genet. *48*, 1284–1287.

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

23. McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet *48*, 1279–1283.

24. Varshney, A., Scott, L.J., Welch, R.P., Erdos, M.R., Chines, P.S., Narisu, N., Albanus, R.D., Orchard, P., Wolford, B.N., Kursawe, R., et al. (2017). Genetic regulatory signatures underlying islet gene expression and type 2 diabetes. Proc Natl Acad Sci USA *114*, 2301–2306.

25. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinforma. Oxf. Engl. *30*, 2114–2120.

26. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinforma. Oxf. Engl. *29*, 15–21.

27. Jun, G., Flickinger, M., Hetrick, K.N., Romm, J.M., Doheny, K.F., Abecasis, G.R., Boehnke, M., and Kang, H.M. (2012). Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am J Hum Genet *91*, 839–848.

28. Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. *47*, D766–D773.

29. Hartley, S.W., and Mullikin, J.C. (2015). QoRTs: a comprehensive toolset for quality control and data processing of RNA-Seq experiments. BMC Bioinformatics *16*, 224.

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

31. Scott, L.J., Erdos, M.R., Huyghe, J.R., Welch, R.P., Beck, A.T., Wolford, B.N., Chines, P.S., Didion, J.P., Narisu, N., Stringham, H.M., et al. (2016). The genetic regulatory signature of type 2 diabetes in human skeletal muscle. Nat Commun *7*, 11764.

32. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods *10*, 1213–1218.

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

34. Rai, V., Quang, D.X., Erdos, M.R., Cusanovich, D.A., Daza, R.M., Narisu, N., Zou, L.S., Didion, J.P., Guan, Y., Shendure, J., et al. (2020). 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.

35. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and and, R.D. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

36. Orchard, P., Kyono, Y., Hensley, J., Kitzman, J.O., and Parker, S.C.J. (2020). Quantification, Dynamic Visualization, and Validation of Bias in ATAC-Seq Data with ataqv. Cell Syst. *10*, 298-306.e4.

37. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol *9*, R137.

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

39. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinforma. Oxf. Engl. *30*, 923–930.

40. R Core Team (2015). R: A Language and Environment for Statistical Computing (Vienna, Austria: R Foundation for Statistical Computing).

41. Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J. (2004). The UCSC Table Browser data retrieval tool. Nucleic Acids Res *32*, D493-6.

42. Weirauch, M.T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P., Najafabadi, H.S., Lambert, S.A., Mann, I., Cook, K., et al. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. Cell *158*, 1431–1443.

43. Castro-Mondragon, J.A., Jaeger, S., Thieffry, D., Thomas-Chollier, M., and van Helden, J. (2017). RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections. Nucleic Acids Res. *45*, e119.

44. McLeay, R.C., and Bailey, T.L. (2010). Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. BMC Bioinformatics *11*, 165.

45. Ramaker, R.C., Savic, D., Hardigan, A.A., Newberry, K., Cooper, G.M., Myers, R.M., and Cooper, S.J. (2017). A genome-wide interactome of DNA-associated proteins in the human liver. Genome Res. *27*, 1950–1960.

46. Jou, J., Gabdank, I., Luo, Y., Lin, K., Sud, P., Myers, Z., Hilton, J.A., Kagda, M.S., Lam, B., O'Neill, E., et al. (2019). The ENCODE Portal as an Epigenomics Resource. Curr. Protoc. Bioinforma. *68*, e89.

47. Huang, Q.Q., Ritchie, S.C., Brozynska, M., and Inouye, M. (2018). Power, false discovery rate and Winner's Curse in eQTL studies. Nucleic Acids Res. *46*, e133.

48. Davis, J.R., Fresard, L., Knowles, D.A., Pala, M., Bustamante, C.D., Battle, A., and Montgomery, S.B. (2016). An Efficient Multiple-Testing Adjustment for eQTL Studies that Accounts for Linkage Disequilibrium between Variants. Am. J. Hum. Genet. *98*, 216–224.

49. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Methodol. *57*, 289–300.

50. Geijn, B. van de, McVicker, G., Gilad, Y., and Pritchard, J.K. (2015). WASP: allele-specific software for robust molecular quantitative trait locus discovery. Nat Methods *12*, 1061–1063.

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

52. Yee, T.W. (2015). Vector generalized linear and additive models: with an implementation in R (Springer).

53. Iotchkova, V., Ritchie, G.R.S., Geihs, M., Morganella, S., Min, J.L., Walter, K., Timpson, N.J., UK10K Consortium, Dunham, I., Birney, E., et al. (2019). GARFIELD classifies disease-relevant genomic features through integration of functional annotations with association signals. Nat. Genet. *51*, 343–353.

54. Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics *27*, 1017–1018.

55. Mitchelmore, J., Grinberg, N.F., Wallace, C., and Spivakov, M. (2020). Functional effects of variation in transcription factor binding highlight long-range gene regulation by epromoters. Nucleic Acids Res. *48*, 2866–2879.

56. de la Torre-Ubieta, L., Stein, J.L., Won, H., Opland, C.K., Liang, D., Lu, D., and Geschwind, D.H. (2018). The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis. Cell *172*, 289-304.e18.

57. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47.

58. Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. *47*, D1005–D1012.

59. Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F., et al. (2006). The UCSC Genome Browser Database: update 2006. Nucleic Acids Res. *34*, D590-598.

60. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308–311.

61. Fogarty, M.P., Cannon, M.E., Vadlamudi, S., Gaulton, K.J., and Mohlke, K.L. (2014). Identification of a regulatory variant that binds FOXA1 and FOXA2 at the CDC123/CAMK1D type 2 diabetes GWAS locus. PLoS Genet. *10*, e1004633.

62. Nagaki, M., and Moriwaki, H. (2008). Transcription factor HNF and hepatocyte differentiation. Hepatol. Res. Off. J. Jpn. Soc. Hepatol. *38*, 961–969.

63. Kim, S., Yu, N.-K., and Kaang, B.-K. (2015). CTCF as a multifunctional protein in genome regulation and gene expression. Exp. Mol. Med. *47*, e166.

64. Oishi, Y., and Manabe, I. (2018). Krüppel-Like Factors in Metabolic Homeostasis and Cardiometabolic Disease. Front. Cardiovasc. Med. *5*, 69.

65. Mayran, A., and Drouin, J. (2018). Pioneer transcription factors shape the epigenetic landscape. J. Biol. Chem. *293*, 13795–13804.

66. Lau, E., and Ronai, Z.A. (2012). ATF2 – at the crossroad of nuclear and cytosolic functions. J. Cell Sci. *125*, 2815–2824.

67. Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N.E., Ahfeldt, T., Sachs, K.V., Li, X., Li, H., Kuperwasser, N., Ruda, V.M., et al. (2010). From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature *466*, 714–719.

68. Chambers, J.C., Zhang, W., Sehmi, J., Li, X., Wass, M.N., Van der Harst, P., Holm, H., Sanna, S., Kavousi, M., Baumeister, S.E., et al. (2011). Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma. Nat. Genet. *43*, 1131–1138.

69. Kanai, M., Akiyama, M., Takahashi, A., Matoba, N., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K., et al. (2018). Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. Nat. Genet. *50*, 390–400.

70. Klarin, D., Damrauer, S.M., Cho, K., Sun, Y.V., Teslovich, T.M., Honerlaw, J., Gagnon, D.R., DuVall, S.L., Li, J., Peloso, G.M., et al. (2018). Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. Nat. Genet. *50*, 1514–1523.

71. Myokai, F., Takashiba, S., Lebo, R., and Amar, S. (1999). A novel lipopolysaccharide-induced transcription factor regulating tumor necrosis factor alpha gene expression: molecular cloning, sequencing, characterization, and chromosomal assignment. Proc. Natl. Acad. Sci. U. S. A. *96*, 4518–4523.

72. Kichaev, G., Bhatia, G., Loh, P.-R., Gazal, S., Burch, K., Freund, M.K., Schoech, A., Pasaniuc, B., and Price, A.L. (2019). Leveraging Polygenic Functional Enrichment to Improve GWAS Power. Am. J. Hum. Genet. *104*, 65–75.

73. Han, H.-S., Jung, C.-Y., Yoon, Y.-S., Choi, S., Choi, D., Kang, G., Park, K.-G., Kim, S.-T., and Koo, S.-H. (2014). Arginine methylation of CRTC2 is critical in the transcriptional control of hepatic glucose metabolism. Sci. Signal. *7*, ra19.

74. Dütting, S., Brachs, S., and Mielenz, D. (2011). Fraternal twins: Swiprosin-1/EFhd2 and Swiprosin-2/EFhd1, two homologous EF-hand containing calcium binding adaptor proteins with distinct functions. Cell Commun. Signal. CCS *9*, 2.

CHAPTER 4: CHROMATIN ACCESSIBILITY DIFFERENCES DURING ADIPOGENESIS IDENTIFY CONTEXT-DEPENDENT REGULATORY VARIANTS

Introduction

Identifying functional variants, molecular mechanisms, and relevant tissue/cell types for genome-wide association study (GWAS) loci remains challenging, especially at signals without coding variants. Regulatory variants and mechanisms have been detected based on variant location in transcriptional regulatory elements of trait-relevant tissues[1–3], including variants located in regulatory elements present only in certain cellular contexts[4]. GWAS loci are colocalized with gene expression quantitative trait loci (eQTL) in trait-relevant tissues[5–10]. Some GWAS loci are colocalized with context-dependent eQTL, such as those in stimulated, but not naïve, immune cells[11]. Mapping transcriptional regulatory elements and gene expression in additional contexts may help characterize molecular mechanisms of additional GWAS loci.

Adipose tissue influences insulin sensitivity, blood cholesterol levels, inflammation, and related cardiometabolic traits through its roles in lipid storage and hormone secretion[12,13]. Hundreds of GWAS loci for cardiometabolic traits are shared, or colocalized, with gene expression quantitative trait loci (eQTL) in adipose tissue[5–7]. At a subset of colocalized GWAS-eQTL signals, statistical analyses suggest that adipose tissue gene expression mediates the effect of the genetic variant on the GWAS trait[7]. Variants at cardiometabolic GWAS loci are also overrepresented in transcriptional regulatory elements in adipose tissue[1–3]. However, adipose tissue contains multiple cell types, including adipocytes and their precursors (preadipocytes)[14]. Mapping genetic effects on gene regulation in preadipocytes and adipocytes at various stages of differentiation may uncover additional roles for GWAS variants.

Multiple approaches exist to study gene regulation in specific cell types within adipose tissue. Gene expression has been profiled using microarrays in adipocyte, preadipocyte, and immune cells isolated from whole adipose tissue using flow cytometry sorting[15]. Chromatin accessibility differences have been identified between primary preadipocytes and in vitro differentiated adipocytes[16]. The Simpson Golabi-Behmel Syndrome (SGBS) human preadipocyte cell strain is a well-characterized adipocyte cell model[17] that has been used to study differences in gene expression, chromatin accessibility, histone modifications, and transcription factor (TF) binding during adipocyte differentiation and other differences in cell environment[18,19].

In this study, we identified differences in chromatin accessibility and gene expression between adipocyte differentiation states in SGBS cells. We identified variants at GWAS loci that resided in regulatory elements more accessible in preadipocytes or adipocytes and predicted genes likely regulated by these elements. Finally, we identified variants at the *SCD* and *EYA2* loci that showed context-specific and/or allelic effects in transcriptional reporter activity.

Methods

Cell Culture

SGBS cells[20] were generously provided by Dr. Martin Wabitsch (University of Ulm) and cultured as previously described[21]. Briefly, we cultured SGBS preadipocytes in serum-containing medium until confluent, then rinsed in phosphate-buffered-saline (PBS) and differentiated for four days in basal medium (DMEM:F12 + 3.3mM biotin + 1.7mM panthotenate) 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.

ATAC-seq library preparation

We profiled chromatin accessibility in SGBS cells at days 0, 2, 4, and 14 of adipocyte differentiation following the Omni-ATAC protocol[22,23] 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 5uL of Tn5 per library, although some early libraries were prepared with 2.5uL of Tn5 (**Table 4.2**). We cleaned the transposase reaction and final library with Zymo DNA Clean and Concentrator. We visualized and quantified libraries using a TapeStation. Paired-end sequencing was performed using 150-bp reads on an Illumina Novaseq at Novogene (Beijing, China) or with 50-bp reads on an Illumina HiSeq 4000 at the University of North Carolina sequencing core (Chapel Hill, North Carolina; **Table 4.2**).

ATAC-seq read alignment and peak calling

We trimmed sequencing adapters and low quality base calls from the 3' ends of reads using cutadapt[24] with parameters -q 20 –minimum-length 36. We aligned trimmed reads to the hg19 human genome[25] using bowtie2[26] with parameters –minins 36 –maxins 1000 –no-mixed –no-discordant –no-unal and selected nuclear chromosomal alignments with mapq>20 using samtools[27]. We removed alignments overlapping blacklisted regions[28] using BEDTools pairToBed[29] 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[30]. 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)[22] and smoothed signal by extending alignments 100 bp on either side of the Tn5 binding sites using BEDTools slop[29]. We called peaks (FDR<5%) with MACS2[31] with parameters -q 0.05 –nomodel –bdg and generated ATAC signal bigwig files from MACS2 bedGraph files using the bedGraphToBigWig tool from ucsctools[32].

For each analyzed day of SGBS differentiation, we generated a set of representative ATAC peaks using the following method. First, we merged peak genomic coordinates across replicates for a given day using BEDTools merge[29]. Second, we defined representative peaks as merged peaks that overlapped individual replicate peaks in greater than 50% of replicates (at least 3 out of 5 replicates for day 14 and 6 out of 10 replicates for days 0 and 4).

Identification of differentially accessible peaks

We generated a set of consensus peaks to test for differential chromatin accessibility by merging the top 100,000 representative peaks in each day (ranked by median MACS2 p-value across replicates). We quantified the accessibility of these consensus peaks in each library using featureCounts[33]. We computed the GC percent of each peak using BEDTools nuc[29] and generated within-library GC bias normalization factors using full quantile normalization with EDASeq[34] and used DESeq2[35] size factors to control for differences in sequencing depth between libraries. Adjusting peak counts for batch effects (defined as library preparation date) did not improve clustering of replicates within each differentiation day, so we did not adjust for batch in the differential chromatin accessibility analysis. We tested for differential chromatin accessibility using DESeq2[35] 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[36] with the -size 200 option. We used peaks that were not differential in any pairwise day comparison (FDR>50%, absolute value of $\log_2$ fold change < 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/)[37] with the GO Biological Process ontology[38,39]. We ran GREAT 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.

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

We isolated total RNA from SGBS cells at days 0, 2, 4, and 14 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 base pairs (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[24] with parameters -q 20 –minimum-length 36. We aligned reads to the hg19 human genome[25] using STAR[40] 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[41] and corrected for GC bias using salmon[42] with parameters –seqBias –gcBias –gencode. We generated RNA-seq quality metrics using the CollectRnaSeqMetrics tool from Picard (https://github.com/broadinstitute/picard).

To identify differentially expressed genes, we imported salmon transcript quantifications and collapsed to the gene level using tximport (https://github.com/mikelove/tximport). We retained 18,299 genes with median DESeq2-normalized count >= 1 across all libraries. Adjusting gene counts for batch effects (defined as RNA extraction date) improved the clustering of samples by differentiation day, so we included batch as a covariate in the differential gene expression analysis. We tested for differential gene expression using DESeq2[35] and classified genes with FDR<5% and log fold change (LFC)>1 as significantly differential.

Gene ontology enrichment of differential genes

We tested if differentially expressed genes were enriched for specific biological processes using the PANTHER statistical overrepresentation test[43] with the GO-Slim Biological Process ontology[38,39]. We ran PANTHER using the fisher 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.

GWAS loci colocalized with adipose tissue eQTL and overlap with differential peaks

We obtained a published set of 231 cardiometabolic trait GWAS loci colocalized with adipose tissue eQTL (eQTL n=434)[7]. A GWAS locus was considered colocalized with an eQTL if the GWAS and eQTL lead variants were in strong LD (r^2>=0.8) and if the eQTL lead variant was no longer significantly associated with gene expression ($p>9.6x10^{-6}$) when conditioning on the GWAS lead[7]. We restricted to 228 loci where the GWAS lead variant was biallelic and present in the 1000 Genomes phase 3 reference panel. We identified LD proxies of the GWAS lead variants ($r^2>0.8$) using PLINK v1.9[44] and identified loci that had a proxy variant within a preadipocyte-dependent or adipocyte-dependent peak using BEDTools intersect[29].

Transcriptional activity reporter assays

SGBS preadipocytes and adipocytes were maintained and transcriptional reporter luciferase assays were performed as previously described[3,21]. Primers were designed to amplify the entire chromatin accessibility region containing variants of interest. Amplified regions containing the reference and alternate allele for variants of interest were inserted in pGL4.23 firefly luciferase reporter vectors (Promega) in a 'forward' and 'reverse' orientation (with respect to the genome) upstream of a minimal promoter and luciferase gene. Three to five independent clones were cotransfected with *Renilla* luciferase vector in triplicate using Lipofectamine 3000 (Life Technologies) into SGBS cells at differentiation day 0 (preadipocyte) and day 2 (adipocyte). Luciferase activity of experimental clones was normalized to Renilla luciferase 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 pGL4.23 vector. We used a Student's t-test to compare luciferase activity between alleles and between contexts.

Results

Identification of differential chromatin accessibility during adipocyte differentiation

We collected SGBS cells at days 0, 2, 4, and 14 of adipocyte differentiation and profiled chromatin accessibility using Omni-ATAC-seq[22,23] and gene expression using RNA-seq (**Table 4.1**). For ATAC-seq, we identified an average of 76 million high-quality nuclear alignments across all libraries (**Table 4.2**) and identified 147,587 consensus peaks. Using principal component analysis (PCA) of ATAC-seq read counts in consensus peaks, we identified that the ATAC-seq libraries clustered into three distinct groups: libraries from day 0, libraries from days 2 and 4, and libraries from day 14 (**Figure 4.1**). We identified 58,784 peaks that showed differential chromatin accessibility between any two days of differentiation (FDR<5%, absolute value $\log_2$ fold change (LFC)>1). We found that the chromatin accessibility profiles of days 2 and 4 were essentially identical (no differential peaks, **Figure 4.2**), so we removed day 2 from further analyses because it had fewer replicates than day 4. We found many more differential peaks between days 0 and 4 (52,653) than between days 4 and 14 (1,118, **Figure 4.2**). In addition, 86% of the differential peaks between days 0 and 14 were differential between days 0 and 4 with the same direction of effect (34,163 of 39,721). Given the similarity between days 4 and 14, and to reduce the number of chromatin profiles to compare, we defined 'adipocyte-dependent' peaks as the intersection of peaks more accessible on days 4 or 14 relative to day 0 (n=15,919) and 'preadipocyte-dependent' peaks as the intersection of peaks more accessible on day 0 relative to days 4 or 14 (n=18,244).

We tested if adipocyte and preadipocyte-dependent peaks were found near genes involved in cell state-relevant biological processes and contained binding motifs for cell state-relevant transcription factors (TFs). Adipocyte-dependent peaks were found near genes enriched in 24 biological processes, including insulin signaling, white fat cell differentiation, glucose metabolism, and fat metabolism (**Table 4.3**). Preadipocyte-dependent peaks were found near genes enriched in 3 biological processes, which were all involved in cell adhesion (**Table 4.4**). Motifs for TFs that promote adipogenesis, such as *PPAR* and CEBP factors[45], were enriched in adipocyte-dependent peaks but not preadipocyte-dependent peaks (**Table 4.5**). Motifs for TFs that inhibit adipogenesis, such as GATA factors and PU.1[45,46], were enriched in preadipocyte-dependent peaks but not adipocyte-dependent peaks (**Table 4.6**). Taken together, adipocyte and preadipocyte-dependent peaks are found near cell state-relevant genes and contain cell state-relevant TF motifs.

Identification of differential gene expression during adipocyte differentiation

We profiled gene expression across adipocyte differentiation (**Table 4.1**) and identified an average of 46 million transcriptomic alignments across all libraries (**Table 4.7**). Similar to ATAC-seq libraries, we found that RNA-seq libraries clustered into three distinct groups using PCA of gene counts (days 0, 2-4, and 14; **Figure 4.3**) and the profiles of days 2 and 4 were essentially identical (only 3 differential genes; **Figure 4.4**). We identified a similar number of differentially expressed genes (FDR<5%, absolute value LFC>1) between days 0 and 4 (n=2,171) and days 0 and 14 (n=2,107) (**Figure 4.4**). In contrast to differential chromatin accessibility, for which 86% of changes between days 0 and 14 were observed by day 4, only 1,282 of 2,107 (61%) of differential genes between days 0 and 14 were identified by day 4. These results suggest that gene expression changes continue after most accessible chromatin sites have stabilized. As with chromatin accessibility, we defined 'adipocyte-dependent' genes as the intersection of genes more expressed on days 4 or 14 relative to day 0 (n=734) and 'preadipocyte-dependent' genes as the intersection of genes more expressed on day 0 relative to days 4 or 14 (n=548).

We next determined whether adipocyte and preadipocyte-dependent gene sets contained genes with roles relevant to their respective cell states. We identified adipocyte-dependent genes including *PPARG* and *CEBPA*, which promote adipocyte differentiation[45], and *ADIPOQ* and *LEP*, which encode the adiponectin and leptin hormones[12]. Preadipocyte-dependent genes included *WNT2* and *GATA2*, which have been shown to be down-regulated during adipocyte differentiation[45,47]. Adipocyte-dependent genes were enriched for 38 biological process ontology terms, including terms for lipid metabolism, hormone response, and glucose homeostasis (**Table 4.8**). Preadipocyte-dependent genes were enriched for 163 biological process terms, many of which are involved in cell cycle and cell division (**Table 4.9**). These results indicate that we identified differentially expressed genes with functions relevant to the corresponding cell states.

Context-dependent peaks overlap variants at GWAS loci colocalized with adipose tissue eQTL

To link regulatory elements to adipose tissue gene expression and cardiometabolic traits, we tested if context-dependent peaks overlapped proxy variants at a published set of cardiometabolic GWAS loci that colocalized with adipose tissue eQTL signals[7]. Of 228 tested GWAS loci, 59 (26%) harbored a proxy variant within either an adipocyte-dependent or preadipocyte-dependent peak (**Table 4.10**). Of these 59 loci, 26 had a proxy in an adipocyte-dependent peak but not a preadipocyte-dependent peak, 23 had a proxy in a preadipocyte-dependent peak but not an adipocyte-dependent peak, and 10 had a proxy in both types of peaks. Among the cardiometabolic traits, body mass index (BMI) had the most loci harboring context-dependent peaks (15 loci had an adipocyte-dependent peak and 21 loci had a preadipocyte-dependent peak; **Table 4.11**), likely in part because BMI had the largest number of initial GWAS-colocalized eQTL. At 4 loci, we identified 5 eQTL genes that were also adipocyte-dependent genes: *SCD*, *FADS1*, *SYPL2*, *AL139819.1*, and *SLC22A3*. The context-dependent peaks at these 59 GWAS-colocalized eQTL may have context-dependent roles on adipose tissue gene regulation and are candidates for further functional analysis.

GWAS variants show context and allele-dependent effects in transcriptional reporter activity

We tested regulatory elements at two GWAS-colocalized eQTL loci for context-dependent effects on transcriptional reporter activity. At the first locus, a GWAS locus for metabolic traits and plasma phospholipid levels is colocalized with an eQTL for the adipocyte-dependent gene *SCD* (**Table 4.10**). *SCD* encodes the stearoyl-CoA desaturase enzyme, which is involved in fatty acid synthesis[48]. Variant rs603424 is the only proxy variant at this locus based on LD and is found within an adipocyte-dependent peak (peak19405; **Figure 4.5** left). We tested both alleles of a 592-bp region that encompassed the majority of the peak for differences in transcriptional activity between adipocytes and preadipocytes (**Figure 4.5** right). When combining across alleles, we observed higher transcriptional activity in adipocytes compared to preadipocytes in both the forward orientation (adipocyte fold change (FC) relative to empty vector=4.74, preadipocyte FC=0.22, difference between conditions p=1.36x10^{-8}) and reverse orientation (adipocyte FC=10.14, preadipocyte FC=0.84, p=6.83x10^{-5}). In adipocytes we observed significantly higher transcriptional activity for the G allele compared to the A allele in both the forward orientation (G allele FC=5.82, A allele FC=3.66, difference between alleles p=0.003) and the reverse orientation (G allele FC=15.89, A allele FC=4.38, p=0.0001; **Figure 4.5** right). The G allele is associated with higher *SCD* expression in the eQTL data. These results suggest that the regulatory element marked by peak19405 may have both context-dependent and allele-dependent effects on transcription.

At the second locus, a GWAS locus for type 2 diabetes (T2D) and triglycerides (TG) is colocalized with an eQTL for the *EYA2* gene (**Table 4.10**). *EYA2* encodes the eyes absent transcriptional coactivator and phosphatase 2 transcription factor, which is involved in numerous processes including muscle development, hypertrophy, and cancer[49–51]. A proxy variant at this locus (rs55966194) is found within an adipocyte peak in the first intron of *EYA2* (peak81750; **Figure 4.6** left), and none of the 12 other proxy variants at this locus overlap context-dependent peaks. We tested both alleles of a 419-bp region that encompassed the majority of the peak for differences in transcriptional activity between adipocytes and preadipocytes (**Figure 4.6** right). When combining across alleles, we observed significantly higher transcriptional activity in adipocytes compared to preadipocytes in both the forward orientation (adipocyte FC=38.23, preadipocyte FC=0.18, p=1.23x10^{-13}) and reverse orientation (adipocyte FC=6.68, preadipocyte FC=0.17, p=2.20x10^{-16}). We observed significantly higher transcriptional activity for the C allele compared to the G allele in the reverse orientation, but not the forward orientation, for both adipocytes (C allele FC=7.72, G allele FC=5.64, p=0.0001) and preadipocytes (C allele FC=0.21, G allele FC=0.13, p=0.002; **Figure 4.6** right). We observed much lower transcriptional activity in adipocytes in the reverse orientation compared to the forward orientation, so the allelic differences in the reverse orientation should be interpreted with caution. The C allele is associated with higher *EYA2* expression in the eQTL data. These results suggest that the regulatory element marked by peak81750 may impact transcription in a context-dependent, and possibly allele-dependent, manner. At both loci, we observed higher transcriptional activity in the context for which the peak had higher chromatin accessibility, suggesting that context-dependent peaks can identify regulatory elements with context-dependent function.

Discussion

Here we identified differences in chromatin accessibility and gene expression during adipocyte differentiation. Among 147,587 total accessible regions, we defined 15,919 adipocyte-dependent peaks and 18,244 preadipocyte-dependent peaks. We identified variants within these context-dependent accessible chromatin regions at 59 GWAS-adipose tissue eQTL loci, suggesting context-dependent roles for these variants on gene regulation and cardiometabolic traits. Variants at two of these loci showed context-dependent effects on transcriptional reporter activity, suggesting that context-dependent accessible chromatin regions can predict context-dependent regulatory effects. Our results are consistent with previous reports of context-dependent effects of GWAS variants on gene regulation[4,11]. The data presented here help uncover the regulatory elements altered by GWAS variants and in which cellular states these alterations occur.

Context-dependent chromatin accessibility helped prioritize GWAS variants with potential context-dependent effects on gene regulation. We identified a variant, rs603424, at a GWAS-colocalized eQTL for *SCD* that showed higher transcriptional reporter activity in adipocytes compared to preadipocytes. Variant rs603424 was a particularly strong candidate for context-dependent regulatory activity because it overlapped a peak more accessible in adipocytes, it was the only proxy variant at the GWAS locus, and it was an eQTL for the *SCD* gene, which was more expressed in adipocytes. The role of SCD in fatty acid synthesis[48] also suggests a role in adipocytes. However, other loci are more complicated. The T2D GWAS locus colocalized with an eQTL for *EYA2* contains 13 proxy variants, only one of which, rs55966194, overlapped a context-dependent accessible chromatin region. While the regulatory element containing rs55966194 showed higher transcriptional activity in adipocytes, there may be other candidate causal variants at this locus. The context-dependent transcriptional activity of these two regulatory elements are promising and motivate testing of the other 57 loci we predict to have context-dependent effects.

We were able to predict context-dependent effects on gene regulation by mapping chromatin accessibility in only one individual and one change in cellular context. Given that chromatin accessibility can vary by genotype[11,52], mapping chromatin accessibility in additional individuals across adipocyte differentiation may identify additional context-dependent regions targeted by GWAS variants. Chromatin accessibility in adipocytes also changes in response to inflammation[18], so mapping chromatin accessibility in additional contexts may also uncover new context-dependent genetic regulatory mechanisms. Identifying which GWAS variants are causal and in which cellular contexts they act will help guide therapeutic strategies.

Acknowledgements

I thank everyone who contributed to the work in this chapter: Hannah Perrin and Kenneth Ng for generating ATAC-seq and RNA-seq libraries; Hannah Perrin, Gautam Pandey, and Swarooparani Vadlamudi for performing functional assays; Hannah Perrin for helping with computational analyses; Martin Wabitsch for providing SGBS cells; Markku Laakso for his work on the METSIM study; and Karen Mohlke for advising analyses.

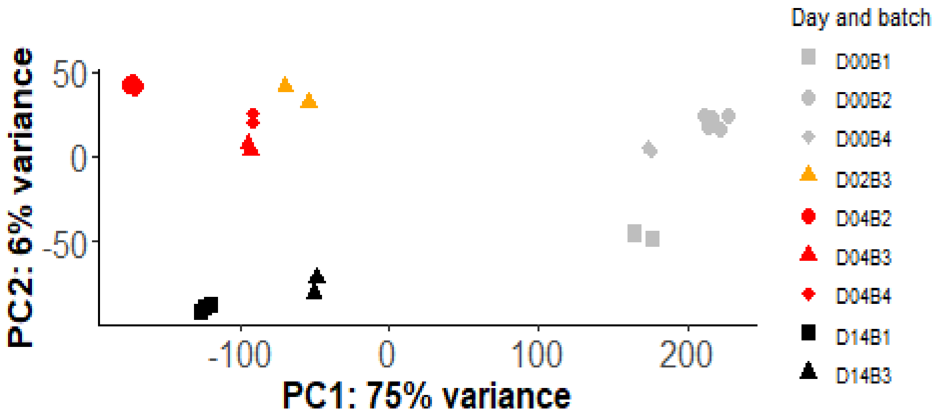
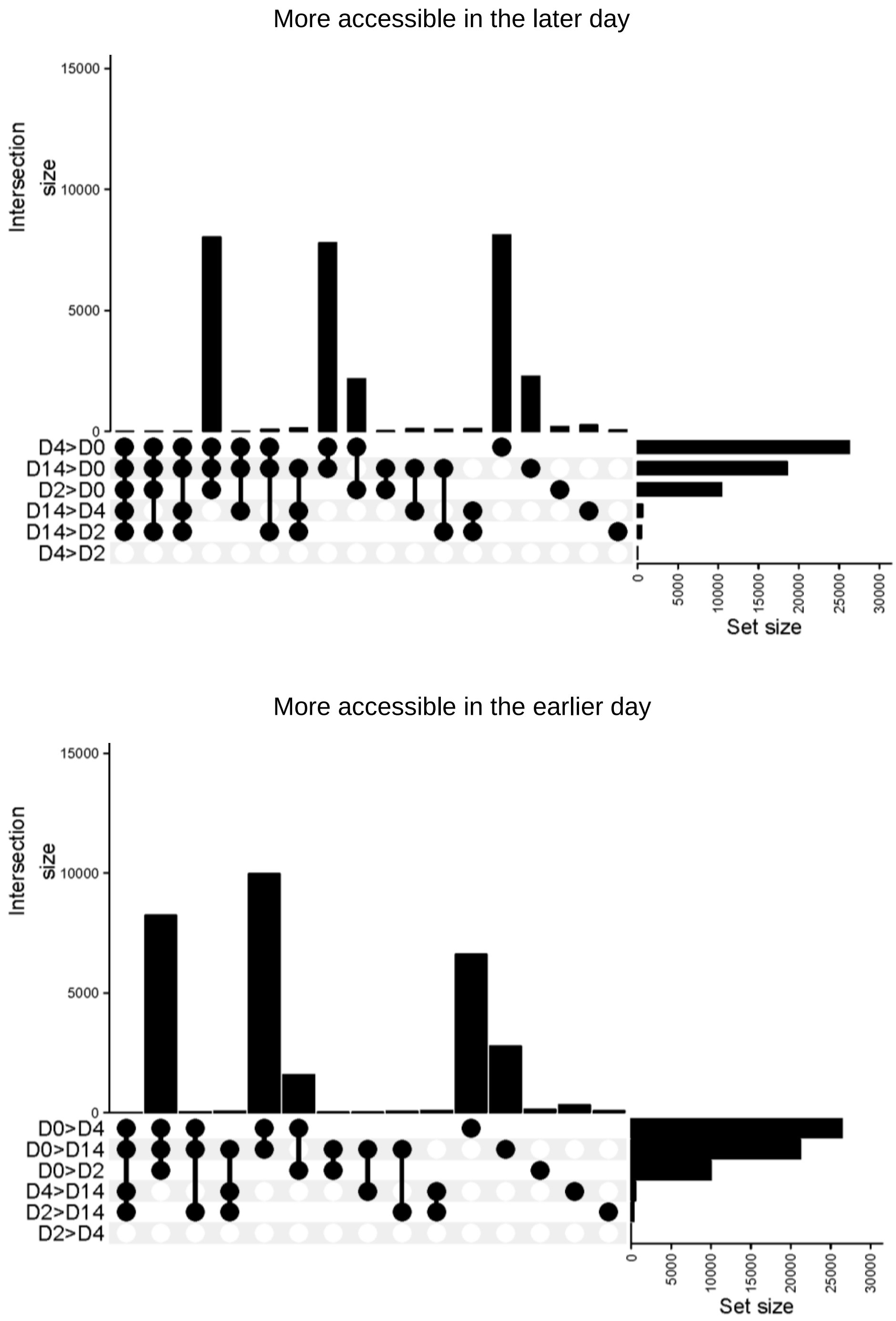


Figure 4.1. Comparison of ATAC-seq libraries using principal component analysis. Principal component analysis (PCA) of ATAC read counts in 147,587 consensus peaks. Peak counts were normalized by the DESeq2 rlog function. Color represents differentiation day and shape indicates ATAC library preparation batch.



More accessible in the later day

More accessible in the earlier day

Figure 4.2. Comparison of differentially accessible peaks between pairwise day comparisons. UpSet plots indicating the number of differentially accessible peaks (FDR<5%, absolute value LFC>1) shared between each pairwise day comparison. The top plot shows peaks more accessible in the later day of a pairwise comparison (day 4 > day 0 for example) and the bottom plot shows peaks more accessible in the earlier day (day 0 > day 4 for example). The horizontal bars on the right side of the plots show the number of total differential peaks in each

pairwise comparison and the vertical bars on the top show the number of differential peaks shared between comparisons.

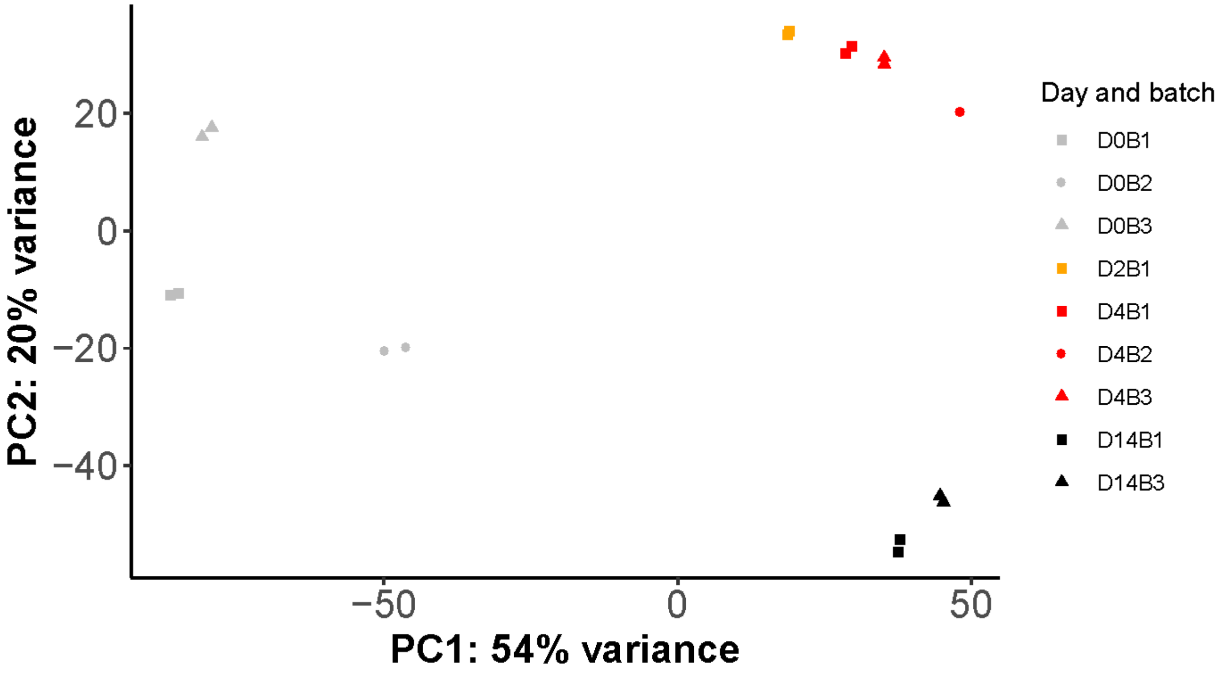
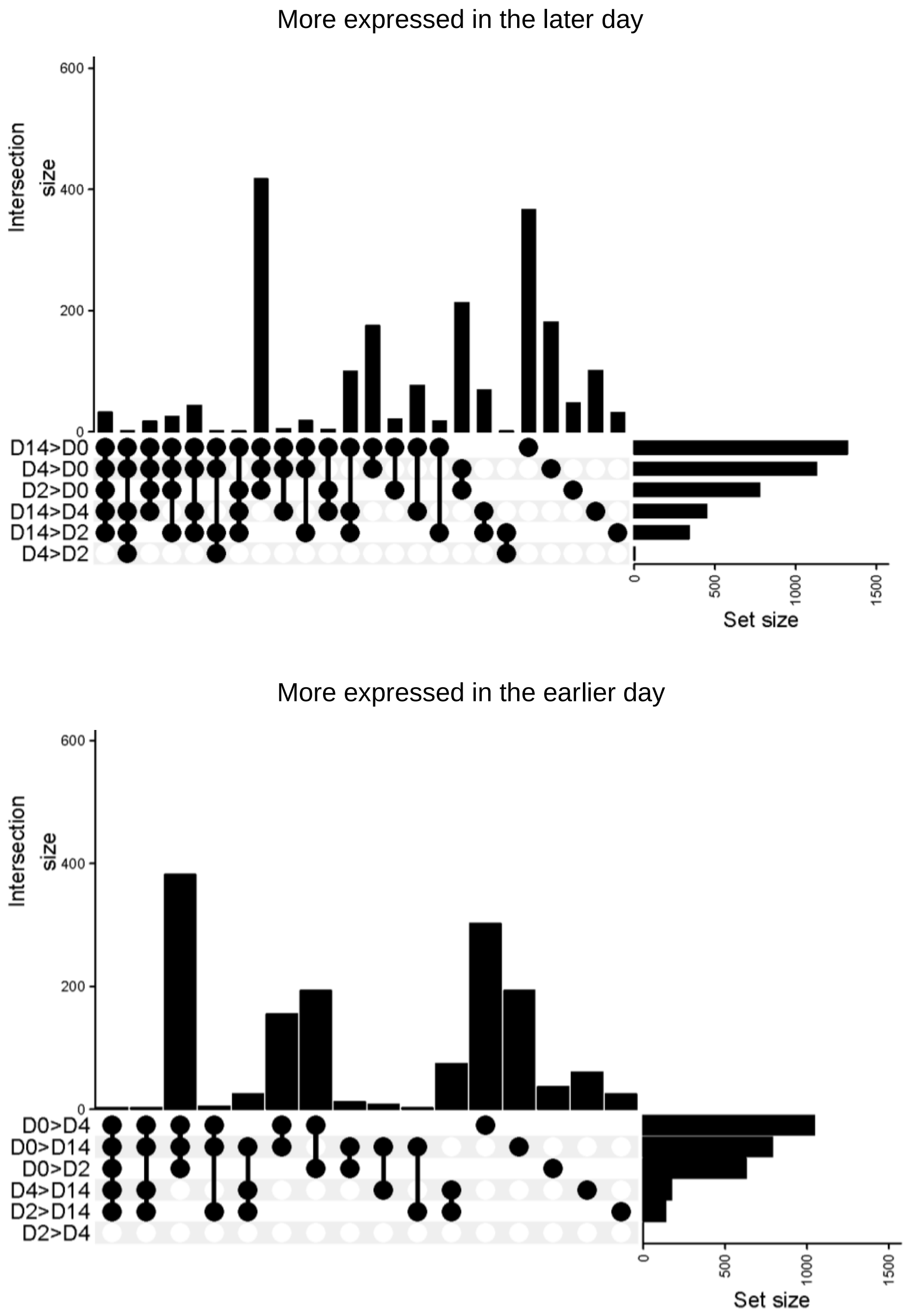


Figure 4.3. Comparison of RNA-seq libraries using principal component analysis. Principal component analysis (PCA) of RNA read counts in 18,299 genes with median DESeq2-normalized count>=1. Gene counts were normalized by the DESeq2 rlog function and adjusted for batch effects (RNA extraction day) using the Limma removeBatchEffect function. Color represents differentiation day and shape indicates RNA extraction batch.


More expressed in the later day

More expressed in the earlier day

Figure 4.4. Comparison of differentially expressed genes between pairwise day comparisons. UpSet plots indicating the number of differentially expressed genes (FDR<5%, absolute value LFC>1) shared between each pairwise day comparison. The top plot shows genes more expressed in the later day of a pairwise comparison (day 4 > day 0 for example) and the bottom plot shows genes more expressed in the earlier day (day 0 > day 4 for example). The horizontal bars on the right side of the plots show the number of total differential genes in each

pairwise comparison and the vertical bars on the top show the number of differential genes shared between comparisons.

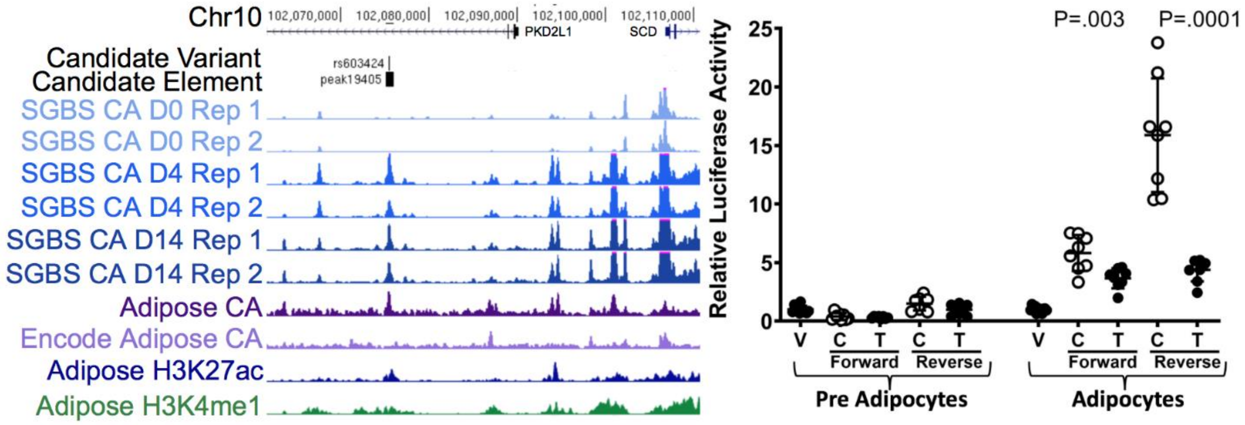


Figure 4.5. A variant near the *SCD* gene is found within a peak more accessible in adipocytes and shows context-dependent transcriptional reporter activity. (left) Genome browser shot of the GWAS locus for metabolic traits and plasma phospholipid levels that is colocalized with an adipose tissue eQTL for *SCD*. Selected ATAC signal tracks for each day of differentiation are shown. (right) Transcriptional activity of a 592-bp DNA element spanning peak19405 and containing rs603424 in SGBS cells at day 0 (preadipocyte) and day 2 (adipocyte) of differentiation. The DNA element was tested in the forward and reverse orientations relative to the genome. V: empty vector, CA: chromatin accessibility, d: differentiation day.

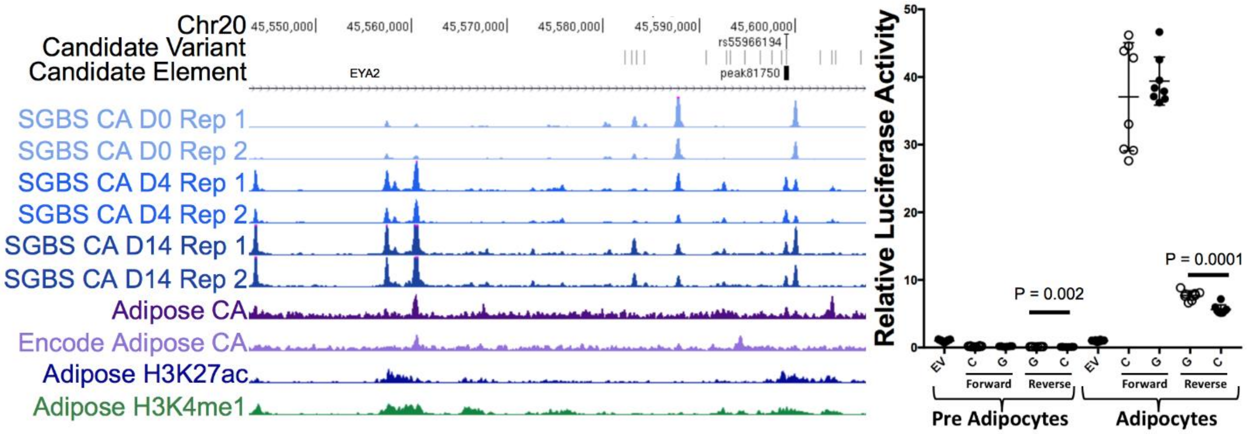


Figure 4.6. A variant near the *EYA2* gene is found within a peak more accessible in adipocytes and shows context-dependent transcriptional reporter activity. (left) Genome browser shot of the GWAS locus for type 2 diabetes (T2D) and triglycerides (TG) that is colocalized with an adipose tissue eQTL for *EYA2*. Selected ATAC signal tracks for each day of differentiation are shown. (right) Transcriptional activity of a 419-bp DNA element spanning peak81750 and containing rs55966194 in SGBS cells at day 0 (preadipocyte) and day 2 (adipocyte) of differentiation. The DNA element was tested in the forward and reverse orientations relative to the genome. V: empty vector, CA: chromatin accessibility, d: differentiation day.

Differentiation day	Number ATAC-seq replicates	Number RNA-seq replicates
0	10	6
2	2	2
4	10	6
14	5	4

Table 4.1. Study design. Number of ATAC-seq and RNA-seq replicates per day of adipocyte differentiation.

Sample	Day	Date	Total reads	Aligned reads	#peaks	%reads in peaks	TSS enrichment
S00b1r1	0	Oct10_2017	97.3	57.5	147,944	64	5.9
S00b1r2	0	Oct10_2017	75.1	45.3	133,522	52	5.9
S00b2r1	0	July2_2018	98.3	77.3	157,596	58	5.9
S00b2r2	0	July2_2018	71.2	51.0	132,142	48	6.5
S00b2r3	0	July2_2018	129.2	97.5	150,754	49	5.9
S00b2r4	0	July2_2018	65.5	43.4	116,318	38	5.8
S00b2r5	0	July2_2018	43.5	33.6	155,237	56	9.6
S00b2r6	0	July2_2018	60.7	40.9	118,621	43	6.2
S00b4r1	0	dec19_2018	189.6	97.2	191,773	49	7.2
S00b4r2	0	dec19_2018	201.2	102.1	162,246	48	5.4
S02b3r1	2	dec17_2018	146.1	77.9	164,274	52	6.3
S02b3r2	2	dec17_2018	92.0	47.2	113,441	26	5.5
S04b2r1	4	July2_2018	99.6	79.7	165,682	52	6.4
S04b2r2	4	July2_2018	101.6	67.9	141,195	45	7.1
S04b2r3	4	July2_2018	91.3	67.6	154,912	50	6.3
S04b2r4	4	July2_2018	86.1	62.3	144,845	48	6.8
S04b2r5	4	July2_2018	45.9	37.7	156,758	52	11.3
S04b2r6	4	July2_2018	147.1	106.9	164,104	49	7.4
S04b3r1	4	dec17_2018	157.4	80.8	163,193	50	6.7
S04b3r2	4	dec17_2018	117.0	72.4	154,669	49	5.8
S04b4r1	4	dec19_2018	159.1	112.9	172,115	53	5.9
S04b4r2	4	dec19_2018	47.5	34.3	124,029	49	5.9
S14b1r1	14	Oct10_2017	382.6	152.1	170,636	48	6.8
S14b1r2	14	Oct10_2017	245.1	137.9	167,758	45	8.5
S14b1r3	14	Oct10_2017	253.7	156.5	171,050	52	8.4
S14b3r1	14	dec17_2018	99.1	62.6	144,757	43	6.2
S14b3r2	14	dec17_2018	90.8	45.7	106,454	23	5.1

Table 4.2. ATAC-seq library quality metrics. Sample names use the following naming scheme: ‘S’ for SGBS, ‘##’ representing day of differentiation (‘00’ for day 0, ‘02’ for day 2, etc), ‘b#’ indicates batch number, and ‘r#’ indicates the replicate within a given batch. ‘Date’ is the date the ATAC-seq library was prepared. Reads are reported in millions of reads. ‘Aligned reads’ is the number of blacklist-filtered and non-duplicated reads aligning to nuclear chromosomes (mapq>20). [a]: These samples were prepared using 2.5uL of Tn5 and sequenced with 50-bp reads on a HiSeq 4000. All other samples were prepared using 5uL Tn5 and sequenced with 150-bp reads on a NovaSeq.

Term name	Binom Rank	Binom Raw P-Value	Binom FDR Q-Val	Binom Fold Enrichment	Binom Observed Region Hits	Binom Region Set Coverage
cellular response to insulin stimulus	81	7.E-57	1.E-54	2.3	482	3.E-02
response to insulin	82	1.E-56	2.E-54	2.0	598	4.E-02
insulin receptor signaling pathway	206	2.E-34	1.E-32	2.3	288	2.E-02
cellular response to hydrogen peroxide	400	4.E-21	1.E-19	2.1	194	1.E-02
white fat cell differentiation	495	6.E-18	2.E-16	2.8	95	6.E-03
positive regulation of glucose metabolic process	549	1.E-16	3.E-15	2.1	147	9.E-03
response to fluid shear stress	583	5.E-16	1.E-14	2.0	160	1.E-02
regulation of cardiac muscle hypertrophy in response to stress	608	1.E-15	3.E-14	2.9	75	5.E-03
regulation of fatty acid oxidation	611	2.E-15	4.E-14	2.3	121	8.E-03
semaphorin-plexin signaling pathway	618	3.E-15	6.E-14	2.1	144	9.E-03
branching in salivary gland morphogenesis	642	8.E-15	2.E-13	2.1	137	9.E-03
positive regulation of gluconeogenesis	709	1.E-13	2.E-12	2.9	68	4.E-03
cellular response to fluid shear stress	744	6.E-13	1.E-11	2.3	99	6.E-03
dichotomous subdivision of an epithelial terminal unit	899	1.E-10	1.E-09	2.2	83	5.E-03
clathrin coat assembly	916	1.E-10	2.E-09	2.4	68	4.E-03
intracellular lipid transport	941	2.E-10	3.E-09	2.3	74	5.E-03
response to laminar fluid shear stress	955	3.E-10	3.E-09	2.3	73	5.E-03
positive regulation of fatty acid oxidation	969	3.E-10	4.E-09	2.2	83	5.E-03
commissural neuron axon guidance	1013	9.E-10	1.E-08	2.4	64	4.E-03
negative regulation of vascular permeability	1463	5.E-07	4.E-06	2.0	63	4.E-03
substrate-dependent cell migration, cell extension	1646	2.E-06	2.E-05	2.0	53	3.E-03
intracellular cholesterol transport	1742	5.E-06	4.E-05	2.3	36	2.E-03
glycogen catabolic process	1916	2.E-05	1.E-04	2.0	45	3.E-03
regulation of glial cell migration	2960	2.E-03	8.E-03	2.0	22	1.E-03

Table 4.3. **Gene ontology enrichment for genes near adipocyte-dependent peaks**. Gene ontology enrichment was performed using the Genomic Regions Enrichment of Annotations Tool (GREAT) with the GO Biological Process ontology. We ran GREAT 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.

Term name	Binom Rank	Binom Raw P-Value	Binom FDR Q-Val	Binom Fold Enrichment	Binom Observed Region Hits	Binom Region Set Coverage
cell adhesion mediated by integrin	635	6.E-13	1.E-11	2.5	82	4.E-03
positive regulation of focal adhesion assembly	704	1.E-11	2.E-10	2.1	108	6.E-03
positive regulation of adherens junction organization	756	4.E-11	6.E-10	2.0	109	6.E-03

Table 4.4. **Gene ontology enrichment for genes near preadipocyte-dependent peaks**. Gene ontology enrichment was performed using the Genomic Regions Enrichment of Annotations Tool (GREAT) with the GO Biological Process ontology. We ran GREAT 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.

Motif	Consensus	P-value
CEBP	ATTGCGCAAC	1e-631
CEBP:AP1	DRTGTTGCAA	1.E-222
GRE	VAGRACAKWCTGTYC	1.E-199
GRE	NRGVACABNVTGTYCY	1.E-166
ARE	RGRACASNSTGTYCYB	1.E-163
PGR	AAGAACATWHTGTTC	1.E-148
Olig2	RCCATMTGTT	1.E-100
Atf4	MTGATGCAAT	1.E-95
Tcf21	NAACAGCTGG	1.E-90
Atoh1	VNRVCAGCTGGY	1.E-80
ZBTB18	AACATCTGGA	1.E-74
PR	VAGRACAKNCTGTBC	1.E-63
NF1-halfsite	YTGCCAAG	1.E-53
PPARE	TGACCTTTGCCCCA	1.E-52
Ap4	NAHCAGCTGD	1.E-50
AR-halfsite	CCAGGAACAG	1.E-49
Chop	ATTGCATCAT	1.E-44
Ptf1a	ACAGCTGTTN	1.E-42
RXR	TAGGGCAAAGGTCA	1.E-39
NeuroD1	GCCATCTGTT	1.E-39
HEB	VCAGCTGBNN	1.E-29
Ascl1	NNVVCAGCTGBN	1.E-28
EBF1	GTCCCCWGGGGA	1.E-27
BMAL1	GNCACGTG	1.E-27
MyoG	AACAGCTG	1.E-26
Myf5	BAACAGCTGT	1.E-22
HNF4a	CARRGKBCAAAGTYCA	1.E-20
Tcf12	VCAGCTGYTG	1.E-20
EBF	DGTCCCYRGGGA	1.E-19
Max	RCCACGTGGYYN	1.E-17
MyoD	RRCAGCTGYTSY	1.E-16
NPAS2	KCCACGTGAC	1.E-16
SCL	AVCAGCTG	1.E-16
E2A	DNRCAGCTGY	1.E-11
n-Myc	VRCCACGTGG	1.E-10
CEBP:CEBP	NTNATGCAAYMNNHTGMAAY	1.E-10
ZNF711	AGGCCTAG	1.E-09
Hoxc9	GGCCATAAATCA	1.E-09
Usf2	GTCACGTGGT	1.E-09
USF1	SGTCACGTGR	1.E-08
Esrrb	KTGACCTTGA	1.E-08
c-Myc	VVCCACGTGG	1.E-07
THRa	GGTCANYTGAGGWCA	1.E-06

TR4	GAGGTCAAAGGTCA	1.E-06
ZFX	AGGCCTRG	1.E-06
Nur77	TGACCTTTNCNT	1.E-05
CLOCK	GHCACGTG	1.E-05
Nr5a2	BTCAAGGTCA	1.E-05
HOXA9	GGCCATAAATCA	1.E-05
Foxa2	CYTGTTTACWYW	1.E-05
ZNF189	TGGAACAGMA	1.E-05
ZBTB12	NGNTCTAGAACCNGV	1.E-05
NF1	CYTGGCABNSTGCCAR	1.E-04
RBPJ:Ebox	GGGRAARRGRMCAGMTG	1.E-04

Table 4.5. **Motif enrichment in adipocyte-dependent peaks using Homer**.

We classified motifs with P-value<1.6x10^{-4} (0.05 / 319 tested motifs) as significantly enriched.

Motif	Consensus	P-value
Atf3	DATGASTCATHN	1E-1327
AP-1	VTGACTCATC	1E-1260
BATF	DATGASTCAT	1E-1250
Fra1	NNATGASTCATH	1E-1204
Fosl2	NATGASTCABNN	1E-640
Jun-AP1	GATGASTCATCN	1E-426
Bach2	TGCTGAGTCA	1E-142
TEAD	YCWGGAATGY	1E-91
Gata4	NBWGATAAGR	1E-88
Gata2	BBCTTATCTS	1E-86
Gata1	SAGATAAGRV	1E-85
GATA3	AGATAASR	1E-82
ETS1	ACAGGAAGTG	1E-77
TEAD4	CCWGGAATGY	1E-70
ERG	ACAGGAAGTG	1E-63
RUNX1	AAACCACARM	1E-62
Fli1	NRYTTCCGGH	1E-58
RUNX2	NWAACCACADNN	1E-57
TEAD2	CCWGGAATGY	1E-54
RUNX	SAAACCACAG	1E-50
Etv2	NNAYTTCCTGHN	1E-49
ETV1	AACCGGAAGT	1E-48
RUNX-AML	GCTGTGGTTW	1E-47
MafK	GCTGASTCAGCA	1E-42
EWS:ERG-fusion	ATTTCCTGTN	1E-42
Ets1-distal	MACAGGAAGT	1E-39
GABPA	RACCGGAAGT	1E-37
EHF	AVCAGGAAGT	1E-33
KLF5	DGGGYGKGGC	1E-32
EWS:FLI1-fusion	VACAGGAAAT	1E-32
MafA	TGCTGACTCA	1E-28
Pdx1	YCATYAATCA	1E-28
Elk1	HACTTCCGGY	1E-21
SPDEF	ASWTCCTGBT	1E-21
ETS:RUNX	RCAGGATGTGGT	1E-19
PU.1	AGAGGAAGTG	1E-17
Klf4	GCCACACCCA	1E-17
Elk4	NRYTTCCGGY	1E-17
NFAT:AP1	SARTGGAAAAWRTGAGTCAB	1E-13
ETS	AACCGGAAGT	1E-13
EKLF	NWGGGTGTGGCY	1E-11
Sox10	CCWTTGTYYB	1E-11
ELF1	AVCCGGAAGT	1E-11

Reverb	GTRGGTCASTGGGTCA	1E-11
PAX5	GTCACGCTCSCTGM	1E-10
Tcf4	ASATCAAAGGVA	1E-10
PAX3:FKHR-fusion	ACCRTGACTAATTNN	1E-10
ELF5	ACVAGGAAGT	1E-09
FOXA1	WAAGTAAACA	1E-09
SpiB	AAAGRGGAAGTG	1E-08
FOXM1	TRTTTACTTW	1E-08
Mef2c	DCYAAAAATAGM	1E-07
Egr2	NGCGTGGGCGGR	1E-07
Mef2b	GCTATTTTTGGM	1E-06
CRX	GCTAATCC	1E-06
GSC	RGGATTAR	1E-06
FOXA1	WAAGTAAACA	1E-05
Hoxb4	TGATTRATGGCY	1E-05
ZFP3	GGGTTTTGAAGGATGARTAGGAGTT	1E-05
Pax8	GTCATGCHTGRCTGS	1E-05
Mef2a	CYAAAAATAG	1E-04
Bapx1	TTRAGTGSYK	1E-04
Srebp1a	RTCACSCCAY	1E-04
Sox6	CCATTGTTNY	1E-04
Brn2	ATGAATATTC	1E-04
Mef2d	GCTATTTTTAGC	1E-04
NF-E2	GATGACTCAGCA	1E-04

Table 4.6. **Motif enrichment in preadipocyte-dependent peaks using Homer**.

We classified motifs with P-value<1.6x10^{-4} (0.05 / 319 tested motifs) as significantly enriched.

Sample	Day	Date	Total reads	Transcript reads	Fraction mRNA bases
S00b1r1	0	dec17_2018	66.0	52.5	0.88
S00b1r2	0	dec17_2018	69.1	55.6	0.89
S00b2r1	0	dec19_2018	58.9	46.3	0.86
S00b2r2	0	dec19_2018	60.1	49.0	0.90
S00b3r1	0	June11_2019	50.2	44.2	0.96
S00b3r2	0	June11_2019	57.5	50.3	0.96
S02b1r1	2	dec17_2018	50.4	36.4	0.79
S02b1r2	2	dec17_2018	60.7	46.2	0.84
S04b1r1	4	dec17_2018	53.2	42.2	0.87
S04b1r2	4	dec17_2018	58.6	45.9	0.87
S04b2r1	4	dec19_2018	61.4	47.9	0.86
S04b2r2	4	dec19_2018	65.5	50.5	0.85
S04b3r1	4	June11_2019	58.1	49.9	0.95
S04b3r2	4	June11_2019	49.5	41.5	0.95
S14b1r1	14	dec17_2018	51.2	41.2	0.90
S14b1r2	14	dec17_2018	52.6	42.1	0.89
S14b3r1	14	June11_2019	56.6	48.0	0.94
S14b3r2	14	June11_2019	54.0	45.4	0.94

Table 4.7. **RNA-seq library quality metrics**. Sample names use the following naming scheme: ‘S’ for SGBS, ‘##’ representing day of differentiation (‘00’ for day 0, ‘02’ for day 2, etc), ‘b#’ indicates batch number, and ‘r#’ indicates the replicate within a given batch. ‘Date’ is the date RNA was extracted. Reads are reported in millions of reads. ‘Transcriptome reads’ is the number of reads aligning to the transcriptome using salmon. ‘Fraction mRNA bases’ is the fraction of bases across aligned reads that fall within mRNA regions (exons and untranslated regions) calculated by the CollectRnaSeqMetrics tool from Picard.

PANTHER GO-Slim Biological Process	REFLIST (13918)	Tested (627)	expected	over/ under	gefold Enrich -ment	raw P-value	FDR
lipid metabolic process (GO:0006629)	260	36	11.7	+	3.1	2.E-08	4.E-05
monocarboxylic acid metabolic process (GO:0032787)	110	20	5.0	+	4.0	7.E-07	7.E-04
response to oxygen-containing compound (GO:1901700)	125	19	5.6	+	3.4	1.E-05	3.E-03
negative regulation of molecular function (GO:0044092)	113	17	5.1	+	3.3	4.E-05	6.E-03
fatty acid metabolic process (GO:0006631)	71	13	3.2	+	4.1	6.E-05	6.E-03
lipid biosynthetic process (GO:0008610)	117	17	5.3	+	3.2	6.E-05	6.E-03
negative regulation of catalytic activity (GO:0043086)	90	15	4.1	+	3.7	4.E-05	6.E-03
carboxylic acid metabolic process (GO:0019752)	244	26	11.0	+	2.4	1.E-04	1.E-02
oxoacid metabolic process (GO:0043436)	254	27	11.4	+	2.4	1.E-04	1.E-02
organic acid metabolic process (GO:0006082)	257	27	11.6	+	2.3	1.E-04	1.E-02
triglyceride metabolic process (GO:0006641)	15	6	0.7	+	8.9	2.E-04	1.E-02
cellular lipid metabolic process (GO:0044255)	223	24	10.1	+	2.4	2.E-04	1.E-02
lipid catabolic process (GO:0016042)	55	10	2.5	+	4.0	4.E-04	2.E-02
steroid metabolic process (GO:0008202)	26	7	1.2	+	6.0	4.E-04	2.E-02
transmembrane receptor protein tyrosine kinase signaling pathway (GO:0007169)	101	14	4.6	+	3.1	4.E-04	2.E-02
response to hormone (GO:0009725)	90	13	4.1	+	3.2	5.E-04	2.E-02
small molecule catabolic process (GO:0044282)	92	13	4.1	+	3.1	6.E-04	3.E-02
response to lipid (GO:0033993)	57	10	2.6	+	3.9	6.E-04	3.E-02
regulation of hormone levels (GO:0010817)	19	6	0.9	+	7.0	5.E-04	3.E-02
organic acid catabolic process (GO:0016054)	58	10	2.6	+	3.8	6.E-04	3.E-02
carboxylic acid catabolic process (GO:0046395)	58	10	2.6	+	3.8	6.E-04	3.E-02
response to endogenous stimulus (GO:0009719)	172	19	7.8	+	2.5	8.E-04	3.E-02
negative regulation of peptidase activity (GO:0010466)	30	7	1.4	+	5.2	9.E-04	3.E-02
negative regulation of endopeptidase activity (GO:0010951)	30	7	1.4	+	5.2	9.E-04	3.E-02
hormone metabolic process (GO:0042445)	14	5	0.6	+	7.9	1.E-03	3.E-02
negative regulation of proteolysis (GO:0045861)	31	7	1.4	+	5.0	1.E-03	3.E-02
negative regulation of cellular protein metabolic process (GO:0032269)	100	13	4.5	+	2.9	1.E-03	3.E-02

response to peptide (GO:1901652)	42	8	1.9	+	4.2	1.E-03	3.E-02
cellular response to hormone stimulus (GO:0032870)	76	11	3.4	+	3.2	1.E-03	3.E-02
response to peptide hormone (GO:0043434)	42	8	1.9	+	4.2	1.E-03	3.E-02
negative regulation of hydrolase activity (GO:0051346)	42	8	1.9	+	4.2	1.E-03	3.E-02
negative regulation of protein metabolic process (GO:0051248)	103	13	4.6	+	2.8	1.E-03	4.E-02
neutral lipid metabolic process (GO:0006638)	24	6	1.1	+	5.6	2.E-03	4.E-02
acylglycerol metabolic process (GO:0006639)	24	6	1.1	+	5.6	2.E-03	4.E-02
enzyme linked receptor protein signaling pathway (GO:0007167)	153	17	6.9	+	2.5	2.E-03	4.E-02
glucose homeostasis (GO:0042593)	16	5	0.7	+	6.9	2.E-03	4.E-02
cellular response to endogenous stimulus (GO:0071495)	157	17	7.1	+	2.4	2.E-03	4.E-02
ammonium ion metabolic process (GO:0097164)	25	6	1.1	+	5.3	2.E-03	4.E-02

Table 4.8. **Gene ontology term enrichment for adipocyte-dependent genes using the Panther overrepresentation test**. Terms with FDR<5% and gefold Enrichment>2 were considered significantly enriched.

PANTHER GO-Slim Biological Process	REFLIST (13918)	Tested (500)	expected	over/ under	gefold Enrich -ment	raw P-value	FDR
organelle fission (GO:0048285)	285	54	10.2	+	5.3	2.E-21	2.E-18
nuclear division (GO:0000280)	267	53	9.6	+	5.5	9.E-22	2.E-18
cell cycle (GO:0007049)	393	63	14.1	+	4.5	2.E-21	2.E-18
mitotic nuclear division (GO:0140014)	240	48	8.6	+	5.6	7.E-20	2.E-17
mitotic cell cycle process (GO:1903047)	240	48	8.6	+	5.6	7.E-20	3.E-17
cell cycle process (GO:0022402)	350	57	12.6	+	4.5	1.E-19	3.E-17
mitotic cell cycle (GO:0000278)	240	48	8.6	+	5.6	7.E-20	3.E-17
regulation of cell cycle (GO:0051726)	200	33	7.2	+	4.6	5.E-12	1.E-09
chromosome segregation (GO:0007059)	94	23	3.4	+	6.8	1.E-11	3.E-09
sister chromatid segregation (GO:0000819)	61	19	2.2	+	8.7	2.E-11	4.E-09
mitotic sister chromatid segregation (GO:0000070)	49	17	1.8	+	9.7	6.E-11	1.E-08
nuclear chromosome segregation (GO:0098813)	77	19	2.8	+	6.9	6.E-10	1.E-07
regulation of cell cycle process (GO:0010564)	84	19	3.0	+	6.3	2.E-09	3.E-07
negative regulation of cell cycle process (GO:0010948)	28	11	1.0	+	10.9	5.E-08	7.E-06
negative regulation of cell cycle (GO:0045786)	74	16	2.7	+	6.0	7.E-08	9.E-06
meiotic cell cycle (GO:0051321)	58	14	2.1	+	6.7	1.E-07	2.E-05
meiotic nuclear division (GO:0140013)	58	14	2.1	+	6.7	1.E-07	2.E-05
regulation of mitotic cell cycle (GO:0007346)	102	18	3.7	+	4.9	2.E-07	2.E-05
meiotic cell cycle process (GO:1903046)	58	14	2.1	+	6.7	1.E-07	2.E-05
negative regulation of nuclear division (GO:0051784)	13	8	0.5	+	17.1	3.E-07	2.E-05
multicellular organismal process (GO:0032501)	697	54	25.0	+	2.2	3.E-07	3.E-05
anatomical structure development (GO:0048856)	677	53	24.3	+	2.2	4.E-07	3.E-05
regulation of nuclear division (GO:0051783)	30	10	1.1	+	9.3	8.E-07	6.E-05
cytoskeleton organization (GO:0007010)	490	41	17.6	+	2.3	1.E-06	1.E-04
multicellular organism development (GO:0007275)	544	44	19.5	+	2.3	1.E-06	1.E-04
mitotic cell cycle phase transition (GO:0044772)	73	14	2.6	+	5.3	2.E-06	1.E-04
developmental process (GO:0032502)	760	55	27.3	+	2.0	2.E-06	2.E-04
cell cycle phase transition (GO:0044770)	78	14	2.8	+	5.0	3.E-06	2.E-04
regulation of mitotic cell cycle phase transition (GO:1901990)	46	11	1.7	+	6.7	3.E-06	2.E-04

cell cycle checkpoint (GO:0000075)	47	11	1.7	+	6.5	4.E-06	2.E-04
regulation of mitotic nuclear division (GO:0007088)	29	9	1.0	+	8.6	4.E-06	2.E-04
nervous system development (GO:0007399)	302	29	10.9	+	2.7	5.E-06	2.E-04
system development (GO:0048731)	473	39	17.0	+	2.3	6.E-06	3.E-04
reproduction (GO:0000003)	84	14	3.0	+	4.6	7.E-06	3.E-04
reproductive process (GO:0022414)	84	14	3.0	+	4.6	7.E-06	3.E-04
negative regulation of mitotic cell cycle (GO:0045930)	50	11	1.8	+	6.1	7.E-06	3.E-04
regulation of cell cycle phase transition (GO:1901987)	51	11	1.8	+	6.0	8.E-06	3.E-04
cellular developmental process (GO:0048869)	525	41	18.9	+	2.2	9.E-06	4.E-04
mitotic cell cycle checkpoint (GO:0007093)	35	9	1.3	+	7.2	2.E-05	6.E-04
antimicrobial humoral immune response mediated by antimicrobial peptide (GO:0061844)	7	5	0.3	+	19.9	3.E-05	1.E-03
cell differentiation (GO:0030154)	476	37	17.1	+	2.2	3.E-05	1.E-03
cellular response to lipopolysaccharide (GO:0071222)	21	7	0.8	+	9.3	4.E-05	1.E-03
regulation of organelle organization (GO:0033043)	238	23	8.6	+	2.7	4.E-05	1.E-03
supramolecular fiber organization (GO:0097435)	206	21	7.4	+	2.8	4.E-05	2.E-03
cellular response to molecule of bacterial origin (GO:0071219)	22	7	0.8	+	8.9	5.E-05	2.E-03
regulation of chromosome segregation (GO:0051983)	23	7	0.8	+	8.5	6.E-05	2.E-03
regulation of sister chromatid segregation (GO:0033045)	23	7	0.8	+	8.5	6.E-05	2.E-03
microtubule cytoskeleton organization involved in mitosis (GO:1902850)	42	9	1.5	+	6.0	6.E-05	2.E-03
cellular response to biotic stimulus (GO:0071216)	23	7	0.8	+	8.5	6.E-05	2.E-03
negative regulation of cell cycle phase transition (GO:1901988)	24	7	0.9	+	8.1	7.E-05	2.E-03
negative regulation of mitotic cell cycle phase transition (GO:1901991)	24	7	0.9	+	8.1	7.E-05	2.E-03
actin filament-based process (GO:0030029)	216	21	7.8	+	2.7	8.E-05	2.E-03
myeloid leukocyte migration (GO:0097529)	25	7	0.9	+	7.8	9.E-05	3.E-03
negative regulation of mitotic nuclear division (GO:0045839)	10	5	0.4	+	13.9	1.E-04	3.E-03
defense response (GO:0006952)	110	14	4.0	+	3.5	1.E-04	3.E-03
negative regulation of chromosome segregation (GO:0051985)	10	5	0.4	+	13.9	1.E-04	3.E-03
negative regulation of mitotic sister chromatid separation (GO:2000816)	10	5	0.4	+	13.9	1.E-04	3.E-03
negative regulation of sister chromatid segregation (GO:0033046)	10	5	0.4	+	13.9	1.E-04	3.E-03

negative regulation of mitotic sister chromatid segregation (GO:0033048)	10	5	0.4	+	13.9	1.E-04	3.E-03
negative regulation of chromosome separation (GO:1905819)	10	5	0.4	+	13.9	1.E-04	3.E-03
anatomical structure morphogenesis (GO:0009653)	343	28	12.3	+	2.3	1.E-04	3.E-03
actin cytoskeleton organization (GO:0030036)	207	20	7.4	+	2.7	1.E-04	3.E-03
regulation of chromosome separation (GO:1905818)	18	6	0.7	+	9.3	1.E-04	3.E-03
regulation of mitotic sister chromatid separation (GO:0010965)	18	6	0.7	+	9.3	1.E-04	3.E-03
neuron differentiation (GO:0030182)	193	19	6.9	+	2.7	2.E-04	4.E-03
mitotic sister chromatid separation (GO:0051306)	19	6	0.7	+	8.8	2.E-04	4.E-03
response to lipopolysaccharide (GO:0032496)	29	7	1.0	+	6.7	2.E-04	5.E-03
humoral immune response (GO:0006959)	12	5	0.4	+	11.6	2.E-04	5.E-03
establishment of chromosome localization (GO:0051303)	12	5	0.4	+	11.6	2.E-04	5.E-03
chromosome localization (GO:0050000)	12	5	0.4	+	11.6	2.E-04	5.E-03
regulation of mitotic sister chromatid segregation (GO:0033047)	20	6	0.7	+	8.4	2.E-04	5.E-03
response to molecule of bacterial origin (GO:0002237)	30	7	1.1	+	6.5	2.E-04	5.E-03
meiotic chromosome segregation (GO:0045132)	21	6	0.8	+	8.0	3.E-04	6.E-03
chemokine-mediated signaling pathway (GO:0070098)	13	5	0.5	+	10.7	3.E-04	6.E-03
leukocyte migration (GO:0050900)	31	7	1.1	+	6.3	3.E-04	6.E-03
neurogenesis (GO:0022008)	221	20	7.9	+	2.5	3.E-04	6.E-03
cell morphogenesis (GO:0000902)	205	19	7.4	+	2.6	3.E-04	6.E-03
generation of neurons (GO:0048699)	207	19	7.4	+	2.6	3.E-04	7.E-03
meiosis I cell cycle process (GO:0061982)	32	7	1.2	+	6.1	3.E-04	7.E-03
leukocyte chemotaxis (GO:0030595)	22	6	0.8	+	7.6	3.E-04	7.E-03
mitotic spindle organization (GO:0007052)	32	7	1.2	+	6.1	3.E-04	7.E-03
positive regulation of cell population proliferation (GO:0008284)	43	8	1.5	+	5.2	3.E-04	7.E-03
response to chemokine (GO:1990868)	14	5	0.5	+	9.9	4.E-04	7.E-03
cellular response to chemokine (GO:1990869)	14	5	0.5	+	9.9	4.E-04	7.E-03
chromosome separation (GO:0051304)	33	7	1.2	+	5.9	4.E-04	7.E-03
inflammatory response (GO:0006954)	57	9	2.1	+	4.4	4.E-04	8.E-03
chemotaxis (GO:0006935)	114	13	4.1	+	3.2	5.E-04	8.E-03
taxis (GO:0042330)	114	13	4.1	+	3.2	5.E-04	8.E-03
cellular component morphogenesis (GO:0032989)	222	20	8.0	+	2.5	5.E-04	8.E-03

granulocyte chemotaxis (GO:0071621)	15	5	0.5	+	9.3	5.E-04	9.E-03
cell division (GO:0051301)	46	8	1.7	+	4.8	5.E-04	9.E-03
mitotic metaphase plate congression (GO:0007080)	8	4	0.3	+	13.9	6.E-04	1.E-02
cellular response to lipid (GO:0071396)	36	7	1.3	+	5.4	6.E-04	1.E-02
response to other organism (GO:0051707)	88	11	3.2	+	3.5	6.E-04	1.E-02
response to external biotic stimulus (GO:0043207)	88	11	3.2	+	3.5	6.E-04	1.E-02
meiosis I (GO:0007127)	26	6	0.9	+	6.4	7.E-04	1.E-02
meiotic telophase I (GO:0007134)	26	6	0.9	+	6.4	7.E-04	1.E-02
response to biotic stimulus (GO:0009607)	89	11	3.2	+	3.4	7.E-04	1.E-02
telophase (GO:0051326)	26	6	0.9	+	6.4	7.E-04	1.E-02
response to bacterium (GO:0009617)	37	7	1.3	+	5.3	7.E-04	1.E-02
meiotic cell cycle phase (GO:0098762)	26	6	0.9	+	6.4	7.E-04	1.E-02
meiosis I cell cycle phase (GO:0098764)	26	6	0.9	+	6.4	7.E-04	1.E-02
M phase (GO:0000279)	26	6	0.9	+	6.4	7.E-04	1.E-02
mitotic spindle assembly checkpoint (GO:0007094)	9	4	0.3	+	12.4	8.E-04	1.E-02
DNA replication initiation (GO:0006270)	17	5	0.6	+	8.2	8.E-04	1.E-02
metaphase/anaphase transition of cell cycle (GO:0044784)	17	5	0.6	+	8.2	8.E-04	1.E-02
metaphase/anaphase transition of mitotic cell cycle (GO:0007091)	17	5	0.6	+	8.2	8.E-04	1.E-02
biological phase (GO:0044848)	27	6	1.0	+	6.2	8.E-04	1.E-02
cell cycle phase (GO:0022403)	27	6	1.0	+	6.2	8.E-04	1.E-02
actin filament organization (GO:0007015)	139	14	5.0	+	2.8	9.E-04	1.E-02
DNA biosynthetic process (GO:0071897)	93	11	3.3	+	3.3	1.E-03	1.E-02
regulation of cellular component organization (GO:0051128)	357	26	12.8	+	2.0	1.E-03	1.E-02
regulation of actin filament-based process (GO:0032970)	94	11	3.4	+	3.3	1.E-03	1.E-02
DNA metabolic process (GO:0006259)	278	22	10.0	+	2.2	1.E-03	2.E-02
cytokinesis (GO:0000910)	40	7	1.4	+	4.9	1.E-03	2.E-02
metaphase plate congression (GO:0051310)	10	4	0.4	+	11.1	1.E-03	2.E-02
membrane fission (GO:0090148)	40	7	1.4	+	4.9	1.E-03	2.E-02
negative regulation of organelle organization (GO:0010639)	54	8	1.9	+	4.1	1.E-03	2.E-02
negative regulation of molecular function (GO:0044092)	113	12	4.1	+	3.0	1.E-03	2.E-02
cell chemotaxis (GO:0060326)	30	6	1.1	+	5.6	1.E-03	2.E-02
actin polymerization or depolymerization (GO:0008154)	70	9	2.5	+	3.6	2.E-03	2.E-02

extracellular matrix organization (GO:0030198)	44	7	1.6	+	4.4	2.E-03	2.E-02
multi-organism process (GO:0051704)	118	12	4.2	+	2.8	2.E-03	2.E-02
DNA replication (GO:0006260)	72	9	2.6	+	3.5	2.E-03	2.E-02
negative regulation of cell death (GO:0060548)	45	7	1.6	+	4.3	2.E-03	3.E-02
peripheral nervous system development (GO:0007422)	5	3	0.2	+	16.7	2.E-03	3.E-02
positive regulation of supramolecular fiber organization (GO:1902905)	45	7	1.6	+	4.3	2.E-03	3.E-02
response to external stimulus (GO:0009605)	253	20	9.1	+	2.2	2.E-03	3.E-02
locomotion (GO:0040011)	254	20	9.1	+	2.2	2.E-03	3.E-02
positive regulation of cytoskeleton organization (GO:0051495)	46	7	1.7	+	4.2	2.E-03	3.E-02
spindle organization (GO:0007051)	60	8	2.2	+	3.7	2.E-03	3.E-02
plasma membrane bounded cell projection morphogenesis (GO:0120039)	123	12	4.4	+	2.7	3.E-03	3.E-02
neuron projection morphogenesis (GO:0048812)	123	12	4.4	+	2.7	3.E-03	3.E-02
cell projection morphogenesis (GO:0048858)	123	12	4.4	+	2.7	3.E-03	3.E-02
negative regulation of cellular component organization (GO:0051129)	76	9	2.7	+	3.3	3.E-03	3.E-02
cell part morphogenesis (GO:0032990)	124	12	4.5	+	2.7	3.E-03	3.E-02
extracellular structure organization (GO:0043062)	48	7	1.7	+	4.1	3.E-03	3.E-02
DNA-dependent DNA replication (GO:0006261)	62	8	2.2	+	3.6	3.E-03	3.E-02
negative regulation of chromosome organization (GO:2001251)	24	5	0.9	+	5.8	3.E-03	3.E-02
cell-cell adhesion (GO:0098609)	93	10	3.3	+	3.0	3.E-03	3.E-02
system process (GO:0003008)	93	10	3.3	+	3.0	3.E-03	3.E-02
regulation of actin cytoskeleton organization (GO:0032956)	93	10	3.3	+	3.0	3.E-03	3.E-02
cellular response to cytokine stimulus (GO:0071345)	65	8	2.3	+	3.4	4.E-03	4.E-02
regulation of actin polymerization or depolymerization (GO:0008064)	65	8	2.3	+	3.4	4.E-03	4.E-02
regulation of G2/M transition of mitotic cell cycle (GO:0010389)	15	4	0.5	+	7.4	4.E-03	4.E-02
regulation of actin filament length (GO:0030832)	65	8	2.3	+	3.4	4.E-03	4.E-02
cell development (GO:0048468)	255	19	9.2	+	2.1	4.E-03	4.E-02
microtubule cytoskeleton organization (GO:0000226)	255	19	9.2	+	2.1	4.E-03	4.E-02
cellular response to organic substance (GO:0071310)	290	21	10.4	+	2.0	4.E-03	4.E-02
regulation of neuron death (GO:1901214)	7	3	0.3	+	11.9	4.E-03	4.E-02
cell motility (GO:0048870)	201	16	7.2	+	2.2	4.E-03	4.E-02

mitotic DNA replication checkpoint (GO:0033314)	7	3	0.3	+	11.9	4.E-03	4.E-02
localization of cell (GO:0051674)	201	16	7.2	+	2.2	4.E-03	4.E-02
neuron apoptotic process (GO:0051402)	7	3	0.3	+	11.9	4.E-03	4.E-02
chromosome condensation (GO:0030261)	16	4	0.6	+	7.0	4.E-03	5.E-02
regulation of cytoskeleton organization (GO:0051493)	132	12	4.7	+	2.5	4.E-03	5.E-02
regulation of cytokinesis (GO:0032465)	16	4	0.6	+	7.0	4.E-03	5.E-02
immune system process (GO:0002376)	182	15	6.5	+	2.3	5.E-03	5.E-02
regulation of cell cycle G2/M phase transition (GO:1902749)	16	4	0.6	+	7.0	4.E-03	5.E-02
regulation of cell death (GO:0010941)	99	10	3.6	+	2.8	5.E-03	5.E-02
regulation of cell division (GO:0051302)	16	4	0.6	+	7.0	4.E-03	5.E-02
cellular response to growth factor stimulus (GO:0071363)	68	8	2.4	+	3.3	5.E-03	5.E-02
response to growth factor (GO:0070848)	69	8	2.5	+	3.2	5.E-03	5.E-02

Table 4.9. **Gene ontology term enrichment for preadipocyte-dependent genes using the Panther overrepresentation test**. Terms with FDR<5% and gefold Enrichment>2 were considered significantly enriched.

GWAS trait/s	GWAS lead variant	eQTL gene	Peak ID	Peak context	Peak proxy variant	r2 with lead
Metabolic traits, Phospholipid levels	rs603424	_SCD*_	peak19405	adipocyte	rs603424	1.00
Metabolic traits, Phospholipid levels	rs603424	_AL139819.1*_	peak19405	adipocyte	rs603424	1.00
Coronary heart disease	rs2048327	_SLC22A3*_	peak122426	preadipocyte	rs3106162	0.98
Lipid metabolism phenotypes, HDL, Metabolic traits, Sphingolipid levels, TG, LDL, Metabolic syndrome, Phospholipid levels, TC, FGlu, FGlu-related traits	rs174547	_FADS1*_	peak24179	adipocyte	rs174541	0.91
eGFRcrea	rs1933182	_SYPL2*_	peak6999	adipocyte	rs3768495	0.86
eGFRcrea	rs1933182	_SYPL2*_	peak6999	adipocyte	rs3768493	0.86
eGFRcrea	rs1933182	_SYPL2*_	peak6999	adipocyte	rs10858091	0.86
eGFRcrea	rs1933182	_SYPL2*_	peak6999	adipocyte	rs3768494	0.85
eGFRcrea	rs1933182	_SYPL2*_	peak6999	adipocyte	rs2140924	0.85
eGFRcrea	rs1933182	_SYPL2*_	peak6989	adipocyte	rs370088	0.85
Coronary heart disease	rs2048327	_SLC22A3*_	peak122422	adipocyte	rs2661839	0.82
HDL, LDL, TG	rs12748152	_RP5-968P14.2_	peak2035	preadipocyte	rs57217461	1.00
HDL, LDL, TG	rs12748152	_PIGV_	peak2035	preadipocyte	rs57217461	1.00
HDL, LDL, TG	rs12748152	_RP5-968P14.2_	peak2042	preadipocyte	rs58421016	1.00
HDL, LDL, TG	rs12748152	_PIGV_	peak2042	preadipocyte	rs58421016	1.00
HDL, LDL, TG	rs12748152	_RP5-968P14.2_	peak2045	preadipocyte	rs34618114	1.00
HDL, LDL, TG	rs12748152	_PIGV_	peak2045	preadipocyte	rs34618114	1.00
BMI	rs2275426	_MAST2_	peak3612	preadipocyte	rs7540325	1.00
BMI	rs2275426	_MAST2_	peak3618	preadipocyte	rs4134386	1.00
Glycated hemoglobin	rs6684514	_C1orf85_	peak8580	adipocyte	rs2277871	1.00
Glycated hemoglobin	rs6684514	_CCT3_	peak8580	adipocyte	rs2277871	1.00
Glycated hemoglobin	rs6684514	_C1orf85_	peak8580	adipocyte	rs2277870	1.00
Glycated hemoglobin	rs6684514	_CCT3_	peak8580	adipocyte	rs2277870	1.00
Glycated hemoglobin	rs6684514	_C1orf85_	peak8587	preadipocyte	rs111850227	1.00
Glycated hemoglobin	rs6684514	_CCT3_	peak8587	preadipocyte	rs111850227	1.00
BMI	rs912768	_PRDX6_	peak9673	adipocyte	rs1886638	1.00
BMI	rs912768	_PRDX6_	peak9673	adipocyte	rs912768	1.00
WHRadjBMI, WHR	rs17326656	_LHCGR_	peak69622	adipocyte	rs17326656	1.00
blood urea nitrogen	rs11123170	_PAX8_	peak72540	adipocyte	rs10175462	1.00
WHR	rs1789882	_ADH1A_	peak101238	adipocyte	rs1693458	1.00
WHR	rs1789882	_ADH1A_	peak101238	adipocyte	rs3133155	1.00
WHRadjBMI	rs1544474	_LAMB1_	peak128118	adipocyte	rs1544474	1.00
Metabolic traits	rs2066938	_UNC119B_	peak35333	adipocyte	rs2066938	1.00
WHRadjBMI, WHR, HDL, TG, BMI, T2D	rs7307277	_RP11-380L11.4_	peak35619	adipocyte	rs7978610	1.00
BMI	rs912768	_PRDX6_	peak9679	adipocyte	rs1461025	1.00
WHRadjBMI	rs11917361	_CCDC12_	peak89520	preadipocyte	rs4018905	1.00
WHRadjBMI	rs11917361	_SETD2_	peak89520	preadipocyte	rs4018905	1.00

WHRadjBMI	rs11917361	*ELP6*	peak89520	preadipocyte	rs4018905	1.00
WHRadjBMI	rs11917361	*CCDC12*	peak89520	preadipocyte	rs9311403	1.00
WHRadjBMI	rs11917361	*SETD2*	peak89520	preadipocyte	rs9311403	1.00
WHRadjBMI	rs11917361	*ELP6*	peak89520	preadipocyte	rs9311403	1.00
T2D, TG	rs6063048	*EYA2*	peak81750	adipocyte	rs55966194	0.99
BMI	rs1020548	*BEND6*	peak117576	preadipocyte	rs17685277	0.99
eGFRcrea	rs2928148	*INO80*	peak45130	adipocyte	rs11856848	0.99
BMI	rs9846123	*WDR6*	peak89666	preadipocyte	rs9311433	0.99
BMI	rs9846123	*ARIH2*	peak89666	preadipocyte	rs9311433	0.99
BMI	rs9846123	*NCKIPSD*	peak89666	preadipocyte	rs9311433	0.99
BMI	rs9846123	*CCDC36*	peak89666	preadipocyte	rs9311433	0.99
BMI	rs9846123	*RP11-3B7.1*	peak89666	preadipocyte	rs9311433	0.99
WHRadjBMI, WHR, HDL, TG, BMI, T2D	rs7307277	*RP11-380L11.4*	peak35626	adipocyte	rs11057413	0.99
BMI	rs886444	*CDK5RAP3*	peak56838	preadipocyte	rs4794333	0.99
BMI	rs2814992	*UHRF1BP1*	peak116157	preadipocyte	rs2814969	0.99
BMI	rs2814992	*SPC*	peak116157	preadipocyte	rs2814969	0.99
BMI	rs2814992	*UHRF1BP1*	peak116157	preadipocyte	rs2814970	0.99
BMI	rs2814992	*SPC*	peak116157	preadipocyte	rs2814970	0.99
BMI	rs2814992	*UHRF1BP1*	peak116159	preadipocyte	rs9469863	0.99
BMI	rs2814992	*SPC*	peak116159	preadipocyte	rs9469863	0.99
BMI	rs2820311	*LMOD1*	peak11035	adipocyte	rs2820316	0.99
BMI	rs2820311	*IPO9*	peak11035	adipocyte	rs2820316	0.99
BMI	rs6494481	*TRIP4*	peak46646	adipocyte	rs2470895	0.99
TC	rs2854322	*EVI2A*	peak55460	preadipocyte	rs2905872	0.99
BMI	rs498240	*NELFE*	peak115932	adipocyte	rs623529	0.98
BMI	rs498240	*NELFE*	peak115932	adipocyte	rs621701	0.98
HDL, LDL, TG	rs12748152	*RP5-968P14.2*	peak2055	adipocyte	rs12760759	0.98
HDL, LDL, TG	rs12748152	*PIGV*	peak2055	adipocyte	rs12760759	0.98
WHRadjBMI, WHR	rs7798002	*AC003090.1*	peak124070	preadipocyte	rs10260677	0.98
WHRadjBMI, WHR, HDL, TG, BMI, T2D	rs7307277	*RP11-380L11.4*	peak35626	adipocyte	rs11057412	0.98
BMI	rs2814992	*UHRF1BP1*	peak116142	adipocyte	rs2814972	0.98
BMI	rs2814992	*SPC*	peak116142	adipocyte	rs2814972	0.98
BMI	rs2814992	*UHRF1BP1*	peak116157	preadipocyte	rs2764207	0.98
BMI	rs2814992	*SPC*	peak116157	preadipocyte	rs2764207	0.98
BMI	rs2814992	*UHRF1BP1*	peak116157	preadipocyte	rs2814968	0.98
BMI	rs2814992	*SPC*	peak116157	preadipocyte	rs2814968	0.98
BMI	rs2814992	*UHRF1BP1*	peak116159	preadipocyte	rs9462015	0.98
BMI	rs2814992	*SPC*	peak116159	preadipocyte	rs9462015	0.98
BMI	rs2275426	*MAST2*	peak3625	adipocyte	rs785481	0.98
BMI	rs2275426	*MAST2*	peak3626	adipocyte	rs6675726	0.98
Metabolic traits	rs2066938	*UNC119B*	peak35335	adipocyte	rs34673751	0.98
WHRadjBMI, WHR	rs7798002	*AC003090.1*	peak124071	preadipocyte	rs4722530	0.98
WHRadjBMI	rs17154889	*PPIP5K2*	peak109353	preadipocyte	rs35100629	0.97
T2D	rs4932265	*C15orf38-AP3S2*	peak48620	adipocyte	rs12594774	0.97
T2D	rs4932265	*AP3S2*	peak48620	adipocyte	rs12594774	0.97
WHRadjBMI, WHR	rs7798002	*AC003090.1*	peak124071	preadipocyte	rs10262483	0.97
Glycated hemoglobin	rs6684514	*C1orf85*	peak8578	preadipocyte	rs2273832	0.97
Glycated hemoglobin	rs6684514	*CCT3*	peak8578	preadipocyte	rs2273832	0.97
Glycated hemoglobin	rs6684514	*C1orf85*	peak8578	preadipocyte	rs2273833	0.97
Glycated hemoglobin	rs6684514	*CCT3*	peak8578	preadipocyte	rs2273833	0.97
WHRadjBMI	rs17154889	*PPIP5K2*	peak109353	preadipocyte	rs17154825	0.97
BMI	rs6738445	*AC068039.4*	peak74899	preadipocyte	rs10200608	0.96

WHR	rs9988	*MRPS7*	peak58722	adipocyte	rs1005714	0.96
HDL, BMI, TC	rs7941030	*UBASH3B*	peak27761	preadipocyte	rs7118212	0.96
HDL, BMI, TC	rs7941030	*UBASH3B*	peak27761	preadipocyte	rs7101940	0.96
WHR	rs9988	*MRPS7*	peak58722	adipocyte	rs1005713	0.96
WHRadjBMI	rs1544474	*LAMB1*	peak128118	adipocyte	rs41281051	0.96
BMI	rs4886506	*SCAPER*	peak47745	preadipocyte	rs11629727	0.96
BMI	rs12574668	*C11orf49*	peak23743	adipocyte	rs7125907	0.96
HDL, BMI, TC	rs7941030	*UBASH3B*	peak27756	preadipocyte	rs10790517	0.95
WHRadjBMI, WHR	rs672356	*CCDC144B*	peak54881	adipocyte	rs4924750	0.95
WHRadjBMI, WHR	rs672356	*RP1-178F10.3*	peak54881	adipocyte	rs4924750	0.95
T2D	rs4932265	*C15orf38-AP3S2*	peak48620	adipocyte	rs2165069	0.95
T2D	rs4932265	*AP3S2*	peak48620	adipocyte	rs2165069	0.95
WHRadjBMI, WHR	rs672356	*CCDC144B*	peak54881	adipocyte	rs7211382	0.94
WHRadjBMI, WHR	rs672356	*RP1-178F10.3*	peak54881	adipocyte	rs7211382	0.94
HDL, BMI, TC	rs7941030	*UBASH3B*	peak27761	preadipocyte	rs10790519	0.94
HDL, BMI	rs2013208	*RBM6*	peak89770	preadipocyte	rs2252833	0.94
BMI	rs6738445	*AC068039.4*	peak74904	preadipocyte	rs6758704	0.94
BMI, HDL	rs9931407	*CTC-277H1.7*	peak51879	preadipocyte	rs115328599	0.94
BMI, HDL	rs9931407	*FHOD1*	peak51879	preadipocyte	rs115328599	0.94
BMI	rs10497807	*PLCL1*	peak76404	preadipocyte	rs10192466	0.93
WHRadjBMI	rs17764730	*CTC-228N24.3*	peak110427	preadipocyte	rs1560637	0.93
WHRadjBMI	rs17154889	*PPIP5K2*	peak109359	preadipocyte	rs62362544	0.93
BMI	rs10497807	*PLCL1*	peak76405	adipocyte	rs9288281	0.93
WHRadjBMI	rs17154889	*PPIP5K2*	peak109353	preadipocyte	rs6862616	0.92
Metabolic traits, eGFRcrea	rs13391552	*ALMS1P*	peak70992	adipocyte	rs7604682	0.92
BMI	rs10268050	*NUDCD3*	peak125300	adipocyte	rs10246459	0.92
BMI	rs10268050	*NUDCD3*	peak125300	adipocyte	rs10231203	0.92
BMI	rs10268050	*NUDCD3*	peak125300	adipocyte	rs10249846	0.92
BMI	rs10268050	*NUDCD3*	peak125301	preadipocyte	rs10229330	0.92
HDL, BMI, TC	rs7941030	*UBASH3B*	peak27761	preadipocyte	rs10892873	0.92
WHRadjBMI	rs11917361	*CCDC12*	peak89520	preadipocyte	rs13098228	0.92
WHRadjBMI	rs11917361	*SETD2*	peak89520	preadipocyte	rs13098228	0.92
WHRadjBMI	rs11917361	*ELP6*	peak89520	preadipocyte	rs13098228	0.92
WHRadjBMI	rs11917361	*CCDC12*	peak89520	preadipocyte	rs13061071	0.92
WHRadjBMI	rs11917361	*SETD2*	peak89520	preadipocyte	rs13061071	0.92
WHRadjBMI	rs11917361	*ELP6*	peak89520	preadipocyte	rs13061071	0.92
T2D	rs12681990	*ZNF703*	peak132216	adipocyte	rs10955009	0.92
LDL, TC, Phytosterol levels	rs72875462	*ABCG5*	peak69153	preadipocyte	rs114938914	0.91
HDL, BMI, TC	rs7941030	*UBASH3B*	peak27761	preadipocyte	rs61679561	0.91
WHRadjBMI	rs7479183	*PIDD*	peak21210	adipocyte	rs11246319	0.91
WHRadjBMI	rs7479183	*AP006621.6*	peak21210	adipocyte	rs11246319	0.91
eGFRcrea	rs2928148	*INO80*	peak45118	adipocyte	rs4923890	0.91
BMI	rs12964689	*NPC1*	peak60585	preadipocyte	rs1788783	0.91
BMI	rs12964689	*C18orf8*	peak60585	preadipocyte	rs1788783	0.91
BMI, WHR	rs998732	*YJEFN3*	peak64504	adipocyte	rs2905433	0.91
BMI, WHR	rs998732	*YJEFN3*	peak64504	adipocyte	rs2905433	0.91
BMI, WHR	rs998732	*YJEFN3*	peak64521	adipocyte	rs76095338	0.91
BMI, WHR	rs998732	*YJEFN3*	peak64521	adipocyte	rs76095338	0.91
Metabolic traits	rs6499165	*PLA2G15*	peak52001	adipocyte	rs3961283	0.91

BMI, HDL	rs9931407	*CTC-277H1.7*	peak51851	adipocyte	rs114556591	0.90
BMI, HDL	rs9931407	*FHOD1*	peak51851	adipocyte	rs114556591	0.90
BMI, HDL	rs9931407	*CTC-277H1.7*	peak51856	preadipocyte	rs115992891	0.90
BMI, HDL	rs9931407	*FHOD1*	peak51856	preadipocyte	rs115992891	0.90
Metabolic traits, eGFRcrea	rs13391552	*ALMS1P*	peak70992	adipocyte	rs7580750	0.90
WHRadjBMI, WHR, HDL, TG, BMI, T2D	rs7307277	*RP11-380L11.4*	peak35616	preadipocyte	rs34854841	0.90
BMI	rs2230590	*MST1R*	peak89770	preadipocyte	rs2252833	0.90
BMI	rs9838283	*RP11-804H8.6*	peak89841	preadipocyte	rs34889917	0.89
BMI	rs9838283	*HEMK1*	peak89841	preadipocyte	rs34889917	0.89
BMI	rs62034325	*SULT1A2*	peak50669	adipocyte	rs12446550	0.89
LDL	rs9438900	*TMEM50A*	peak1900	preadipocyte	rs10903129	0.88
TG, HDL, Hypertriglyceridemia, CRP, BMI	rs17145738	*BCL7B*	peak126180	adipocyte	rs13231516	0.88
T2D	rs1061810	*HSD17B12*	peak23483	preadipocyte	rs10838175	0.88
T2D, HDL, TG, BMI, WC, Adiposity, WHR, Adiponectin, FIns	rs2972144	*IRS1*	peak78188	preadipocyte	rs2943656	0.87
BMI	rs919433	*AC013264.2*	peak76364	preadipocyte	rs7574271	0.87
WHR, BMI, TG	chr10:65318766	*BF2*	peak17144	adipocyte	rs10761756	0.87
BMI	rs1075901	*ZSWIM7*	peak54626	preadipocyte	rs1065822	0.87
BMI	rs1075901	*ADORA2B*	peak54626	preadipocyte	rs1065822	0.87
Metabolic traits, eGFRcrea	rs13391552	*ALMS1P*	peak70995	adipocyte	rs13538	0.87
Glycated hemoglobin levels	rs6980507	*SMIM19*	peak132658	adipocyte	rs2923447	0.86
BMI	rs6494481	*TRIP4*	peak46654	preadipocyte	rs28635082	0.86
Metabolic traits, eGFRcrea	rs13391552	*ALMS1P*	peak70995	adipocyte	rs4547554	0.86
LDL	rs9438900	*TMEM50A*	peak1902	preadipocyte	rs6699113	0.85
BMI	rs2814992	*UHRF1BP1*	peak116159	preadipocyte	rs9394248	0.85
BMI	rs2814992	*SPC*	peak116159	preadipocyte	rs9394248	0.85
Cardiac hypertrophy	rs1320448	*OBFC1*	peak19765	preadipocyte	rs73329737	0.84
BMI, WHR	rs524281	*PACS1*	peak24632	preadipocyte	rs7942894	0.81
BMI, WHR	rs524281	*RP11-755F10.1*	peak24632	preadipocyte	rs7942894	0.81
WHRadjBMI	rs11917361	*CCDC12*	peak89532	preadipocyte	rs11710322	0.80
WHRadjBMI	rs11917361	*SETD2*	peak89532	preadipocyte	rs11710322	0.80
WHRadjBMI	rs11917361	*ELP6*	peak89532	preadipocyte	rs11710322	0.80

Table 4.10. **Proxy variants found within context-dependent peaks at cardiometabolic GWAS loci colocalized with adipose tissue eQTL signals**. GWAS loci colocalized with adipose tissue eQTL signals were obtained from Raulerson et al[7]. 'r^2 with lead' is the linkage disequilibrium (LD) r^2 between the GWAS lead variant and the proxy variant found within the peak. The table is first sorted by whether the eQTL gene is adipocyte context-dependent (indicated by asterisks), and then by LD 'r^2 with lead) in decreasing order. BMI: body mass index, WC: waist

circumference, WHR: waist-hip ratio, WHRadjBMI: waist-hip ratio adjusted for body mass index, HDL: high density lipoprotein cholesterol, LDL: low density lipoprotein cholesterol, TC: total cholesterol, TG: triglycerides, T2D: type 2 diabetes, FGlu: fasting glucose, Fins: fasting insulin, CRP: C-reactive protein, eGFRcrea: estimated glomerular filtration rate for creatinine.

GWAS trait	Total colocalized loci	#with adipocyte peak	%with adipocyte peak	#with preadipocyte peak	%with preadipocyte peak
BMI	84	15	17.9	21	25
WHRadjBMI	47	5	10.6	5	10.6
WHR	39	7	17.9	4	10.3
T2D	28	4	14.3	3	10.7
HDL	25	5	20	6	24
TG	18	6	33.3	3	16.7
LDL	15	2	13.3	3	20
Metabolic traits	12	5	41.7	0	0
TC	8	1	12.5	3	37.5
eGFRcrea	5	3	60	0	0

Table 4.11. **Number of GWAS-colocalized eQTL signals with a proxy variant in a context-dependent peak divided by GWAS trait**. Only traits with at least 5 GWAS-colocalized eQTL signals are shown. The table is sorted by the total number of loci per trait in decreasing order. BMI: body mass index, WHRadjBMI: waist-hip ratio adjusted for body mass index, WHR, waist-hip ratio, T2D: type 2 diabetes, HDL: high density lipoprotein cholesterol, TG: triglycerides, LDL: low density lipoprotein cholesterol, TC: total cholesterol, eGFRcrea: estimated glomerular filtration rate for creatinine.

REFERENCES

1. Shungin, D., Winkler, T.W., Croteau-Chonka, D.C., Ferreira, T., Locke, A.E., Magi, R., Strawbridge, R.J., Pers, T.H., Fischer, K., Justice, A.E., et al. (2015). New genetic loci link adipose and insulin biology to body fat distribution. Nature 518, 187–196.

2. Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature 518, 317.

3. Cannon, M.E., Currin, K.W., Young, K.L., Perrin, H.J., Vadlamudi, S., Safi, A., Song, L., Wu, Y., Wabitsch, M., Laakso, M., et al. (2019). Open chromatin profiling in adipose tissue marks genomic regions with functional roles in cardiometabolic traits. G3: Genes, Genomes, Genetics 9, 2521–2533.

4. Calderon, D., Nguyen, M.L.T., Mezger, A., Kathiria, A., Müller, F., Nguyen, V., Lescano, N., Wu, B., Trombetta, J., Ribado, J.V., et al. (2019). Landscape of stimulation-responsive chromatin across diverse human immune cells. Nat. Genet. 51, 1494–1505.

5. Grundberg, E., Small, K.S., Hedman, Å.K., Nica, A.C., Buil, A., Keildson, S., Bell, J.T., Yang, T.-P., Meduri, E., Barrett, A., et al. (2012). Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat. Genet. 44, 1084–1089.

6. Civelek, M., Wu, Y., Pan, C., Raulerson, C.K., Ko, A., He, A., Tilford, C., Saleem, N.K., Stancakova, A., Scott, L.J., et al. (2017). Genetic Regulation of Adipose Gene Expression and Cardio-Metabolic Traits. Am. J. Hum. Genet. 100, 428–443.

7. Raulerson, C.K., Ko, A., Kidd, J.C., Currin, K.W., Brotman, S.M., Cannon, M.E., Wu, Y., Spracklen, C.N., Jackson, A.U., Stringham, H.M., et al. (2019). Adipose Tissue Gene Expression Associations Reveal Hundreds of Candidate Genes for Cardiometabolic Traits. Am. J. Hum. Genet. 105, 773–787.

8. Scott, L.J., Erdos, M.R., Huyghe, J.R., Welch, R.P., Beck, A.T., Wolford, B.N., Chines, P.S., Didion, J.P., Narisu, N., Stringham, H.M., et al. (2016). The genetic regulatory signature of type 2 diabetes in human skeletal muscle. Nat Commun 7, 11764.

9. Strunz, T., Grassmann, F., Gayán, J., Nahkuri, S., Souza-Costa, D., Maugeais, C., Fauser, S., Nogoceke, E., and Weber, B.H.F. (2018). 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.

10. Etheridge, A.S., Gallins, P.J., Jima, D., Broadaway, K.A., Ratain, M.J., Schuetz, E., Schadt, E., Schroder, A., Molony, C., Zhou, Y., et al. (2020). 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.

11. Alasoo, K., Rodrigues, J., Mukhopadhyay, S., Knights, A.J., Mann, A.L., Kundu, K., Hale, and C., Dougan, G., and Gaffney, D.J. (2018). Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. Nat. Genet.

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

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

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

15. Ehrlund, A., Acosta, J.R., Björk, C., Hedén, P., Douagi, I., Arner, P., and Laurencikiene, J. (2017). The cell-type specific transcriptome in human adipose tissue and influence of obesity on adipocyte progenitors. Sci Data 4, 170164.

16. Garske, K.M., Pan, D.Z., Miao, Z., Bhagat, Y.V., Comenho, C., Robles, C.R., Benhammou, J.N., Alvarez, M., Ko, A., Ye, C.J., et al. (2019). Reverse gene-environment interaction approach to identify variants influencing body-mass index in humans. Nat Metab 1, 630–642.

17. Fischer-Posovszky, P., Newell, F.S., Wabitsch, M., and Tornqvist, H.E. (2008). Human SGBS cells - a unique tool for studies of human fat cell biology. Obes Facts 1, 184–189.

18. Schmidt, S.F., Larsen, B.D., Loft, A., Nielsen, R., Madsen, J.G.S., and Mandrup, S. (2015). Acute TNF-induced repression of cell identity genes is mediated by NFκB-directed redistribution of cofactors from super-enhancers. Genome Res. 25, 1281–1294.

19. Galhardo, M., Sinkkonen, L., Berninger, P., Lin, J., Sauter, T., and Heinäniemi, M. (2014). Integrated analysis of transcript-level regulation of metabolism reveals disease-relevant nodes of the human metabolic network. Nucleic Acids Res. 42, 1474–1496.

20. Wabitsch, M., Brenner, R., Melzner, I., Braun, M., Möller, P., Heinze, E., Debatin, K., and Hauner, H. (2001). Characterization of a human preadipocyte cell strain with high capacity for adipose differentiation. International Journal of Obesity 25, 8.

21. Cannon, M.E., Duan, Q., Wu, Y., Zeynalzadeh, M., Xu, Z., Kangas, A.J., Soininen, P., Ala-Korpela, M., Civelek, M., Lusis, A.J., et al. (2017). Trans-ancestry Fine Mapping and Molecular Assays Identify Regulatory Variants at the ANGPTL8 HDL-C GWAS Locus. G3 (Bethesda) 7, 3217–3227.

22. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods 10, 1213–1218.

23. Corces, M.R., Trevino, A.E., Hamilton, E.G., Greenside, P.G., Sinnott-Armstrong, N.A., Vesuna, S., Satpathy, A.T., Rubin, A.J., Montine, K.S., Wu, B., et al. (2017). An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. Nat. Methods 14, 959–962.

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

25. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature 409, 860–921.

26. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359.

27. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and and, R.D. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079.

28. Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J. (2004). The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 32, D493-6.

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

30. Orchard, P., Kyono, Y., Hensley, J., Kitzman, J.O., and Parker, S.C.J. (2020). Quantification, Dynamic Visualization, and Validation of Bias in ATAC-Seq Data with ataqv. Cell Syst 10, 298-306.e4.

31. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. 9, R137.

32. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. 12, 996–1006.

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

34. Risso, D., Schwartz, K., Sherlock, G., and Dudoit, S. (2011). GC-content normalization for RNA-Seq data. BMC Bioinformatics 12, 480.

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

36. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell 38, 576–589.

37. McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol. 28, 495–501.

38. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25, 25–29.

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

40. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21.

41. Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 47, D766–D773.

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

43. Mi, H., Muruganujan, A., Ebert, D., Huang, X., and Thomas, P.D. (2019). PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res. 47, D419–D426.

44. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575.

45. Sarjeant, K., and Stephens, J.M. (2012). Adipogenesis. Cold Spring Harb Perspect Biol 4, a008417.

46. Wang, F., and Tong, Q. (2008). Transcription factor PU.1 is expressed in white adipose and inhibits adipocyte differentiation. Am. J. Physiol., Cell Physiol. 295, C213-220.

47. Shen, L., Glowacki, J., and Zhou, S. (2011). Inhibition of adipocytogenesis by canonical WNT signaling in human mesenchymal stem cells. Exp. Cell Res. 317, 1796–1803.

48. Ntambi, J.M., and Miyazaki, M. (2004). Regulation of stearoyl-CoA desaturases and role in metabolism. Prog. Lipid Res. 43, 91–104.

49. Heanue, T.A., Reshef, R., Davis, R.J., Mardon, G., Oliver, G., Tomarev, S., Lassar, A.B., and Tabin, C.J. (1999). Synergistic regulation of vertebrate muscle development by Dach2, Eya2, and Six1, homologs of genes required for Drosophila eye formation. Genes Dev. 13, 3231–3243.

50. Lee, S.H., Kim, J., Ryu, J.Y., Lee, S., Yang, D.K., Jeong, D., Kim, J., Lee, S.-H., Kim, J.M., Hajjar, R.J., et al. (2012). Transcription coactivator Eya2 is a critical regulator of physiological hypertrophy. J. Mol. Cell. Cardiol. 52, 718–726.

51. Li, Z., Qiu, R., Qiu, X., and Tian, T. (2017). EYA2 promotes lung cancer cell proliferation by downregulating the expression of PTEN. Oncotarget 8, 110837–110848.

52. Degner, J.F., Pai, A.A., Pique-Regi, R., Veyrieras, J.-B., Gaffney, D.J., Pickrell, J.K., Leon, S.D., Michelini, K., Lewellen, N., Crawford, G.E., et al. (2012). DNase?I sensitivity QTLs are a major determinant of human expression variation. Nature 482, 390–394.

CHAPTER 5: DISCUSSION

GWAS have identified thousands of loci associated with cardiometabolic traits[1–5], but GWAS do not identify which variants at these loci are functional or the molecular mechanisms of functional variants[6]. It is well established that GWAS variants are overrepresented in transcriptional regulatory elements, which are typically marked by chromatin accessibility, in trait-relevant tissues and cell types[7–9] and that regulatory elements are useful in predicting functional variants that alter transcription[10,11]. Integrating chromatin accessibility with other genomic data types, such as gene expression, chromatin contacts, and TF binding sites, provides further mechanistic insights at GWAS loci. However, the chromatin accessibility profiles of many primary tissues are under-annotated. Chromatin accessibility profiles across multiple individuals and environmental contexts are needed to fully characterize genetic effects on transcription and disease. In this dissertation, I presented chromatin accessibility profiles in 3 subcutaneous adipose tissue samples, 20 liver tissue samples, and replicates of preadipocyte and adipocyte cells from the SGBS adipocyte cell model. I identified accessible chromatin regions that differ by genotype in liver tissue and that differ across states of adipocyte differentiation. In all cell and tissue types, I identified candidate functional variants found within accessible chromatin at cardiometabolic GWAS loci. I integrated chromatin accessibility data with additional genomic datasets to identify target genes and disrupted TF motifs at GWAS loci. The work in this dissertation contributes to the regulatory characterization of cardiometabolic-relevant tissues and identifies potential molecular mechanisms at GWAS loci.

Whole adipose tissue and SGBS cells both have benefits and drawbacks for studying adipose chromatin accessibility. ATAC peaks in adipose tissue may better reflect in vivo chromatin structure and thus may sometimes be more useful in identifying disease-relevant regulatory elements compared to ATAC peaks from cell models. For example, GWAS variants for insulin traits were enriched in adipose tissue peaks but not SGBS adipocyte or preadipocyte peaks. While GWAS variants for waist-hip ratio were significantly enriched in both adipose tissue and SGBS peaks, they were much more strongly enriched in tissue peaks. However, we found that generating consistent, high-quality ATAC data in SGBS cells is easier than in tissue. We identified more ATAC peaks and generally higher signal-to-noise in SGBS cells, likely due to the homogenous cell content and controlled environment of cell models compared to tissue. ATAC peaks in SGBS cells also originate from a single cell state: either preadipocytes or adipocytes depending on differentiation state. However, some peaks in SGBS may reflect the effects of cells growing in a dish and may not be relevant to in vivo conditions. ATAC peaks in adipose tissue could originate from one or more cell types within heterogeneous adipose tissue. Identifying ATAC peaks present in both adipose tissue and SGBS cells overcomes some of the drawbacks of the two approaches and may identify regulatory elements present in adipocytes or preadipocytes within tissue.

Through working with wet-lab biologists during my graduate research, I learned how difficult it is to generate chromatin accessibility profiles in frozen adipose tissue. We hoped to use ATAC-seq to map chromatin accessibility in 400 individuals. However, we could not reliably generate high-quality data. Wet-lab biologists in the lab optimized multiple assay parameters and tracked various experimental indicators. I processed the resulting sequencing data and assessed data quality. However, we were unable to identify any experimental metrics that would predict data quality prior to sequencing. We suspect that the high lipid content of adipocytes and/or the tissue freezing protocol may disrupt chromatin structure or interfere with the ATAC protocol. Future improvements to the ATAC protocol, new chromatin accessibility assays, or the use of fresh, rather than frozen, tissue may help generation of more consistent, high quality chromatin accessibility profiles in adipose tissue.

We were more successful in profiling chromatin accessibility using frozen liver tissue compared to adipose tissue. We identified more ATAC peaks and observed higher signal-to-noise in liver compared to adipose. ATAC-seq may work better in liver because liver typically has a lower lipid content than adipose tissue. Comparison of ATAC-seq data between liver samples with high vs. low lipid content would help determine if lipid content negatively impacts the ATAC protocol. Differences in tissue extraction, handling, and storage could also contribute to differences in ATAC-seq quality. Importantly, the increased success of ATAC-seq in liver tissue allowed us to generate chromatin profiles in enough individuals to test for genetic differences in chromatin accessibility.

The caQTL I presented in CHAPTER 3 provided mechanistic insight at GWAS loci, similar to findings in previous studies[12–16]. Compared to simple location of GWAS proxy variants within accessible chromatin regions, colocalization of GWAS and caQTL signals provides stronger evidence that GWAS variants may alter regulatory element activity. Colocalization of caQTL, GWAS, and eQTL signals provides even more insight by identifying the putative gene/s targeted by the altered regulatory element. caQTL are just associations however, and additional experiments are needed to prove that caQTL variants alter chromatin accessibility in vivo. Additionally, colocalization is just a measure of sharing between two association signals and it does not indicate whether a variant mediates its effect on one of the traits through the other trait. Statistical mediation tests can help infer causal directions at colocalized loci, such as a variant mediating effect on gene expression through chromatin accessibility changes, but wet-lab experiments are truly needed to demonstrate causality. For example, CRISPR-Cas9 could be used to inactivate a regulatory element and the gene expression of the linked gene could be measured to determine if the regulatory element regulates the gene[17]. Mapping caQTL followed by functional experiments is a promising approach for identifying variants that may influence gene expression and disease by altering chromatin accessibility.

We identified 3,123 significant liver caQTL despite a modest sample size of 20 individuals. The reasonable number of caQTL is partly due to the high sequencing depth and signal-to-noise of the liver ATAC libraries. Compared to mapping QTL using simple linear regression on transformed counts, the RASQUAL model has increased detection power through including allelic imbalance and through modeling count data directly[13,18,19]. Whether more caQTL can be detected compared to eQTL at a similar sample size remains to be determined. We did not investigate this in our study given our modest sample size. A recent study from Alasoo et al.[14] mapped caQTL and eQTL in multiple immune cell contexts and consistently identified more caQTL (~11,000-20,000 across cell types) compared to eQTL (~2,500-3,000 across cell types) despite smaller caQTL sample sizes. Alasoo et al. used the same statistical model and multiple testing correction method for caQTL and eQTL, but they tested for association of variants within a smaller window for caQTL (within 50kb of peak edges) compared to eQTL (within 500kb of gene bodies). We found that decreasing the window size for tested variants from 100kb to 1kb increased the number of identified caQTL, likely due to a reduced multiple testing burden. Therefore, the smaller window size for caQTL compared to eQTL in Alasoo et al. could partly explain the increased number of caQTL. Another technical explanation for this is that many more peaks (n=296,220) than genes (n=15,797) were used for QTL mapping; a higher percent of tested genes had an eQTL compared to the percent of peaks with a caQTL. Another explanation for this is that caQTL may have higher effect sizes than eQTL. However, Keele et al.[20] mapped caQTL and eQTL in three mouse tissues and found that caQTL effect sizes were generally lower than eQTL effect sizes. In contrast to Alasoo et al.[14], Keele et al.[20] identified a smaller number of caQTL relative to eQTL using the same sample size for both analyses and a more similar number of tested peaks (~11,000-24,000 across tissues) and genes (~8,000-11,000 across tissues) compared to Alasoo et al. Keele et al. used the same statistical model, multiple testing correction procedure, and variant window sizes for caQTL and eQTL, which makes the comparison of caQTL and eQTL number more straightforward than in Alasoo et al. However, the sample sizes in both studies were not particularly large (eQTL n=84 and maximum caQTL n=42 for Alasoo et al. and eQTL n=47 and caQTL n=47 for Keele et al.). Larger studies that jointly map eQTL and caQTL in the same samples will help determine if the effect sizes of caQTL and eQTL differ from each other.

Although caQTL can be mapped in small sample sizes, mapping caQTL with more samples has numerous benefits. The power to detect low frequency variants will increase with sample size. We found that some peaks, such as peak9372 at the *SORT1* locus, vary dramatically by genotype, with the less accessible homozygote not exhibiting a peak. Therefore, larger sample sizes may allow identification of peaks that vary strongly by genotype that are absent in the 20 samples we analyzed. Inclusion of diverse samples may allow identification of caQTL that have different effects based on sex, genetic ancestry, age, or other characteristics.

We suspect that the ATAC peak calling strategy impacts caQTL discovery. Unlike genomic locations of genes, chromatin accessibility regions are not well-defined and must be inferred from peak calling algorithms. Consequently, the sizes and boundaries of chromatin accessibility regions are hard to determine and multiple nearby regions can be called as one large peak. Through manual inspection on the UCSC genome browser, we identified peaks that appeared to vary by genotype in part, but not all, of the peak. Some of these peaks were not classified as caQTL, potentially because the change in one part of the peak was masked by the static region. If peak sizes are made too small however, then there may not be enough ATAC-seq counts in the region to detect caQTL. We also identified large peaks that showed strong differences by genotype across the entire peak. Consequently, refining peak calling is not as simple as breaking peaks into smaller regions. Further research is needed to determine the optimal peak calling strategy for caQTL discovery.

In CHAPTER 3, we found that SGBS cells are a particularly useful resource for identifying GWAS variants that may have context-dependent roles on gene regulation. In addition to mapping differences in chromatin accessibility and gene expression across differentiation state, we used SGBS cells to identify context-dependent and allelic effects on transcription using reporter assays. The combined evidence from genomic experiments and reporter assays helps prioritize variants that have functional roles in gene regulation, although further experiments are needed to prove function. Moving forward, CRISPR-cas9[17] could be used to inactivate regulatory elements in SGBS cells, or another adipocyte model, to test for effects on gene regulation in vivo. We could also use SGBS cells to measure changes in disease-relevant cellular phenotypes, such as insulin resistance, in response to altered regulatory element activity. Although we found that SGBS chromatin accessibility does not completely mirror that of adipose tissue in CHAPTER 2, SGBS cells are still valuable to study how GWAS variants impact gene regulation and disease-relevant cellular phenotypes in adipocytes.

Mapping chromatin accessibility and gene expression in additional disease-relevant contexts and across additional individuals may identify additional context-dependent genetic effects. Gene regulation in adipocytes is altered by various stimuli relevant to metabolic disease, such as high insulin, inflammation, and hypoxia[21,22]. Context-dependent genetic associations with chromatin accessibility have also been identified using induced pluripotent stem cells (iPSCs) derived from multiple donors[23]. Therefore, iPSCs could be a useful way to study both genetic and environmental effects on adipose function. Single nuclei ATAC-seq could also be used to help identify GWAS variants that influence cell type-specific regulatory elements in adipose tissue, as has been done in other tissues[24]. All of these strategies will be useful in identifying context-dependent effects of GWAS variation.

The work I presented in this dissertation contributes to the understanding of how disease-associated genetic variants influence gene regulation in cardiometabolic-relevant tissues. I identified candidate functional variants, target regulatory elements, and target genes that can be tested for causal relationships in future experiments. Identifying genes and regulatory elements that cause disease may lead to therapeutic strategies. Therapies could be designed that target individual genes or networks of genes involved in similar biological pathways[25]. In addition, regulatory elements could be targeted for therapy using epigenome editing[26]. Context-dependent genetic effects can be used to identify the precise cell type or cell context in which a therapy may be most effective. Future large-scale genetic, epigenomic, transcriptomic, and functional studies in diverse participants will hopefully lead to a detailed understanding of the molecular mechanisms underlying cardiometabolic diseases and lead to effective pharmaceutical therapies and public health initiatives.

REFERENCES

1. Mahajan, A., Taliun, D., Thurner, M., Robertson, N.R., Torres, J.M., Rayner, N.W., Payne, A.J., Steinthorsdottir, V., Scott, R.A., Grarup, N., et al. (2018). Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nat. Genet. *50*, 1505–1513.

2. Evangelou, E., Warren, H.R., Mosen-Ansorena, D., Mifsud, B., Pazoki, R., Gao, H., Ntritsos, G., Dimou, N., Cabrera, C.P., Karaman, I., et al. (2018). Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. Nat. Genet. *50*, 1412–1425.

3. Yengo, L., Sidorenko, J., Kemper, K.E., Zheng, Z., Wood, A.R., Weedon, M.N., Frayling, T.M., Hirschhorn, J., Yang, J., Visscher, P.M., et al. (2018). Meta-analysis of genome-wide association studies for height and body mass index in ∼700000 individuals of European ancestry. Hum. Mol. Genet. *27*, 3641–3649.

4. Klarin, D., Damrauer, S.M., Cho, K., Sun, Y.V., Teslovich, T.M., Honerlaw, J., Gagnon, D.R., DuVall, S.L., Li, J., Peloso, G.M., et al. (2018). Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. Nat. Genet. *50*, 1514–1523.

5. Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. *47*, D1005–D1012.

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

7. Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190–1195.

8. Absher, Devin (2012). An Integrated Encyclopedia of DNA Elements in the Human Genome. Nature *489*, 57–74.

9. Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317.

10. Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N.E., Ahfeldt, T., Sachs, K.V., Li, X., Li, H., Kuperwasser, N., Ruda, V.M., et al. (2010). From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature *466*, 714–719.

11. Fogarty, M.P., Cannon, M.E., Vadlamudi, S., Gaulton, K.J., and Mohlke, K.L. (2014). Identification of a regulatory variant that binds FOXA1 and FOXA2 at the CDC123/CAMK1D type 2 diabetes GWAS locus. PLoS Genet. *10*, e1004633.

12. Degner, J.F., Pai, A.A., Pique-Regi, R., Veyrieras, J.-B., Gaffney, D.J., Pickrell, J.K., Leon, S.D., Michelini, K., Lewellen, N., Crawford, G.E., et al. (2012). DNase?I sensitivity QTLs are a major determinant of human expression variation. Nature *482*, 390–394.

13. Kumasaka, N., Knights, A.J., and Gaffney, D.J. (2016). Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. Nat. Genet. *48*, 206–213.

14. Alasoo, K., Rodrigues, J., Mukhopadhyay, S., Knights, A.J., Mann, A.L., Kundu, K., Hale, and C., Dougan, G., and Gaffney, D.J. (2018). Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. Nat. Genet.

15. Calderon, D., Nguyen, M.L.T., Mezger, A., Kathiria, A., Müller, F., Nguyen, V., Lescano, N., Wu, B., Trombetta, J., Ribado, J.V., et al. (2019). Landscape of stimulation-responsive chromatin across diverse human immune cells. Nat. Genet. *51*, 1494–1505.

16. Gate, R.E., Cheng, C.S., Aiden, A.P., Siba, A., Tabaka, M., Lituiev, D., Machol, I., Gordon, M.G., Subramaniam, M., Shamim, M., et al. (2018). Genetic determinants of co-accessible chromatin regions in activated T cells across humans. Nature Genetics *50*, 1140–1150.

17. Pulecio, J., Verma, N., Mejía-Ramírez, E., Huangfu, D., and Raya, A. (2017). CRISPR/Cas9-based engineering of the epigenome. Cell Stem Cell *21*, 431–447.

18. Sun, W. (2012). A statistical framework for eQTL mapping using RNA-seq data. Biometrics *68*, 1–11.

19. Geijn, B. van de, McVicker, G., Gilad, Y., and Pritchard, J.K. (2015). WASP: allele-specific software for robust molecular quantitative trait locus discovery. Nat. Methods *12*, 1061–1063.

20. Keele, G.R., Quach, B.C., Israel, J.W., Chappell, G.A., Lewis, L., Safi, A., Simon, J.M., Cotney, P., Crawford, G.E., Valdar, W., et al. (2020). Integrative QTL analysis of gene expression and chromatin accessibility identifies multi-tissue patterns of genetic regulation. PLoS Genet. *16*, e1008537.

21. Lo, K.A., Labadorf, A., Kennedy, N.J., Han, M.S., Yap, Y.S., Matthews, B., Xin, X., Sun, L., Davis, R.J., Lodish, H.F., et al. (2013). Analysis of in vitro insulin-resistance models and their physiological relevance to in vivo diet-induced adipose insulin resistance. Cell Rep *5*, 259–270.

22. Schmidt, S.F., Larsen, B.D., Loft, A., Nielsen, R., Madsen, J.G.S., and Mandrup, S. (2015). Acute TNF-induced repression of cell identity genes is mediated by NFκB-directed redistribution of cofactors from super-enhancers. Genome Res. *25*, 1281–1294.

23. Banovich, N.E., Li, Y.I., Raj, A., Ward, M.C., Greenside, P., Calderon, D., Tung, P.Y., Burnett, J.E., Myrthil, M., Thomas, S.M., et al. (2018). Impact of regulatory variation across human iPSCs and differentiated cells. Genome Res. *28*, 122–131.

24. Rai, V., Quang, D.X., Erdos, M.R., Cusanovich, D.A., Daza, R.M., Narisu, N., Zou, L.S., Didion, J.P., Guan, Y., Shendure, J., et al. (2020). 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.

25. Shu, L., Blencowe, M., and Yang, X. (2018). Translating GWAS Findings to Novel Therapeutic Targets for Coronary Artery Disease. Front Cardiovasc Med *5*, 56.

26. Xie, N., Zhou, Y., Sun, Q., and Tang, B. (2018). Novel Epigenetic Techniques Provided by the CRISPR/Cas9 System. Stem Cells Int *2018*, 7834175.