Title

Identification of Context-Specific Genomic Regulatory Mechanisms Contributing to Human Cardiometabolic Disorders

Author

Garske, Kristina Marie

Publication Date

2021

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Identification of Context-Specific Genomic Regulatory Mechanisms Contributing to Human Cardiometabolic Disorders

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Human Genetics

by

Kristina Marie Garske

2021

ABSTRACT OF THE DISSERTATION

Identification of Context-Specific Genomic Regulatory Mechanisms Contributing to Human Cardiometabolic Disorders

by

Kristina Marie Garske

Doctor of Philosophy in Human Genetics

University of California, Los Angeles, 2021

Professor Päivi Elisabeth Pajukanta, Chair

Obesity is an important driver of many cardiometabolic disorders (CMDs), including type 2 diabetes (T2D) and multiple types of dyslipidemias. Genome-wide association studies (GWAS) have identified genetic variants associated with body mass index (BMI) in humans, but the genes and genomic regulatory mechanisms underlying these associations are not well-understood. Furthermore, gene-environment interactions (GxEs) likely play a role in driving the population variance in BMI, as well as downstream comorbidities. Understanding the biology underlying genetic associations requires studying the relevant cell- and tissue-types, as variants will likely only function in cell-types in which they regulate gene expression. However, determining in which cell-types and under what environmental conditions the variants function is not a trivial task. Functional fine-mapping of individual loci is laborious and time-intensive. This issue has produced a bottleneck in understanding the hundreds of loci that have been associated with CMDs. One way to prioritize strong candidates for future characterization of gene function

is to first fine-map the cell- and context-specific genomic regulatory mechanisms through which variants function, thereby providing evidence of the variant having a role in a given context. Adipose tissue is the main site of fat storage in the body, and is thus highly responsive to the obesogenic environment. It must expand to accommodate excess nutrients, and it produces signaling molecules that regulate food intake and energy expenditure. Thus, genetic variants that regulate these and other important adipose tissue processes likely play a role in the etiology and pathophysiology of obesity.

The projects in this dissertation were designed with the goal of identifying genes and genomic regulatory mechanisms underlying the genetic risk for obesity and related comorbidities. I studied the key adipose tissue cell-types that are responsible for energy homeostasis: adipocytes, the energy-storing fat cells, and their progenitors, preadipocytes. In the second chapter, we fine-mapped BMI GWAS variants to those that are likely functioning in adipocytes. We performed promoter Capture Hi-C in human primary adipocytes to identify the physical interactions between gene promoters and their regulatory elements. We then linked the BMI GWAS variants to adipose gene expression by identifying *cis*-eQTLs in 335 subcutaneous adipose tissue biopsies from the METabolic Syndrome In Men (METSIM) cohort. By screening the adipocyte-specific promoter-interacting regions for these variants, we identified four examples of BMI GWAS variants and 38 additional candidate genes that likely function in adipocytes via promoter interactions to affect BMI in humans.

Work in the third chapter was motivated by the hypothesis that GxEs affect adipose tissue expansion and thus contribute to variation in BMI. Given the known correlation of high dietary saturated fat intake with BMI, I sought to extend this knowledge and demonstrate that causal genomic regulatory mechanisms in response to saturated fat intake in adipocytes affect variation

in BMI. Genetic variants exhibiting significant interaction effects are difficult to detect in humans, for reasons including the multiple-testing burden for genome-wide GxE scans and the heterogeneity of our environments. To circumvent these issues, I first performed a controlled, *in vitro* treatment of human primary adipocytes with dietary saturated or monounsaturated fatty acids. I quantitatively assessed the genomic responses to this lipid challenge via changes in the accessible chromatin landscape in the adipocytes. Only the genetic variants that landed in the lipid-responsive genomic regulatory elements were selected for a GxE scan in the large UK Biobank (UKB) cohort. By prioritizing and restricting the GxE search space in this way, we identified 38 significant GxE variants that exhibit adipocyte-origin genomic regulatory mechanisms responding to dietary saturated fat to affect BMI in humans.

It is important to address not just the etiology of obesity, but its pathophysiological mechanisms as well. A pro-inflammatory environment develops in the adipose tissue in obesity, which is then thought to contribute to systemic low-grade inflammation and downstream obesity comorbidities. In chapter four, I leveraged a BMI-discordant monozygotic (MZ) twin cohort (ΔBMI$\geq$3 kg/m^2) to determine the contribution of preadipocyte genomic dysregulation to systemic inflammation in humans. Preadipocytes can mount an immune response to the pro-inflammatory signals from adipose tissue macrophages, at the expense of differentiating into adipocytes. However, the mechanisms underlying this dysregulation are not fully known. Furthermore, it is not clear whether the preadipocyte mechanisms are causal for inflammation, or simply reactive to the environment. By studying the chromatin accessibility and gene expression profiles in the MZ twin PAd, I showed that increased BMI alters the higher-order genomic programming of PAd. The reprogrammed regions exhibit a stronger accumulation of low *p*-value GxE variants that interact with BMI to affect the inflammatory marker C-reactive protein (CRP)

in the UKB, thus providing evidence of PAd-origin genetic and genomic mechanisms contributing to systemic inflammation in humans.

In summary, by integrating various levels of epigenomic, transcriptomic, and genetic information across multiple cohorts with deep phenotyping, we have prioritized genes and genomic regulatory mechanisms in human adipose tissue cell-types that are important for obesity and obesity pathophysiology. The improved molecular understanding of genetic causes and GxEs underlying obesity and its comorbidities will inform prevention and treatment methods in precision medicine, toward reducing cardiovascular disease risk in humans.

The dissertation of Kristina Marie Garske is approved.

Jason Ernst

Bogdan Pasaniuc

Janet S. Sinsheimer

Päivi Elisabeth Pajukanta, Committee Chair

University of California, Los Angeles

2021

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

Firstly, I would like to thank my advisor, Professor Päivi Pajukanta, for providing guidance and support through an environment that allowed me to grow personally and as a scientist. I have learned so much from you and appreciate you immensely. Your unwavering excitement and passion for your science always inspired and encouraged me to keep moving forward.

Thank you to my committee members, Professors Jason Ernst, Bogdan Pasaniuc, and Janet S. Sinsheimer, for providing insight and guidance throughout my dissertation work. In particular, to Professor Janet S. Sinsheimer, thank you for your close collaboration and for being so generous with your time to help with the analytical approaches in all of my major projects. Thank you also for your mentorship and support for my professional development over these past 6 years.

To the Pajukanta lab current and past group members, thank you all for your collaboration, support, and friendship. Thank you to Elina Nikkola, for welcoming me into the lab and being such a good friend throughout your time at UCLA. To David Pan and Zong Miao, I learned a ton from you both, and was always so appreciative of your willingness to collaborate in a heartbeat. Marcus Alvarez and Jihane Benhammou, your passion and dedication for your work are truly inspiring. It has been so fun to get to know you both in and outside of the lab throughout my time at UCLA, and I look forward to continuing our friendship beyond grad school.

Mentoring throughout my graduate career has been such a pleasure, and I have had the privilege of working with very talented UCLA undergraduate students. Elliot Kim, Yash Bhagat, Caroline Comenho, Sandhya Rajkumar, and Amogha Koka, thank you for all of the time and energy you have dedicated to these projects. It is so rewarding to watch mentees grow, and I am

proud of all of your work and accomplishments. Yash, you provided input on all of my projects, encouraged me when I was stumped, and have so much passion and excitement for the research. It has truly been a pleasure to get to know you and I look forward to our continued friendship. Caroline, you started off as my mentee, and our relationship has blossomed into one of my most cherished friendships. I am so proud of you and look forward to seeing you succeed in grad school and beyond.

I would also like thank the broader Human Genetics and Computational Medicine community at UCLA. In particular, I would like to thank Jake Lusis, Karen Reue, Janet S. Sinsheimer, and all other members of the Lusis NIH Program Project Grant. Your input and advice have been so important in contributing to the quality and robustness of the projects in this dissertation. To Jake and Karen, thank you also for your mentorship during my TAships for the Human Genetics & Genomics course, and for your support for my professional development throughout my PhD career.

My deepest and warmest thanks go out to my family and friends. To my parents, Wayne and Marla Garske, who have supported me in such a way that I can take risks, make mistakes, and grow. They were there for me throughout my PhD, celebrating the wins and helping me get through the tougher times. My grandmother, Judy Turner, has been my rock since childhood. Open-minded, loving, and brilliant, she never stops learning and growing and I find so much inspiration in that.

This dissertation was partially supported by the NIH-NHLBI grant F31HL142180. The research has been conducted using the UK Biobank Resource under Application Number 33934. Thank you to the UNGC sequencing core for performing the RNA and DNA sequencing.

Chapter 2 was originally published in *Nature Communications* in 2018 with the title "Integration of human adipocyte chromosomal interactions with adipose gene expression prioritizes obesity-related genes from GWAS" by David Z. Pan, Kristina M. Garske, Marcus Alvarez, Yash V. Bhagat, James Boocock, Elina Nikkola, Zong Miao, Chelsea K. Raulerson, Rita M. Cantor, Mete Civelek, Craig A. Glastonbury, Kerrin S. Small, Michael Boehnke, Aldons J. Lusis, Janet S. Sinsheimer, Karen L. Mohlke, Markku Laakso, Päivi Pajukanta and Arthur Ko (Volume 9, p. 1512). As the co-first author, I designed the study and wrote the first draft of the manuscript along with David Z. Pan, Päivi Pajukanta, and Arthur Ko. I performed the library preparation for the human primary adipocyte promoter Capture Hi-C data (pCHi-C). Yash V. Bhagat performed the electrophoretic mobility shift assays (EMSAs). Myself, David Z. Pan, Marcus Alvarez, Zong Miao, James Boocock, Arthur Ko, Janet S. Sinsheimer, Päivi Pajukanta and Rita M. Cantor designed the analytical approach and performed the various computational analyses. Markku Laakso performed the METabolic Syndrome In Men (METSIM) cohort phenotyping. Arthur Ko, Elina Nikkola, Marcus Alvarez, Karen L. Mohlke, Chelsea Raulerson, and Päivi Pajukanta performed the RNA-seq, quality control and data cleaning of the METSIM adipose samples used for the *cis*-eQTL identification and gene expression correlation with BMI. Mete Civelek, Aldons J. Lusis, Elina Nikkola, Michael Boehnke, Karen L. Mohlke, and Markku Laakso are our METSIM collaborators responsible for the data collection and genotyping. Craig A. Glastonbury and Kerrin S. Small performed the replication analysis of the gene expression correlation with BMI in the Twins UK cohort.

Chapter 3 was originally published in *Nature Metabolism* in 2019 with the title "Reverse gene-environment interaction approach to identify variants influencing body-mass index in humans" by Kristina M. Garske, David Z. Pan, Zong Miao, Yash V. Bhagat, Caroline Comenho,

Christopher R. Robles, Jihane N. Benhammou, Marcus Alvarez, Arthur Ko, Chun Jimmie Ye, Joseph R. Pisegna, Karen L. Mohlke, Janet S. Sinsheimer, Markku Laakso and Päivi Pajukanta (Volume 1, p 630-642). I was the primary author of this work and performed the human primary adipocyte lipid challenge, collected the ATAC-seq data and performed most computational analyses. I interpreted and synthesized the results, constructed the figures, and wrote the first draft of the manuscript, with Päivi Pajukanta. Myself, David Z. Pan, Zong Miao, Joeseph R. Pisegna, Chun Jimmie Ye, Janet S. Sinsheimer, and Päivi Pajukanta designed the analytical and statistical approaches. David Z. Pan, Zong Miao, Marcus Alvarez, and Christopher R. Robles contributed to the computational analyses. Yash V. Bhagat, Marcus Alvarez, Caroline Comenho, and Jihane N. Benhammou performed the EMSAs and the oil red o staining and quantification in the human primary preadipocytes and adipocytes. Karen L. Mohlke and Markku Laakso are our METSIM collaborators responsible for the data collection and genotyping, along with Päivi Pajukanta. Arthur Ko performed the RNA-seq, quality control and data cleaning of the METSIM adipose samples used for the *cis*-eQTL identification.

Chapter 4 has been submitted for publication with the title "BMI-discordant monozygotic twin preadipocytes exhibit altered compartmentalization in regions with BMI-interacting variants affecting inflammation" by Kristina M. Garske, Caroline Comenho, Brunilda Baillu, David Z. Pan, Yash V. Bhagat, Gregory Rosenberg, Amogha Koka, Zong Miao, Janet S. Sinsheimer, Kirsi H. Pietiläinen and Päivi Pajukanta. I was the primary author of this work, performed all major computational analyses, and performed or supervised the RNA-seq and Assay for Transposase Accessible Chromatin (ATAC) -sequencing data collection in the body mass index (BMI) -discordant monozygotic (MZ) twin preadipocytes and differentiating preadipocytes. Yash V. Bhagat and Gregory Rosenberg contributed to this data collection. Caroline Comenho performed

the EMSAs. Brunilda Baillu and David Z. Pan contributed to the computational analyses. Zong Miao, Janet S. Sinsheimer, Päivi Pajukanta, Brunilda Baillu and I designed the analytical and statistical approaches. I interpreted and synthesized the results, constructed the figures, and wrote the first draft of the manuscript, with Päivi Pajukanta.

Finally, thank you to all of the individuals who gave their time and biological samples to this research. Without them, none of these discoveries would have been possible.

VITA

EDUCATION

Sept 2014 – June 2021	**University of California, Los Angeles, CA** Doctor of Philosophy in Genetics & Genomics
June 2011 – June 2012	**University of California, San Diego, CA** Master of Science in Biology
Sept 2008 – August 2011	**University of California, San Diego, CA** Bachelor of Science in Biochemistry & Cell Biology

HONORS AND AWARDS

2012	Excellence in Teaching Award, Graduate TA, Genetics, UCSD
2013	Travel Award, Weinstein Cardiovascular Conference
2017	Professor Nomination, Distinguished Teaching Award for Graduate Teaching Assistants, Human Genetics & Genomics, UCLA
2017	Reviewer's Choice Abstract Award, American Society of Human Genetics Meeting
2018	Reviewer's Choice Abstract Award, American Society of Human Genetics Meeting
2020	Reviewer's Choice Abstract Award and Selected Poster Talk, American Society of Human Genetics Virtual Meeting

PUBLICATIONS

Yocheved L. Schindler, **Kristina M. Garske**, Jinhu Wang, Beth A. Firulli, Anthony B. Firulli, Kenneth D. Poss, Deborah Yelon. Hand2 elevates cardiomyocyte production during zebrafish heart development and regeneration. *Development* **141,** 3112-3122 (2014).

Elina Nikkola, Arthur Ko, Marcus Alvarez, Rita M. Cantor, **Kristina M. Garske**, Elliot Kim, Stephanie Gee, Alejandra Rodriguez, Reinhard Muxel, Niina Matikainen, Sanni Söderlund, Mahdi M. Motazacker, Jan Borén, Claudia Lamina, Florian Kronenberg, Wolfgang J. Schneider, Aarno Palotie, Markku Laakso, Marja-Riitta Taskinen, Päivi Pajukanta. Family-specific aggregation of lipid GWAS variants confers the susceptibility to familial hypercholesterolemia in a large Austrian family. *Atherosclerosis* **264**, 58-66 (2017).

Malika Kumar Freund, Kathryn S. Burch, Huwenbo Shi, Nicholas Mancuso, Gleb Kichaev, **Kristina M. Garske**, David Z. Pan, Zong Miao, Karen L. Mohlke, Markku Laakso, Päivi Pajukanta, Bogdan Pasaniuc, Valerie A. Arboleda. Phenotype-specific enrichment of Mendelian disorder genes near GWAS regions across 62 complex traits. *American Journal of Human Genetics* **103**, 535-552 (2018).

David Z. Pan*, **Kristina M. Garske***, Marcus Alvarez, Yash V. Bhagat, James Boocock, Elina Nikkola, Chelsea K. Raulerson, Rita M. Cantor, Mete Civelek, Craig A. Glastonbury, Kerrin S. Small, Michael Boehnke, Aldons J. Lusis, Janet S. Sinsheimer, Karen L. Mohlke, Markku Laakso, Päivi Pajukanta, Arthur Ko. Integration of human adipocyte chromosomal interactions

with local adipose gene expression identifies obesity genes beyond GWAS. *Nature Communications* **9**, 1512 (2018). *Authors contributed equally

Kristina M. Garske, David Z. Pan, Zong Miao, Yash V. Bhagat, Caroline Comenho, Christopher R. Robles, Jihane N. Benhammou, Marcus Alvarez, Arthur Ko, Chun Jimmie Ye, Joseph R. Pisegna, Karen L. Mohlke, Janet S. Sinsheimer, Markku Laakso, Päivi Pajukanta. Reverse gene–environment interaction approach to identify variants influencing body-mass index in humans. *Nature Metabolism* **1**, 630-642 (2019).

Jihane N. Benhammou, Arthur Ko, Marcus Alvarez, Minna U. Kaikkonen, Carl Rankin, **Kristina M. Garske**, David Padua, Yash Bhagat, Dorota Kaminska, Vesa Kärjä, Jussi Pihlajamäki, Joseph R. Pisegna and Päivi Pajukanta. Novel lipid lincRNA OLMALINC regulates the liver steatosis gene, SCD, as an enhancer RNA. *Hepatology Communications* **3**, 1356-1372 (2019).

Brandon Jew*, Marcus Alvarez*, Elior Rahmani, Zong Miao, Arthur Ko, **Kristina M. Garske**, Jae Hoon Sul, Kirsi H. Pietiläinen, Päivi Pajukanta**, Eran Halperin**. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nature Communications*, **11**, 1 (2020). *Authors contributed equally; **Co-corresponding authors

Marcus Alvarez*, Elior Rahmani*, Brandon Jew, **Kristina M. Garske**, Zong Miao, Jihane N. Benhammou, Chun Jimmie Ye, Joseph R. Pisegna, Kirsi H. Pietiläinen, Eran Halperin, Päivi Pajukanta. Enhancing droplet-based single-nucleus RNA-seq resolution using the semi-supervised machine learning classifier DIEM. *Scientific Reports*, **10**, 11019 (2020). *Authors contributed equally

Liza D. Morales, Douglas T. Cromack, Devjit Tripathy, Marcel Fourcaudot, Satish Kumar, Joanne E. Curran, Melanie Carless, Harald H. H. Göring, Shirley L. Hu, Juan Carlos Lopez Alvarenga, **Kristina M. Garske**, Päivi Pajukanta, Kerrin Small, Craig Glastonbury, Swapan K. Das, Carl Langefeld, Robert Hanson, Wen-Chi Shueh, Luke Norton, Rector Arya, Srinivas Mummidi, John Blangero, Ralph A. DeFronzo, Ravindranath Duggirala, Christopher P. Jenkinson. Further evidence supporting a potential role for ADH1B in obesity. *Scientific Reports*, **11**, 1932 (2021).

Ilakya Selvarajan, Anu Toropainen*, **Kristina M. Garske***, Maykel López Rodríguez, Arthur Ko, Zong Miao, Dorota Kaminska, Tiit Örd, Aarthi Ravindran, Oscar H Liu, Pierre R Moreau, Ashik Jawahar Deen, Calvin Pan, Anna-Liisa Levonen, Aldons J Lusis, Sami Heikkinen, Casey E Romanoski, Jussi Pihlajamäki, Päivi Pajukanta, Minna U Kaikkonen. Integrative analysis of liver-specific noncoding regulatory variants associated with the risk of coronary artery disease. *American Journal of Human Genetics*, **108**, 411-430 (2021). *Authors contributed equally

Kristina M. Garske, Caroline Comenho, Bruna Baillu, David Z. Pan, Yash V. Bhagat, Gregory Rosenberg, Amogha Koka, Zong Miao, Janet S. Sinsheimer, Kirsi H. Pietiläinen, Päivi Pajukanta. Preadipocytes from BMI-discordant monozygotic twins harbor large genomic loci with altered subnuclear compartmentalization and variants interacting with BMI on inflammation. *Submitted*.

Chapter 1

Introduction and background

1.1 Obesity and cardiometabolic disorders

Obesity is associated with an increased lifetime risk of cardiovascular disease (CVD) and an increased prevalence of type 2 diabetes (T2D), hypertension, and multiple types of dyslipidemias (Virani *et al.*, 2021). The CDC estimates that ~40% of United States adults are obese, with ~70% of adults being at least overweight (Virani *et al.*, 2021). Obesity is defined as having a body mass index (BMI) $\geq$30 kg/m^2, and overweight as 25$\leq$BMI<30. Obesity is a complex trait that develops as a result of genetic and environmental factors, as well as their interactions. Some of the complexity in understanding obesity comes from incomplete knowledge of which factors are causal and which are reactive to the metabolic dysregulation present in the disease state. Various metabolic abnormalities are associated with each other and can affect each other bidirectionally, which can make it difficult to differentiate causation from reaction (Hotamisligil, 2006; Nordestgaard *et al.*, 2007; Tchang, Saunders and Igel, 2021). Both causation and reaction mechanisms contribute to disease, but methods for identifying them and the biological implications of the results may be different. Because our genetics are set at birth, genomic approaches using genetic variants can help provide evidence of causality. This will be discussed in section 1.2 below. On the other hand, the effects of obesity on the development of other CMDs and some cancers (Fock and Khoo, 2013) clearly highlight the need for understanding the pathophysiological mechanisms that occur in obesity to drive these comorbidities, including the body's response to the obese state. This will be discussed in section 1.4 below.

The cause of obesity can be mainly attributed to higher energy intake than energy expenditure, leading to the storage of this excess energy as fat in the adipose tissue. Exceptions include hypothyroidism, Cushing's disease, growth hormone deficiency, and some medications, among others (Fock and Khoo, 2013). The most commonly prescribed treatment for class I

obesity (30 kg/m^2≤BMI<35 kg/m^2) is a healthy diet and exercise (Fock and Khoo, 2013; Tchang, Saunders and Igel, 2021). However, this regimen is difficult for patients to comply with, in an environment that makes unhealthy food more affordable and accessible than healthy food, combined with the relatively sedentary lifestyles of Western or urban cultures. Insights from genetic studies surrounding the biological mechanisms that drive obesity can aid in the development of treatments, which can be combined with diet and exercise plans to reduce the prevalence of obesity.

Defining obesity as a state of disease can be complicated. The cut point of a BMI≥30 kg/m^2 to define obesity is sensitive to differences in ethnicities and the lean-to-fat mass ratio (Virani *et al.*, 2021). Not everyone with a BMI at or above this cut point will exhibit metabolically unhealthy profiles (e.g., insulin resistance or dyslipidemias) (Blüher, 2020). This may be due to a delay in the development of obesity comorbidities, rather than the complete protection from them (Blüher, 2020). One of the main and initial indicators of poor metabolic health in obesity is chronic low-grade inflammation, which is thought to originate at least in part as a result of pro-inflammatory signals from adipose tissue in response to chronic overnutrition (Hotamisligil, 2006). However, the mechanisms underlying the initial pro-inflammatory state are unknown. Given the close link between inflammation and the development of obesity comorbidities, the identification of drivers of the initial inflammatory state in obesity is of particular importance (Hotamisligil, 2006).

In summary, obesity is a complex trait that develops as a result of both genetics and environment, plus their interactions. It is associated with an increased risk for CMDs and CVD (Virani *et al.*, 2021). The alarming prevalence of obesity and overweight in the U.S. strongly

urges the development of additional treatment and prevention options to reduce obesity rates and related cardiovascular risk.

1.2 Genome-wide association studies and the biological underpinnings of obesity

Estimates for the heritability of BMI range from between 30 and 75% (Virani *et al.*, 2021). In one of the largest genome-wide association studies (GWAS) for BMI to date, 97 genome-wide significant (GWS) variants were deemed to account for only 2.7% of the variance in BMI (Locke *et al.*, 2015). Even when the contributions from all HapMap3 SNPs are assessed, only 31-54% of the heritability of BMI can be explained (Locke *et al.*, 2015).

Despite the small amount of variance explained by individual variants contributing to BMI, polygenic risk scores (PRSs) developed using either just 97 GWS variants (Locke *et al.*, 2015), or the genome-wide summary statistics from the same GWAS from Locke *et al.* (Khera *et al.*, 2019), showed a significant positive correlation between high PRS groups and the true mean BMI of the individuals in that group. Furthermore, Khera *et al.* showed that the individuals in the highest decile of the BMI PRS are at an increased risk of having type 2 diabetes (72% increased risk), hypertension (38% increased risk), or coronary artery disease (28% increased risk). Importantly, the main clinical utility of a PRS lies in the prediction power and the availability of a treatment or prevention plan based on those predictions. The utility of a BMI PRS may currently be limited by the existence of few prevention and treatment options aside from diet and exercise. Thus, fine-mapping of BMI GWAS loci to understand the causal mechanisms underlying obesity and related CMDs is necessary for the development of additional, personalized treatment and prevention plans.

Most GWAS variants are located within non-coding regions of the genome, which suggests altered gene regulation, rather than changes in the amino acid sequence of the protein, as a mechanism largely contributing to the etiology of complex traits (Maurano *et al.*, 2012). In the simplest form of gene prioritization in a GWAS locus, the index gene is the nearest gene to the associated variant, or a candidate gene in the locus with prior knowledge indicating it might be relevant for the associated trait. However, the nearest gene is not always the most likely causal gene at the locus. To fine-map GWAS loci, a number of factors must be considered. First, linkage disequilibrium (LD) among genetic variants occurs as a result of low recombination frequencies between variants that are located close together on the linear genome. This causes them to be inherited together and leads to a correlation in genotypes among the variants in LD. This means a GWAS variant "tags" other SNPs that are in tight LD (usually defined as $R^2>0.8$) with it, making it difficult to uncover which variant(s) is the true causal variant underlying the GWAS signal.

Second, demonstrating the causality of the gene(s) through mechanistic studies is important but time-consuming and laborious. GWAS provide the first clues that some gene or set of genes in an associated locus are important for a trait, yet there remains a lot of work to be done to truly identify and confirm causal mechanisms. Prioritizing one or a very small number of genes for the mechanistic studies at a GWAS locus is thus an important goal.

Third, the highly polygenic nature of complex traits has led to the identification of tens to hundreds of associated loci in well-powered GWAS for different traits, making a locus-by-locus approach to following up GWAS inefficient. There has been a huge effort to define the genome-wide gene regulatory landscapes (functional genomics, i.e., histone modifications and transcription factor binding) and assess the effects of genetic variation on these molecular

signatures across human cell and tissue types (Dunham *et al.*, 2012; Roadmap Epigenomics Consortium *et al.*, 2015). These data provide an essential genome-wide, biological layer of information between a GWAS signal and the downstream mechanistic gene studies. They not only provide clues as to which cell-types may be functioning in disease, but they can also aid in the filtering of many variants in LD with each other, by restricting to only those variants that land in the functional elements. Converging evidence from the integration of various levels of these functional datasets can be a powerful way to identify highly likely causal mechanisms at a locus, which can then be assessed at the gene mechanism level.

In summary, GWAS have revealed that complex traits are affected by genetic variants with very small effects on the population variance in the trait. Most of these variants are implicated in gene regulation, given their enrichment in non-coding, functional elements of the genome (Maurano *et al.*, 2012). Demonstrating the causality of a given variant or set of variants in a trait-associated locus is time-consuming and laborious, and given the number of associations for these traits, fine-mapping on a locus-by-locus manner is prohibitively slow. Instead, functional genomics assays provide a more high-throughput functional fine-mapping approach to identify strong molecular priors for genomic regulatory mechanisms and gene pathways that lead to disease.

1.3 Functional genomics and cellular genomic programming

As GWAS strongly implicate gene regulation in the genetic etiology of CMDs and other complex traits, a burst of technologies that use high-throughput sequencing approaches to assay for functional elements in the genome has dominated recent advancements in genetics research. These functional elements range from local epigenetic signatures that drive or inhibit gene

expression to the higher-order genome structure. The combined effects of these functional elements ensure that genes are properly expressed for cells to perform their specialized functions. This is referred to as cellular genomic programming, and it is hypothesized that many GWAS variants act through altering these mechanisms, discussed in section 1.2 above.

One mechanism that is important for gene regulation is the physical interactions of distal regulatory elements with their target gene promoters (Mifsud *et al.*, 2015; Schoenfelder *et al.*, 2015). To identify these interactions, various forms of chromosome conformation capture (3C) techniques have been developed (Grob and Cavalli, 2018). Hi-C is a variant of 3C that involved the high-throughput sequencing of all interactions at the genome-wide level (Lieberman-Aiden *et al.*, 2009). This is powerful as it provides an agnostic view into the genome-wide organization of the chromatin. Because Hi-C libraries essentially represent the entire genome, it is expensive to sequence deep enough for high resolution and regional enrichment of regulatory sequences. This makes Hi-C better suited for understanding the higher-order regulatory environment at genomic loci, rather than individual gene regulation (Lieberman-Aiden *et al.*, 2009; Schoenfelder *et al.*, 2015). This includes identification of blocks of regions that tend to self-interact, topologically associating domains (TADs) (Dixon *et al.*, 2012; Nora *et al.*, 2012); or the broadly active (A) and inactive (B) compartments that the genome is separated into (Lieberman-Aiden *et al.*, 2009; Rao *et al.*, 2014). While TADs have been shown to be largely invariant across cell-types and even across species at megabase-scale (Dixon *et al.*, 2012), the A/B compartments have been shown to be more cell-type-specific (Dixon *et al.*, 2015). Extensive A/B compartment switching occurs across differentiation and exists between cell-types (Dixon *et al.*, 2015), suggesting that they represent functional regulatory units. However, we cannot reliably predict the expression level of an individual gene solely based on it landing in an A or B compartment, suggesting that A/B

compartments reflect only broad differences in the underlying gene regulatory activity in a locus (Lin *et al.*, 2012; Dixon *et al.*, 2015). Furthermore, current costs of sequencing have precluded the ability perform these experiments in large cohorts, and so we do not know if genetic variants affect this level of genome organization. Characterization of the individual gene regulatory mechanisms is thus still a necessary step for fine-mapping trait-associated genetic loci.

Promoter Capture Hi-C (pCHi-C) is an extension of the Hi-C experiment that uses DNA hybridization and pulldown to enrich the genome-wide interaction data for those interactions that involve promoters, and thus are likely important for local gene regulation (Mifsud *et al.*, 2015; Schoenfelder *et al.*, 2015). Early pCHi-C studies showed that these interactions are enriched for regulatory histone marks relative to non-interacting regions and for the motifs of transcription factors (TFs) that are relevant for the studied cell-type (Mifsud *et al.*, 2015). Furthermore, it was shown that genes that are involved in pCHi-C interactions in a given cell-type are more highly expressed on average in that cell-type than genes not involved in interactions (Schoenfelder *et al.*, 2015). This suggests that promoter interactions identified through pCHi-C represent an important mechanism in local gene regulation, and that they can be used as a way to fine-map genomic regulatory mechanisms at a locus. However, similar to Hi-C data, it is unclear whether genetic variants would affect the pCHi-C interactions themselves, given that the cost of creating and sequencing the libraries has prevented the creation of pCHi-C datasets in large numbers of individuals so far. Instead, it is likely that genetic variants within the interactions affect the gene regulatory element that is being brought into contact with its target promoter via the interactions.

Gene regulatory elements need to be accessible to DNA-binding proteins that modulate gene expression, such as TFs and chromatin modifiers. The major exception is for pioneer factors, which are known to bind inaccessible DNA and recruit chromatin modifying enzymes to

unwind the DNA from histones and make it accessible for downstream transcriptional regulation (Cirillo *et al.*, 2002). In general, however, the accessible chromatin landscape in a cell can be viewed as the collection of putative gene regulatory regions for that cell-type. The Assay for Transposase-Accessible Chromatin (ATAC) -sequencing (Buenrostro *et al.*, 2015) is a method to identify accessible chromatin elements in the genome. It is quick to perform and relatively affordable to sequence, which has contributed to its widespread use across biomedical and basic science research (Yan *et al.*, 2020). Trait-associated genetic variants are enriched in accessible chromatin regions (Maurano *et al.*, 2012), which strongly supports the use of chromatin accessibility assays in profiling a cell-type's gene regulatory landscape as it pertains to disease.

Cellular genomic programming represents the unique combination of epigenetic marks that enable both general and cell-type-specific genomic regulatory mechanisms to function. These mechanisms ensure that genes are expressed at the correct level at the correct time. Two functional genomics marks that are useful in characterizing these genomic regulatory mechanisms are promoter interactions and accessible chromatin. The combination of multiple layers of functional genomics assays will enable a more refined understanding of the cell-type-specific mechanisms that function in genomic loci that are associated with obesity and related comorbidities.

1.4 Adipose tissue and obesity

Obesity develops as a result of increased energy intake relative to energy expenditure. GWAS variants for BMI are strongly implicated as likely functioning in brain regulatory regions (Locke *et al.*, 2015), suggesting that genetic risk for increased BMI acts in large part through neural pathways and behavioral mechanisms. Metabolic pathways that function in adipose tissue, such

as lipid biology and adipogenesis, were also implicated as putative mechanisms functioning in the BMI GWAS loci (Locke *et al.*, 2015). Importantly, crosstalk between the brain and adipose occurs via signaling molecules (adipokines) from the adipose tissue, depending on the metabolic status of the body (Cercato and Fonseca, 2019). These signals help the brain know whether food seeking behavior is needed or not. Furthermore, dysfunctional adipose tissue associated with obesity is implicated in a number of metabolic abnormalities such as altered adipokine signaling, insulin resistance, and inflammation (Ghaben and Scherer, 2019). Thus, altered adipose tissue gene regulation is likely important for driving mechanisms in both increasing BMI and obesity comorbidities.

The cell-type composition of adipose tissue is heterogeneous. Adipocytes are the fat storing cell that must take up fatty acids from the bloodstream and store them as triglycerides in response to feeding. In response to fasting, adipocytes will initiate lipolysis pathways to release fatty acids for energy throughout the body. Adipocytes release adipokines that are responsible for maintaining energy homeostasis throughout the body (Ouchi *et al.*, 2011). For instance, leptin, which is an adipocyte-specific adipokine, is a satiety signal from adipocytes to the brain (Ghaben and Scherer, 2019). Rare variants in leptin are associated with monogenic forms of obesity (Huvenne *et al.*, 2016), highlighting the importance of the adipocyte-brain signaling loop.

Both adipocytes and their progenitors, preadipocytes, secrete pro-inflammatory cytokines and are responsive to inflammatory signals (Constant *et al.*, 2006; Lacasa *et al.*, 2007; Poulain-Godefroy and Froguel, 2007). However, in the mouse 3T3-L1 cell line model of adipogenesis, preadipocytes express a higher basal level of IL-6 than adipocytes (Poulain-Godefroy and Froguel, 2007). Given that IL-6 is one of the main cytokines that stimulates the production of CRP by the liver, this suggests an important role for preadipocytes in the chronic low-grade

inflammation seen in obesity. Furthermore, preadipocytes exhibit impaired differentiation *in vitro* in response to the inflammatory secretome produced by macrophages (Constant *et al.*, 2006; Lacasa *et al.*, 2007). When preadipocytes do not differentiate to make new adipocytes, the adipocytes that already exist must take up and store the excess fat in obesity. This leads to adipocyte hypertrophy, which may affect adipokine secretion and insulin sensitivity (Ghaben and Scherer, 2019), implicating a potential mechanism underlying the obesity predisposition to T2D. This suggests a link between preadipocyte differentiation, inflammation, and their dysregulation in obesity.

Resident macrophages are present in the adipose tissue of both lean and obese individuals, but there seems to be a positive correlation between the number of macrophages and BMI (Ouchi *et al.*, 2011). Furthermore, it is thought that in obesity, these macrophages are activated to a pro-inflammatory state (Ouchi *et al.*, 2011). The initial mechanisms that are involved in this activation are not known. In addition, other pro-inflammatory immune cells involved in both the innate and adaptive immune system infiltrate the inflamed adipose tissue associated with obesity (Ouchi *et al.*, 2011). Other functional cell-types in adipose tissue include vascular cells, fibroblasts, and mesenchymal stem cells (Ouchi *et al.*, 2011), and they also respond to an increase in adiposity. For example, as adipocytes expand in obesity, they become hypoxic, which leads to an increase in profibrotic mechanisms in resident fibroblasts (Ouchi *et al.*, 2011; Ghaben and Scherer, 2019).

Taken together, a balance between the proportions of the many adipose cell-types and the inflammatory status of those cells is disturbed in obesity. Whether and which of these mechanisms are causal for obesity and related comorbidities is not well-understood. Adipocytes and their progenitors must sense the metabolic status of the organism and elicit downstream

signaling and differentiation pathways that are strongly linked to the inflammatory status of the adipose tissue.

1.5 Prioritizing genetic variants for gene-environment interaction scans

Because the cellular environment is always changing within an organism, gene expression needs to be constantly regulated to ensure that the cell can perform its necessary functions at any given time. Genetic variants landing in regions that mediate the genomic responses to environmental stimuli during normal cellular processes could cause the pathway response to be altered, and this is one mechanism that can be causal for a trait. However, while this is likely the underlying mechanism at a number of regulatory GWAS loci, showing causality still necessitates mechanistic studies for the gene. Thus, additional approaches can be used to prioritize regions that are truly causal for, rather than simply reactive to, a trait. To this end, genetic associations can be utilized to provide evidence of causality. By performing a gene-environment interaction (GxE) analysis, we can also test whether genetic variants mediate responses to certain environmental stimuli and contribute to the variance in a trait through those responses.

Unlike traditional GWAS scans, which currently have the power to detect genome-wide significant signals (i.e., significant after correction for multiple testing), genome-wide GxE scans are underpowered at the current human cohort sizes (McAllister *et al.*, 2017). One way to circumvent this issue is to prioritize candidate variants for testing, to reduce the multiple testing burden of a genome-wide scan. Often, variants are selected for GxE testing by identifying those that surround a candidate gene, selected either based on previous GWAS or knowledge of the gene function. However, this approach likely leads to the testing of many variants that have no actual evidence of being responsive to the environmental stimulus of study. Selecting variants

that exhibit evidence of functioning in the genomic signaling pathway in response to the environmental stimulus will help filter out those variants with no functional evidence attributed to them.

Identifying variants with prior functional evidence of being involved in a genomic regulatory response mechanism is becoming more feasible with the continued decreasing cost of sequencing. Quantitative functional genomics assays, such as ATAC-seq, can now be performed in many individuals (Gate *et al.*, 2018) or in different *in vitro* conditions (Ramos-Rodríguez *et al.*, 2019) to show a correlation of an epigenetic mark with a trait or treatment, respectively. Variants landing within the regions that are altered can then be tested for GxEs to support a causal role for that environmental response in driving variance in the outcome trait of interest. This effectively reduces the multiple testing burden present in genome-wide GxE scans. The genes that are regulated by those variants then represent strong candidates for mechanistic studies to understand how our genes interact with our environment to affect risk for obesity and related CMDs.

1.6 Integrating the current state of knowledge to this thesis

The first chapter outlines the motivation and approaches taken in this dissertation for studying the genetic basis of obesity and related comorbidities, with a focus on adipose tissue cell-types. Because our genetics and environment do not exist independently of each other, context is important for understanding the effects of causal genetic variation on obesity and related CMDs, even for marginal genetic associations. At a broad level, context can refer to which cell-type the associated variants are functioning in. However, cells execute different processes at different times, depending on the status of the external cues they receive and integrate in order to function

properly. This means that understanding the biological mechanisms underlying obesity and related CMDs will likely require information on the environmental stimuli that activate or repress these mechanisms. By focusing on environmental stimuli that are known to be strongly associated with the trait outcomes, we can prioritize and fine-map regulatory signals that are likely important for these traits.

In Chapter 2, I begin with an assessment of adipocyte-origin genomic regulation that can be linked to putative causality for increased BMI. We created human primary adipocyte pCHi-C data to interrogate the regulatory promoter interactions in this important cell-type for fat storage and endocrine signaling to the brain. We then identified adipose *cis*-eQTL variants that land within the regions that interact with the *cis*-eQTL target gene, to provide evidence that the gene is likely under genetic control via gene regulatory mechanisms in adipocytes. Correlating the target gene expression with BMI then suggests an important role for this genetically-regulated gene in the variance in BMI. However, the expression correlation could be driven by a reaction to, rather than driving, BMI differences. We therefore checked whether the interacting *cis*-eQTLs are also GWAS variants for obesity traits, providing evidence that the genomic regulatory circuitry we identified in adipocytes is causal for these traits. Our approach provided functional genomics fine-mapping of four obesity GWAS loci, suggesting that adipocyte genomic regulatory mechanisms are causal for obesity traits at these loci. This work was published in *Nature Communications* in 2018.

In Chapter 3, I discuss our approach using context-specific assessment of chromatin accessibility across a dietary lipid challenge in human primary adipocytes to identify diet-driven effects on BMI. Open chromatin sites that mediate the responses to saturated or monounsaturated fatty acid intake in adipocytes were identified through differential accessibility across the

conditions. We tested genetic variants that landed in the lipid-responsive regions for interactions with saturated fat intake (based on 24-hr dietary recall) on BMI in the UK Biobank (UKB) cohort. This identified 14 promoter and 24 enhancer GxE SNPs affecting BMI in humans, providing evidence that the known correlation between saturated fat intake and BMI is mediated in part by causal, rather than simply reactive mechanisms. This work was published in *Nature Metabolism* in 2019.

In Chapter 4, I move beyond the study of the genetic causes of obesity and address the mechanisms through which obesity drives related CMDs. Given the link between preadipocytes and inflammation, discussed in section 1.4 above, we studied the preadipocytes from BMI-discordant monozygotic (MZ) twins to identify BMI-driven differences in their accessible chromatin landscapes. We found that the higher-order structure of the genome, assessed through A/B compartments, is altered in the higher-BMI subset of twins compared to the lower-BMI twins. The regions that are reprogrammed in response to BMI are enriched for the heritability of CRP and have a higher accumulation of small-effect variants interacting with BMI to affect CRP in the UKB. This work provides evidence for preadipocyte-origin mechanisms responding to increased BMI and contributing to the chronic low-grade inflammation seen in obesity. It has been submitted for publication in 2021 (Garske *et al.*, submitted).

References

Blüher, M. (2020) 'Metabolically healthy obesity', *Endocrine Reviews*, 41(3), pp. 405–420. doi: 10.1210/endrev/bnaa004.

Buenrostro, J. D. *et al.* (2015) 'ATAC-seq: A method for assaying chromatin accessibility genome-wide', *Current Protocols in Molecular Biology*, 2015, pp. 21.29.1-21.29.9. doi: 10.1002/0471142727.mb2129s109.

Cercato, C. and Fonseca, F. A. (2019) 'Cardiovascular risk and obesity', *Diabetology & Metabolic Syndrome*, 11(1), p. 74. doi: 10.1186/s13098-019-0468-0.

Cirillo, L. A. *et al.* (2002) 'Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4', *Molecular Cell*, 9(2), pp. 279–289. doi: 10.1016/S1097-2765(02)00459-8.

Constant, V. A. *et al.* (2006) 'Macrophage-conditioned medium inhibits the differentiation of 3T3-L1 and human abdominal preadipocytes', *Diabetologia*, 49(6), pp. 1402–1411. doi: 10.1007/s00125-006-0253-0.

Dixon, J. R. *et al.* (2012) 'Topological domains in mammalian genomes identified by analysis of chromatin interactions', *Nature*, 485(7398), pp. 376–380. doi: 10.1038/nature11082.

Dixon, J. R. *et al.* (2015) 'Chromatin architecture reorganization during stem cell differentiation', *Nature*, 518(7539), pp. 331–336. doi: 10.1038/nature14222.

Dunham, I. *et al.* (2012) 'An integrated encyclopedia of DNA elements in the human genome', *Nature*, 489, pp. 57–74. doi: 10.1038/nature11247.

Fock, K. M. and Khoo, J. (2013) 'Diet and exercise in management of obesity and overweight', *Journal of Gastroenterology and Hepatology (Australia)*, 28(S4), pp. 59–63. doi: 10.1111/jgh.12407.

Gate, R. E. *et al.* (2018) 'Genetic determinants of co-accessible chromatin regions in activated T cells across humans', *Nature Genetics*, 50, pp. 1140–1150. doi: 10.1038/s41588-018-0156-2.

Ghaben, A. L. and Scherer, P. E. (2019) 'Adipogenesis and metabolic health', *Nature Reviews Molecular Cell Biology*, 20, pp. 242–258. doi: 10.1038/s41580-018-0093-z.

Grob, S. and Cavalli, G. (2018) 'Technical review: A Hitchhiker's guide to chromosome conformation capture', in *Methods in Molecular Biology*, pp. 233–246. doi: 10.1007/978-1-4939-7318-7_14.

Hotamisligil, G. S. (2006) 'Inflammation and metabolic disorders', *Nature*, pp. 860–867. doi: 10.1038/nature05485.

Huvenne, H. *et al.* (2016) 'Rare Genetic Forms of Obesity: Clinical Approach and Current Treatments in 2016', *Obesity Facts*, pp. 158–173. doi: 10.1159/000445061.

Khera, A. V. *et al.* (2019) 'Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood', *Cell*, 177(3), pp. 587–596. doi: 10.1016/j.cell.2019.03.028.

Lacasa, D. *et al.* (2007) 'Macrophage-secreted factors impair human adipogenesis: Involvement of proinflammatory state in preadipocytes', *Endocrinology*, 148(2), pp. 868–877. doi: 10.1210/en.2006-0687.

Lieberman-Aiden, E. *et al.* (2009) 'Comprehensive mapping of long-range interactions reveals folding principles of the human genome.', *Science (New York, N.Y.)*, 326(5950), pp. 289–293. doi: 10.1126/science.1181369.

Lin, Y. C. *et al.* (2012) 'Global changes in the nuclear positioning of genes and intra-and interdomain genomic interactions that orchestrate B cell fate', *Nature Immunology*, 13(12), pp. 1196–1204. doi: 10.1038/ni.2432.

Locke, A. *et al.* (2015) 'Genetic studies of body mass index yield new insights for obesity

biology', *Nature*, 518(7538), pp. 197–206. doi: 10.1038/nature14177.Genetic.

Maurano, M. T. *et al.* (2012) 'Systematic localization of common disease-associated variation in regulatory DNA', *Science*, 337(6099), pp. 1190–1195. doi: 10.1126/science.1222794.

McAllister, K. *et al.* (2017) 'Current Challenges and New Opportunities for Gene-Environment Interaction Studies of Complex Diseases', in *American Journal of Epidemiology*, pp. 753–761. doi: 10.1093/aje/kwx227.

Mifsud, B. *et al.* (2015) 'Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C', *Nature Genetics*, 47(6), pp. 598–606. doi: 10.1038/ng.3286.

Nora, E. P. *et al.* (2012) 'Spatial partitioning of the regulatory landscape of the X-inactivation centre', *Nature*, 485(7398), pp. 381–385. doi: 10.1038/nature11049.

Nordestgaard, B. G. *et al.* (2007) 'Nonfasting triglycerides and risk of myocardial infarction, ischemic heart disease, and death in men and women.', *JAMA : the journal of the American Medical Association*, 298(3), pp. 299–308. doi: 10.1001/jama.298.3.299.

Ouchi, N. *et al.* (2011) 'Adipokines in inflammation and metabolic disease', *Nature Reviews Immunology*, pp. 85–97. doi: 10.1038/nri2921.

Poulain-Godefroy, O. and Froguel, P. (2007) 'Preadipocyte response and impairment of differentiation in an inflammatory environment', *Biochemical and Biophysical Research Communications*, 356(3), pp. 662–667. doi: 10.1016/j.bbrc.2007.03.053.

Ramos-Rodríguez, M. *et al.* (2019) 'The impact of proinflammatory cytokines on the β-cell regulatory landscape provides insights into the genetics of type 1 diabetes', *Nature Genetics*, 51(11), pp. 1588–1595. doi: 10.1038/s41588-019-0524-6.

Rao, S. S. P. *et al.* (2014) 'A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping', *Cell*, 159(7), pp. 1665–1680. doi: 10.1016/j.cell.2014.11.021.

Roadmap Epigenomics Consortium *et al.* (2015) 'Integrative analysis of 111 reference human epigenomes', *Nature*, 518, pp. 317–330. doi: 10.1038/nature14248.

Schoenfelder, S. *et al.* (2015) 'The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements', *Genome Research*, 25(4), pp. 582–597. doi: 10.1101/gr.185272.114.

Tchang, B. G., Saunders, K. H. and Igel, L. I. (2021) 'Best Practices in the Management of Overweight and Obesity', *Medical Clinics of North America*, pp. 149–174. doi: 10.1016/j.mcna.2020.08.018.

Virani, S. S. *et al.* (2021) 'Heart Disease and Stroke Statistics-2021 Update: A Report From the American Heart Association', *Circulation*, 143(8), pp. e254-2743. doi: 10.1161/CIR.0000000000000950.

Yan, F. *et al.* (2020) 'From reads to insight: A hitchhiker's guide to ATAC-seq data analysis', *Genome Biology*, p. 22. doi: 10.1186/s13059-020-1929-3.

Chapter 2

Integration of human adipocyte chromosomal interactions with adipose gene expression prioritizes obesity-related genes from GWAS

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-018-03554-9 OPEN

Integration of human adipocyte chromosomal interactions with adipose gene expression prioritizes obesity-related genes from GWAS

David Z. Pan[1,2], Kristina M. Garske[1], Marcus Alvarez[1], Yash V. Bhagat[1], James Boocock[1], Elina Nikkola[1], Zong Miao[1,2], Chelsea K. Raulerson[3], Rita M. Cantor[1], Mete Civelek[4], Craig A. Glastonbury[5], Kerrin S. Small[6], Michael Boehnke[7], Aldons J. Lusis[1], Janet S. Sinsheimer[1,8], Karen L. Mohlke[3], Markku Laakso[9], Päivi Pajukanta[1,2,10] & Arthur Ko[1,10]

Increased adiposity is a hallmark of obesity and overweight, which affect 2.2 billion people world-wide. Understanding the genetic and molecular mechanisms that underlie obesity-related phenotypes can help to improve treatment options and drug development. Here we perform promoter Capture Hi-C in human adipocytes to investigate interactions between gene promoters and distal elements as a transcription-regulating mechanism contributing to these phenotypes. We find that promoter-interacting elements in human adipocytes are enriched for adipose-related transcription factor motifs, such as PPARG and CEBPB, and contribute to heritability of *cis*-regulated gene expression. We further intersect these data with published genome-wide association studies for BMI and BMI-related metabolic traits to identify the genes that are under genetic *cis* regulation in human adipocytes via chromosomal interactions. This integrative genomics approach identifies four *cis*-eQTL-eGene relationships associated with BMI or obesity-related traits, including rs4776984 and *MAP2K5*, which we further confirm by EMSA, and highlights 38 additional candidate genes.

[1] Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, CA 90095, USA. [2] Bioinformatics Interdepartmental Program, UCLA, Los Angeles, CA 90095, USA. [3] Department of Genetics, University of North Carolina, Chapel Hill, NC 27599, USA. [4] Department of Biomedical Engineering, University of Virginia, Charlottesville, VA 22904, USA. [5] Big Data Institute, University of Oxford, Oxford OX3 7LF, UK. [6] Department of Twin Research and Genetic Epidemiology, King's College, London, UK. [7] Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA. [8] Department of Biomathematics, David Geffen School of Medicine at UCLA, Los Angeles, CA 90095, USA. [9] Institute of Clinical Medicine, Internal Medicine, University of Eastern Finland and Kuopio University Hospital, FI-70210 Kuopio, Finland. [10] Molecular Biology Institute at UCLA, Los Angeles, CA 90095, USA. These authors contributed equally: David Z. Pan, Kristina M. Garske. Correspondence and requests for materials should be addressed to A.K. (email: a5ko@ucla.edu)

Obesity is a serious health epidemic world-wide. A recent study of 195 countries estimated that 2.2 billion people were overweight or obese in 2015[1]. Clinically, obesity is diagnosed by a body mass index (BMI) greater than 30. While a significant proportion of the phenotypic variation in BMI is attributed to genetic variation (heritability of BMI ~0.4–0.7[2]), understanding the mechanisms underlying this heritable component has been challenging. The 97 loci identified in a genome-wide association study (GWAS) for BMI in ~340,000 subjects explain only 2.7% of the variance in BMI, and all HapMap phase 3 genetic variants (~1.5 M single nucleotide polymorphisms (SNPs)) were estimated to account for ~21% of the variance in BMI in 16,275 unrelated individuals[2]. The causal variants and genes are not immediately apparent from GWAS, hindering our ability to understand the biological mechanisms by which genetics contribute to obesity. To address this knowledge gap, we integrate chromosomal interaction data from primary human white adipocytes (HWA) with adipose gene expression and clinical phenotype data (BMI, waist-hip ratio, fasting insulin, and Matsuda index) to elucidate molecular pathways involved in genetic regulation in *cis*.

Combining genotype and RNA-sequencing (RNA-seq) data enables the detection of expression quantitative trait loci (eQTLs) that regulate transcription of near-by genes (i.e., in *cis*). These *cis*-eQTLs often reside in regulatory elements, including promoters, enhancers, and super-enhancers. However, the mechanism by which *cis*-eQTLs regulate their respective eGene(s) is seldom established because identification of the true regulatory variants among SNPs in tight linkage disequilibrium (LD) has proven challenging[3]. Enhancers modulate target gene expression levels via their interaction with promoters, and disruption or improper looping of enhancer sites can contribute to disease risk[4,5]. Promoter Capture Hi–C (pCHi–C) enables detection of promoter interactions at a higher resolution and at lower sequencing depth than that required for Hi–C[6]. Incorporating a chromosomal interaction map constructed from pCHi-C and *cis*-eQTL data can help elucidate the functional mechanisms by which the genetic variants affect gene expression. By overlapping these looping *cis*-eQTLs with trait-associated variants identified in independent, large-scale GWAS, we can assess which GWAS variants could affect expression of regional genes via chromosomal interactions.

To search for genes that are functionally important for adipose tissue biology, we performed a *cis*-eQTL analysis using genome-wide SNP data and adipose RNA-seq data from individuals of the Finnish METabolic Syndrome In Men (METSIM) cohort. We identified 42 genes, regulated by *cis*-eQTLs that reside in regions that physically interact with the promoters of genes. Adipose expression of these 42 genes was robustly correlated with BMI, and among them four genes, *MAP2K5*, *LACTB*, *ORMDL3*, and *ACADS*, were regulated by SNPs (or their tight LD proxies) previously identified in GWAS for BMI or a related metabolic trait, located at the regulatory element-promoter interaction sites. These data provide converging evidence for effects of looping *cis*-eQTL variants on gene expression associated with obesity and related metabolic traits. Our results show that these integrative genomics methods involving pCHi-C data in primary HWA can identify regulatory circuits comprising both regulatory elements and their target gene(s) that operate in a complex obesity-related metabolic trait.

Results

Characterization of the adipocyte chromosomal interactions. Adipose tissue is highly heterogeneous, containing adipocytes, preadipocytes, stem cells, and various immune cells. We performed pCHi-C in primary HWA with the goal of identifying physical interactions between adipose *cis*-eQTLs and target gene promoters. We employed the pCHi-C protocol as described previously[7]. Briefly, we fixed primary HWA to crosslink proteins to DNA, and after digestion with the *Hind*III restriction endonuclease, we performed in-nucleus ligation and biotinylated RNA bait hybridization to pull down only those *Hind*III fragments with annotated gene promoters[6]. To detect the regions that interact with the promoter-containing *Hind*III fragments, we mapped the reads to the genome, and assigned reads to *Hind*III fragments to allow for fragment-level resolution of those regions interacting with the baited fragments containing gene promoters. The key pCHi-C sequencing metrics are shown in Supplementary Table 1.

We first confirmed that the non-promoter regions in adipocyte chromosomal interactions are enriched for enhancer (H3K4me1, H3K4me3, and H3K27ac), repressor (H3K27me3, H3K9me3) histone marks, and DNase I hypersensitive sites (DHSs) (Supplementary Table 2). As there are no publicly available DHS data for adipocytes or adipose tissue, we used the union of DHSs in all cell types from ENCODE and Roadmap rather than DHSs in a single, non-adipocyte cell type[8]. Intersecting the adipocyte and previously published primary $CD34^+$ cell pCHi-C data[6], we found that 68.0% of adipocyte pCHi-C chromosomal interactions were observed in adipocytes but not in $CD34^+$cells. In the following, we used the same public DHS data to focus on open chromatin regions as they are more likely to bind transcription factors (TFs) and, thus, be relevant for chromosomal looping interactions within the interacting *Hind*III fragments.

We examined whether the DHSs are enriched for adipose-related TF motifs, using the Hypergeometric Optimization of Motif EnRichment (HOMER) software[9] that calculates the number of times a TF motif is seen in target and background sequences. The proportion of times the TF motif is seen in the target when compared to the background is then tested for enrichment in the target sequences. We found that when compared to DHSs within $CD34^+$ chromatin interactions, the DHSs within the adipocyte chromatin interactions are enriched for 26 of 332 TF motifs (FDR < 5%) (Supplementary Table 3), including CCAAT/enhancer binding protein beta (CEBPB, p-value $= 1.00 \times 10^{-10}$) and peroxisome proliferator-activated receptor gamma (PPARG, p-value = 0.01), both of which are well-known key players in adipose biology[10]. To address the potential bias of using a different pCHi-C dataset as background, we also performed HOMER analysis comparing the DHSs in adipocyte interactions to DHSs in non-interacting, non-promoter regions in the remainder of the genome. The results were similar, and both CEBPB and PPARG were also enriched in the latter analysis (CEBPB, p-value $= 1.00 \times 10^{-24}$; PPARG, p-value $= 1.00 \times 10^{-6}$; complete enrichment results not shown). These results suggest that the cell-type based pCHi-C interaction data enable the detection of interactions important for that cell type within a heterogeneous human tissue.

Chromosomal interactions explain expression heritability. To investigate whether the variants residing within open chromatin of chromosomal looping regions in adipocytes are enriched for SNPs that contribute to the heritability of *cis* expression regulation, we partitioned the heritability of *cis* regulation of human adipose gene expression into 52 functional categories using a modified partitioned LD Score regression method[11] (see Methods). The 52 functional categories are derived from 26 main annotations that include coding regions, untranslated regions (UTRs), promoters, intronic regions, histone marks, DNase I hypersensitivity sites (DHSs), predicted enhancers, conserved regions, and other annotations[11] (Supplementary

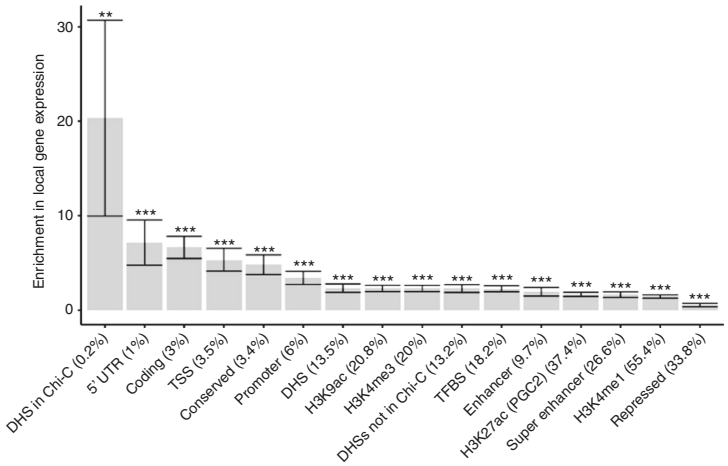


Fig. 1 Open chromatin sites (DHSs) within adipocyte promoter CHi-C chromosomal interactions show significant enrichment in *cis* expression. Enrichments in *cis* expression with error bars for different categories using LD score regression analysis (see Methods). For the horizontal axis labels, the value in parentheses shows the percentage of SNPs contained within the respective annotation category that contributed to the enrichment calculation. For the significance threshold after Bonferroni correction above each bar, * indicates a p-value < 0.05; **, a p-value < 0.001; and ***, a p-value < 0.0001, respectively. The p-values were estimated based on Z scores calculated from the normal distribution. Error bars represent jackknife standard errors around the estimates of enrichment

Figure 1, Supplementary Tables 4–5). The partitioned LD Score regression method[11] utilizes summary association statistics of all variants on gene expression to estimate the degree to which variants in different annotation categories explain the heritability of *cis* and *trans* expression regulation while accounting for the LD among functional annotations. To assess the enrichment of heritability mediated by the variants in the chromosomal interactions detected by pCHi-C on a per-gene basis, we further modified the LD score method, as described in detail in the Methods. Importantly, these modifications did not change the 52 baseline enrichments significantly when compared with the data obtained using the unmodified version[11] (Supplementary Figure 1, Supplementary Tables 4–5). These analyses revealed that open chromatin regions (i.e., DHSs) within the adipocyte chromosomal interactions are enriched for sequences that contribute to heritability of gene expression regulation in *cis* (Fig. 1, p-value < 0.002, enrichment = 20.3 (SD±5.2), average proportion of SNPs = 0.23%). The variants residing within the open chromatin regions within adipocyte chromosomal interactions explain 4.6% of the heritability of adipose tissue gene expression in *cis*, despite only accounting for 0.23% of the SNPs per *cis* gene region on average, indicating the functionality of these SNPs at the DHSs of distal interactions in regulating *cis* expression.

Identification of genes regulated by looping *cis*-eQTL SNPs. To identify adipose-expressed genes regulated by SNPs (eGenes), we performed a *cis*-eQTL analysis using 335 individuals from the METSIM cohort with both genome-wide SNP data and adipose RNA-seq data available (Fig. 2; Methods). Using the published adipose *cis*-eQTL data and criteria for significance from GTEx[12] (see Methods), we found 487,679 *cis*-eQTLs for 4,650 eGenes in the METSIM dataset and confirmed these same SNPs as *cis*-eQTLs by look-up in GTEx. 386,068 of the 487,679 (79.0%) *cis*-eQTL SNPs had the same target gene and direction of effect in both cohorts (Supplementary Figure 2). Only the 386,068 *cis*-eQTL SNPs that were replicated for effect direction and target gene (Supplementary Table 1) in the GTEx adipose RNA-seq data were used in our subsequent downstream analyses (Supplementary Figure 2). Overall, 4,332 of 4,650 of *cis*-eQTL-eGene relationships (93.0%) were replicated using the published adipose *cis*-eQTL data and criteria for significance from GTEx[12] (see Methods). To restrict these adipose *cis*-eQTL SNPs to those that likely function through transcription factor (TF) binding at distal regulatory elements, we determined which of these eGene promoters were involved in looping interactions with the *cis*-eQTLs, assayed through pCHi-C in primary HWA (Fig. 2; Methods). Of the 4,332 eGenes identified in our *cis*-eQTL analysis, 576 (13.4%), were involved in these looping interactions (permutation p-value < 0.00001) (Fig. 2, Methods, Supplementary Figure 2, and Supplementary Table 1).

We next determined the set of 576 looping eGenes with expression levels that are correlated with BMI in METSIM (Pearson correlation, adjusted $p < 1.15 \times 10^{-5}$ to correct for the 4,332 replicated eGenes identified in our *cis*-eQTL analysis). We found 54 of 576 (9.40%) BMI-correlated eGenes with promoters involved in looping interactions with their *cis*-eQTL SNP (Supplementary Table 6). In our subsequent second replication analysis, the expression levels of 42 out of 54 genes (replication rate of 77.8%) were correlated with BMI in adipose RNA-seq data from the TwinsUK cohort ($n = 720$) with the same direction of effect on BMI as in METSIM (Bonferroni adjusted $p < 0.001$) (Table 1, Supplementary Table 6). Another four of the 54 genes were not available in the TwinsUK dataset. The effects sizes and p-values obtained for BMI associations in TwinsUK and METSIM, using a linear regression model in both, show comparable results to those obtained using the Pearson correlations (Table 1, Supplementary Table 6). These 42 BMI-correlated genes are functionally enriched for four pathways with fatty acid metabolism as a top ranking pathway (Supplementary Table 7) based on KEGG pathway enrichment using WebGestalt[13] (Benjamini-Hochberg adjusted $p < 0.05$); however, the small

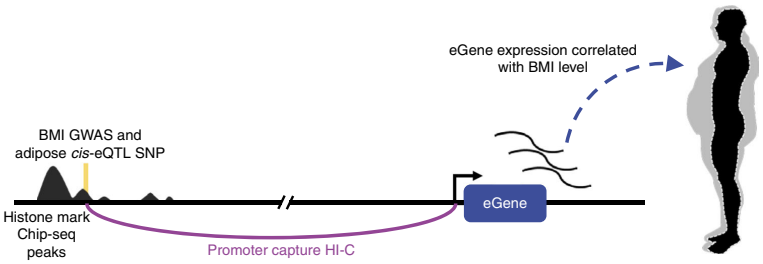


Fig. 2 Overview of the study design targeted to identify new genes for obesity and related metabolic traits. A schematic illustrating the integration of multi-omics data utilized in this study to elucidate genetics of obesity-related traits.

number of genes in these pathway analyses warrant verification in future studies. Only these 42 replicated genes were further investigated in our downstream analyses.

Adipocyte chromosomal interactions prioritize GWAS genes. To investigate which of the 42 BMI-correlated eGenes are regulated by GWAS variants previously identified for BMI and related metabolic traits, we determined which interacting *cis*-eQTL variants are GWAS variants (or their LD proxies, $r^2 > 0.80$), using $p < 5.00 \times 10^{-8}$ as a criterion to select variants. As the goal of the current study was to dissect the molecular contribution of adipose and adipocyte biology to traits that can influence the pathophysiology of obesity, we examined GWAS for BMI and the metabolic traits that have previously been shown to exhibit comorbidities with obesity (e.g., serum lipids and type 2 diabetes) or that are influenced by obesity or correlated with BMI (e.g., metabolites and WHR). We used all GWAS variants (p-value < 5.00×10^{-8}) identified in a previous metabolite GWAS of ~7000 individuals[14], lipid GWAS of ~180,000 individuals[15], an extensive BMI GWAS study of ~340,000 individuals[2], a sequencing-based GWAS for type 2 diabetes[16], and a waist-hip-ratio (WHR) adjusted for BMI GWAS of ~220,000 individuals[17]. We found a GWAS variant for BMI, regulating mitogen-activated protein kinase kinase 5 (*MAP2K5*); a GWAS variant for high-density lipoprotein cholesterol (HDL-C), regulating orosomucoid like sphingolipid biosynthesis regulator 3 (*ORMDL3*); and two GWAS variants for serum metabolites (succinylcarnitine and butyrylcarnitine), regulating lactamase beta (*LACTB*) and acyl-CoA dehydrogenase, C-2 To C-3 short chain (*ACADS*), among the 42 genes (Fig. 3a, b; Supplementary Figure 3a-f), with the looping interactions spanning 287 kb, 16 kb, 151 kb, and 183 kb, respectively. We found that the interacting *cis*-eQTL-containing *Hind*III fragments for *LACTB* and *MAP2K5* are located within the promoter and intron of other genes. Furthermore, using the integrated pCHi-C and *cis*-eQTL data, we found that the SNPs in these regulatory *Hind*III fragments regulate genes that are not their nearest gene for 3 of the 4 BMI-correlated eGenes (Fig. 3a, b, Supplementary Figure 3a–f).

The looping BMI GWAS SNPs regulate *MAP2K5*. For *MAP2K5*, the reported BMI GWAS SNP itself is not located within the regulatory, *cis*-eQTL-containing *Hind*III fragment involved in the looping interaction; however, SNPs in tight LD with the GWAS SNP (using a criterion of $r^2 > 0.80$) are in the regulatory *Hind*III fragment that is interacting with the target gene promoter (Fig. 3b). The regulatory *Hind*III fragment contains 16 *cis*-eQTL SNPs that are LD proxies for the BMI GWAS SNP[2] (rs16951275), which has a total of 62 LD proxies in the METSIM cohort. To prioritize a candidate functional variant within these 16 SNPs within the *Hind*III fragment, we first examined the predicted TF motifs that may be affected by each SNP using the data curated from ChIP-seq by Kheradpour and Kellis[18]. We found that only rs4776984, which is in almost perfect LD with the original BMI GWAS variant rs16951275 ($r^2 = 0.98$), showed a predicted increase in binding of CTCF, which is a TF known to mediate chromosomal interactions (Fig. 4a).

We also used the deep learning–based sequence analyzer (DeepSEA)[19] to examine the allelic effect on protein binding of rs4776984 and the 15 other looping *cis*-eQTLs of *MAP2K5*. Of these 16 looping *cis*-eQTLs, six were potentially functional and of these, two variants passed the functional significance score of <0.05 using DeepSEA. Of the two, our candidate functional eQTL SNP, rs4776984, resulted in the most significant functional score (2.36×10^{-3}) (Supplementary Table 8) and was the only variant passing a functional significance score of <0.01 among the 16 variants. Thus, the DeepSEA result further supports the differential TF binding at the variant site rs4776984 among all possible looping *cis*-eQTLs at the *MAP2K5* locus (Supplementary Table 8). The looping *cis*-eQTL site also shows a ChIP-seq peak for the histone mark H3K4me1 in ENCODE adipose nuclei ChIP-seq data; however, notably it also shows the presence of the histone marks H3K27me3 and H3K9me3 (Fig. 3b), two marks known to be associated with transcriptional repression. Furthermore, the gene expression of *MAP2K5* is negatively correlated with BMI (p-value = 7.83×10^{-6}). These data implicate *MAP2K5* as a gene regulated by the BMI GWAS signal via a repressive chromosomal interaction.

To functionally assess whether there is differential allele-specific binding of proteins at the candidate functional *MAP2K5* eQTL, rs4776984, we performed electrophoretic mobility shift assays (EMSAs) using nuclear protein from primary HWA. The results show reduced protein binding of the reference allele when compared to the alternate allele of rs4776984, consistently in three independent experiments (Fig. 4b, Supplementary Figure 4), in line with the predicted disruption in protein binding for CTCF[18] (Fig. 4a). We performed the supershift experiment using the CTCF antibody and adipocyte nuclear extract, but did not observe a supershift in any of the three replicated experiments (Supplementary Figure 5). We repeated the supershift experiment using a different CTCF antibody (EMD Millipore 07–729), which resulted in the same negative finding (Supplementary Figure 6). To further verify the negative supershift result, we also directly tested the CTCF protein for allele-specific binding at rs4776984 using EMSA in 3 replicate experiments, and did not find evidence of sole CTCF protein binding (Supplementary

Table 1 Thirteen representative eGenes (9 most significant genes and 4 GWAS loci) that correlate with BMI in METSIM and TwinsUK (for the full data on all 54 genes, see Supplementary Table 6)

			Pearson		Linear regression					
Rank[a]	Gene	Chr[f]	METSIM[c]		METSIM[d]			TwinsUK[e]		
			Effect size (r)	*p*-value	Effect size (β)	SE	*p*-value	Effect size (β)	SE	*p*-value
1	*ADH1B*	4	−0.45	7.40×10^{-18}	−0.21	0.02	1.68×10^{-20}	−0.58	0.03	4.47×10^{-71}
2	*ORMDL3*[b]	17	−0.45	8.57×10^{-18}	−0.16	0.02	2.06×10^{-20}	−0.58	0.03	2.65×10^{-70}
3	*AKR1C3*	10	0.33	4.78×10^{-10}	0.13	0.02	2.95×10^{-11}	0.49	0.03	5.19×10^{-54}
4	*CMTM3*	16	0.41	4.32×10^{-15}	0.087	0.01	3.84×10^{-17}	0.50	0.03	6.64×10^{-52}
5	*LPIN1*	2	−0.38	1.49×10^{-13}	−0.14	0.02	2.27×10^{-15}	−0.47	0.03	2.38×10^{-44}
6	*RNF157*	17	−0.29	5.19×10^{-8}	−0.096	0.02	5.87×10^{-9}	−0.47	0.03	8.86×10^{-42}
7	*MYOF*	10	0.32	1.07×10^{-9}	0.086	0.01	7.37×10^{-11}	0.46	0.03	2.59×10^{-40}
8	*NAA40*	11	0.28	1.81×10^{-7}	0.052	0.009	2.67×10^{-8}	0.46	0.03	4.00×10^{-40}
9	*TMEM165*	4	0.33	2.45×10^{-9}	0.045	0.007	1.84×10^{-10}	0.45	0.03	3.52×10^{-37}
10	*RFFL*	11	0.27	1.02×10^{-6}	0.035	0.006	1.84×10^{-7}	0.43	0.03	5.67×10^{-37}
28	*ACADS*[b]	12	−0.37	2.91×10^{-12}	−0.085	0.01	7.12×10^{-14}	−0.24	0.03	6.65×10^{-19}
31	*LACTB*[b]	15	0.30	1.67×10^{-8}	0.069	0.01	1.40×10^{-9}	0.32	0.04	4.94×10^{-18}
34	*MAP2K5*[b]	15	−0.25	7.83×10^{-6}	−0.039	0.01	1.90×10^{-6}	−0.21	0.03	3.81×10^{-10}

[a] Thirteen representative eGenes, including 4 GWAS loci, ranked by their *p*-value in the TwinsUK cohort dataset
[b] GWAS gene
[c] Effect size (r, Pearson rho) and *p*-value calculated from Pearson correlation between gene expression and BMI (see Methods)
[d] Effect size, standard error (SE), and *p*-value calculated using a linear regression model with BMI and age, age^2 and the 14 technical factors as covariates when compared to a null model without BMI. These models were compared using an F-test (see Methods)
[e] Effect size, standard error (SE), and *p*-value calculated from linear mixed effects model. A full model including BMI was compared to a null model in which the same model was fitted, but with BMI omitted. These models were compared using an F-test (see Methods)
[f] Chr is an abbreviation for chromosome

Figure 7). However, a supershift experiment may remain negative even in the presence of true TF binding if a complex instead of a single TF alone is required for the TF binding[20].

Interacting GWAS SNPs implicate three other genes. For *ORMDL3*, there is a single lipid GWAS SNP, rs8076131, in the *Hind*III fragment, which is also the only looping *cis*-eQTL SNP interacting with the *ORMDL3* promoter. Variant rs8076131 lies in a region with enhancer histone marks H3K4me1 and H3K27ac in adipose nuclei (Supplementary Figure 3a,b). The expression of *ORMDL3* is negatively correlated with BMI ($p = 8.57 \times 10^{-18}$), in line with its known role as a negative regulator of sphingolipids that are positively correlated with obesity[21,22].

The regulatory *Hind*III fragment that loops with the *LACTB* promoter contains two reported metabolite GWAS SNPs in tight LD with each other (rs1017546 and rs3784671, $r^2 = 0.97$), both sharing 35 LD proxies ($r^2 > 0.80$) in the METSIM cohort. One of the two index GWAS SNPs within the *Hind*III fragment, rs3784671, resides in a region enriched for the enhancer histone marks H3K4me1 and H3K27ac in adipose nuclei (Supplementary Figure 3c, d). This metabolite GWAS SNP, rs3784671, is associated with succinylcarnitine levels, which have previously been shown to be positively correlated with BMI in KORA ($p = 1.0 \times 10^{-12}$) and TwinsUK ($p = 5.3 \times 10^{-5}$)[23]. Accordingly, the expression of *LACTB* is positively correlated with BMI ($p = 1.19 \times 10^{-8}$). Notably, *LACTB* has been implicated as a causal gene for obesity in mice[24], further supporting our integrated human data that implicates LACTB involvement in an obesity-related metabolic trait.

The most significant metabolite GWAS SNP for *ACADS*, rs10774569, is not located within the regulatory, *cis*-eQTL-containing *Hind*III fragment. Instead, a single *cis*-eQTL SNP rs12310161, in perfect LD ($r^2 = 1.0$) with the GWAS SNP rs10774569, is the only *cis*-eQTL SNP located within the regulatory *Hind*III fragment, looping with the fragment containing the promoter of *ACADS*. This looping *cis*-eQTL SNP also resides in a region enriched for enhancer histone marks H3K4me1 and H3K27ac in adipose nuclei (Supplementary Figure 3e, f). The expression of *ACADS* has a negative correlation with BMI ($p = 2.91 \times 10^{-12}$), and the alternate allele is associated with an increase in expression of *ACADS*, suggesting that this allele has a protective effect against obesity.

Finally, we repeated the pCHi-C experiments in the same HWA cell line in a separate experiment with two replicates and found the same GWAS SNP interactions as in the first experiment (Supplementary Table 9). This validation data thus provides further support for our conclusions and the robustness of interactions we report.

Discussion

BMI is a highly complex trait caused by the poorly characterized interplay between genetic and environmental factors with upper heritability estimates reaching 70%[2]. Understanding how genome-wide signals with small effect sizes contribute to BMI on a molecular level has proven to be difficult. Delineating the underlying biological mechanisms of these signals is crucial to better understand the development of obesity and its concomitant cardiometabolic disorders. In this study, we performed promoter Capture Hi–C (pCHi–C) in primary human white adipocytes (HWA) to identify BMI-correlated adipose-expressed genes that are under genetic regulation in *cis* by variants that physically interact with gene promoters. Through our method of integrating GWAS, *cis*-eQTL analyses, chromosomal interactions, and robust replication of the data from GTEx and TwinsUK, we were able to identify 42 candidate genes for future obesity research.

In the absence of adipocyte DHS information, we used DHS data from all tissues in the ENCODE and Roadmap Epigenomics project to label open chromatin regions within the adipocyte chromosomal interactions[8]. Despite this methodological compromise, our results demonstrate that variants in these regions

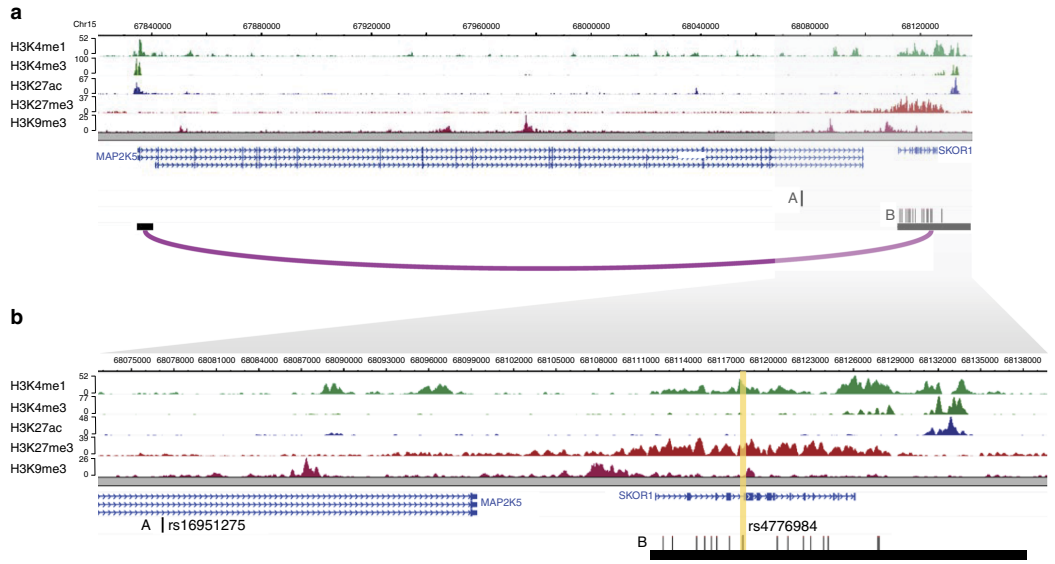


Fig. 3 Promoter Capture Hi-C enables refinement of the BMI GWAS locus that colocalizes with *cis*-eQTLs interacting with the target gene promoter of *MAP2K5*. Genomic landscape of the BMI locus, *MAP2K5* (panels **a**, **b**), modified from the WashU Genome Browser to show the histone mark calls from ChIP-seq data; gene transcripts; promoter and eQTL *Hind*III fragments that interact in primary human white adipocytes (HWA); and GWAS SNPs (A, the rs number indicated in the magnified box) or their LD proxies (B, $r^2 > 0.8$) located in the interacting *Hind*III fragment. The vertical yellow band highlights the *cis*-eQTL variant (the rs number is indicated in the magnified box). **a** Genomic landscape containing *MAP2K5* and the interacting *cis*-eQTL variant and corresponding BMI GWAS SNP. **b** Magnification of the boxed region in (**a**)

explain a significant portion (4.6%) of the heritability of *cis*–regulated expression in human subcutaneous adipose tissue. Even though the total percentage of variants within the intersection of open chromatin regions and adipocyte chromosomal looping sites is small (0.23%), the enrichment implies that these SNPs are functionally relevant for adipocyte biology and gene regulation in *cis*.

The enrichment of TF binding motifs for CEBPB and PPARG in chromosomal interactions found in adipocyte but not in CD34$^+$ cells confirms that the regulatory circuits identified here are relevant to adipose biology. These two TFs have previously been shown to occupy shared regulatory sites. Apart from being an enhancer binding protein, which is in concordance with its presence at chromosomal interaction sites, CEBPB has been demonstrated to precede the binding of PPARG at many regulatory sites[25], suggesting that CEBPB primes the regulatory regions for the binding of the adipose master regulator PPARG.

One of our looping *cis*-eQTL variants is a tight LD proxy ($r^2 = 0.98$) for a regional BMI lead GWAS SNP (rs16951275)[2]. Typical fine mapping techniques such as overlaying histone marks, transcription factor motif scans, or eQTL searches do not necessarily reveal the mechanism through which a SNP might function. We refined the GWAS signal from 64 to 16 LD SNPs within a *Hind*III fragment that interacts with the *MAP2K5* promoter by overlaying *cis*-eQTLs, the promoter-enhancer interaction map, and the expression-BMI correlation. The top candidate, rs4776984, increased HWA nuclear protein binding in an allele-specific way in our EMSA experiment and lies within the repressor histone marks H3K27me3 and H3K9me3 in ENCODE adipose nuclei data. Recent studies have suggested that repressor elements function through looping interactions in a similar manner to enhancer elements[6,26], which would align well with the negative correlation between expression of *MAP2K5* and BMI level.

The region at the *MAP2K5* locus, exhibiting increased binding for the alternate allele for rs4776984, contains predicted motifs for the looping interaction protein, CTCF, and other TFs (Supplementary Table 8). We did not find evidence of CTCF binding at rs4776984 in our supershift and protein binding EMSA experiments. However, a supershift experiment may remain negative even in the presence of true TF binding if a complex instead of a single TF alone is required for the TF binding[20]. Furthermore, using DeepSEA analysis, we confirmed the potential for differential TF binding at the variant site rs4776984 among all possible looping *cis*-eQTLs at the *MAP2K5* locus. Noteworthy, since DeepSEA identified multiple TFs as potential binders of rs4776984 site in an allele-specific way, future studies testing a larger set of TFs are warranted to identify the actual TF that binds this site. We postulate that TF binding at this interaction site would lead to a repressive looping mechanism, in this case altering *MAP2K5* expression in adipocytes.

MAP2K5 is a member of the ERK5 MAP kinase signaling cascade, and the importance of ERK5 signaling in adipose was previously demonstrated in *Erk5* knock-out mice, which exhibit increased adiposity[27]. This suggests that changes in ERK5 signaling in adipocytes could be relevant for human obesity. MAP2K5 is a strong and specific activator of ERK5 in the ERK5 MAP kinase signaling cascade[28], supporting further study of *MAP2K5* in connection with increased adiposity.

The intronic *ORMDL3* GWAS variant rs8076131 is associated with high-density lipoprotein cholesterol (HDL-C)[15] and is the only *cis*-eQTL SNP in the *Hind*III fragment that interacts with the *ORMDL3* promoter in our adipocyte pCHi-C data. ORMDL3 is a negative regulator of the synthesis of sphingolipids that are

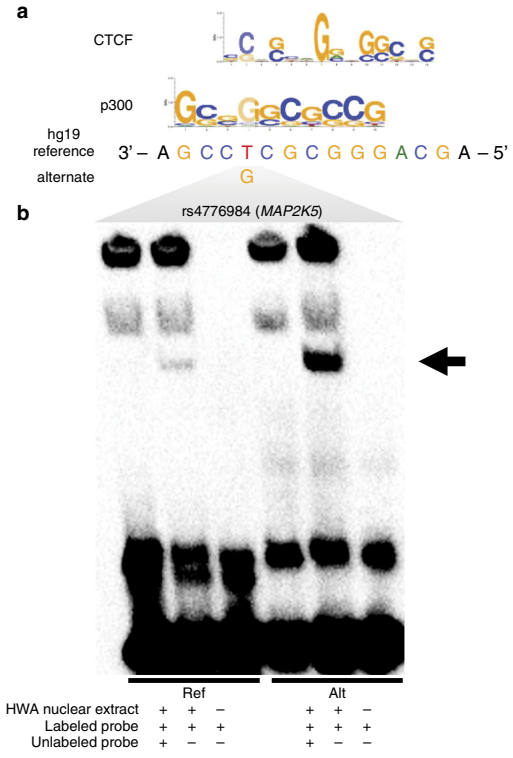


Fig. 4 Predicted TF motifs and electrophoretic mobility shift assay (EMSA) at the rs4776984 site indicate allele-specific binding. **a** Predicted TF motifs for CTCF and p300, as well as the hg19 reference genome sequence. **b** Biotinylated (labeled probe) 31-bp oligonucleotide complexes with ±15 bp flanking the reference or alternate allele for variant rs4776984 were incubated with nuclear protein extracted from primary HWA and resolved on a 6% polyacrylamide gel. Competitor assays were performed by incubating the reaction with ×100 excess of unlabeled (no biotin) oligonucleotide complexes with identical sequence. Arrow denotes specific binding of HWA nuclear protein to reference (left) and alternate (right) allele

produced in response to obesity and related metabolic traits, such as inflammation and insulin resistance[21,22], and that interfere with important signaling pathways associated with these traits[22]. Corroborating this, we show that *ORMDL3* expression is negatively correlated with BMI, and the *cis*-eQTL and risk variant rs8076131 decreases *ORMDL3* expression, potentially through a change in the chromosomal interaction between the enhancer and promoter of *ORMDL3*, as has been shown for this enhancer site previously[29].

We found that the metabolite GWAS SNP, rs3784671, is a looping *cis*-eQTL variant associated with the expression levels of the *LACTB* gene. Although this variant is a *cis*-eQTL for *LACTB* both in our study and the GTEx adipose cohort, it lies within the promoter for the *APH1B* gene, for which it is not a *cis*-eQTL in our study. Through overlap of adipose *cis*-eQTL data and adipocyte pCHi-C data, we established that rs3784671 does not act through the adjacent *APH1B* gene and filtered the 35 *cis*-eQTL variants for *LACTB* down to a single variant, rs3784671. This variant is negatively associated with the levels of succinylcarnitine, a metabolite positively correlated with BMI in two independent cohorts, KORA and TwinsUK, previously[23]. Succinylcarnitine is a molecule in the butanoate metabolism pathway; butanoate has been implicated in anti-inflammation, protection against obesity, and an increase in leptin levels[30]. Furthermore, as the succinylcarnitine GWAS variant rs3784671 is an eQTL for *LACTB*, associated with an increase in *LACTB* expression, we postulate that *LACTB* expression increases succinylcarnitine. This is in agreement with a mouse study that shows that butanoate metabolism is reduced in *Lactb* transgenic mice[24]. Notably, support for *LACTB* as a causal gene for obesity derives from functional studies using transgenic overexpression of *Lactb* in mice, resulting in an increase in the fat-mass-to-lean-mass ratio[24,31]. Although the function of LACTB in adipose has not been fully elucidated, these studies suggest that a reduction in *LACTB* function and, in turn, an increase in butanoate metabolism and decrease of succinylcarnitine levels are beneficial for obesity treatment. Further molecular studies at the protein level are, however, required to determine the function of *ORMDL3* and *LACTB* in connection with obesity.

We identified a perfect LD proxy for a metabolite GWAS SNP that lies within a *Hind*III fragment that regulates the *ACADS* gene and interacts with its promoter. ACADS is a mitochondrial protein that catalyzes the first step of the fatty acid beta-oxidation pathway. Proper mitochondrial function is imperative for adipose function and energy homeostasis. In addition to the METSIM and TwinsUK adipose RNA-seq data sets used in our study, a previous study identified *ACADS* when systematically searching for genes over and under-expressed in obese versus lean adipose tissue[32]. Furthermore, all 3 datasets show a consistent negative correlation between *ACADS* expression and BMI, in support of its well-established mitochondrial function. The interacting *cis*-eQTL and GWAS SNP, rs12310161, is located within enhancer histone marks in adipose nuclei and in the HepG2 liver cell line, with the alternate allele exhibiting a positive effect on gene expression, in line with it being a protective allele. Interestingly, this variant falls within a TEA Domain Transcription Factor 4 (TEAD4) ChIP-seq peak in the HepG2 cells. *TEAD4* expression is regulated by Peroxisome Proliferator Activated Receptor alpha (PPARα)[33], the major regulator of beta-oxidation of fatty acid pathways in liver and brown adipose tissue. Taken together, these results suggest that the interacting *cis*-eQTL and metabolite GWAS SNP, rs12310161, functions within an enhancer to increase *ACADS* expression and mitochondrial fatty acid beta-oxidation in adipose.

As the pCHi-C experiments were performed in primary HWA, we are able to focus on physical chromosomal interactions directly in human adipocytes among all cell types present in adipose tissue. Adipocytes perform central adipose functions, including lipogenesis and lipolysis. Further investigation of the adipose genes, which are under *cis* genetic regulation via chromosomal looping to the promoters and are correlated with BMI, is likely to provide much needed insight into cellular processes contributing to obesity. Our data provide 38 new candidate genes, including some known functionally relevant genes for adiposity such as *LPIN1*[34] and *AKR1C3*[35], that have so far not been highlighted by GWAS for BMI or obesity-related metabolic traits. We postulate that identification of some of these 38 candidates as obesity GWAS genes may require much larger GWA studies, while others may represent genes responding to obesity in human adipose tissue. Our analysis of the looping *cis*-eQTLs for other GWAS traits correlated with BMI, such as serum metabolites and lipids, led to the identification three additional obesity-related metabolic GWAS genes. We recognize that brain and other

tissues likely account for some of the BMI GWAS signals and that GWAS variants may act via other mechanisms, such as *trans* regulation and alternative splicing, that warrant future investigation. Although the four looping *cis*-eQTL variants identified at GWAS loci in our study represent either the GWAS tag SNPs (as is the case at the *ORMDL3* and *LACTB* loci) or they are in perfect or almost perfect LD with the GWAS SNP ($r^2 = 1.0$ at the *ACADS* locus and $r^2 = 0.98$ at the *MAP2K5* locus), we recognize that the looping variants may not always be the strongest *cis*-eQTL SNPs at these loci and, thus, additional fine mapping is needed to fully elucidate all functional regulatory *cis*-eQTL variants.

The current study uses the integration of multi-level genomic and functional data to enhance the understanding of genome-wide molecular signals underlying obesity. GWAS signals often fall within non-coding regulatory regions of the genome, and the affected gene(s) often remain unclear. Similarly, the local LD structure frequently hinders the identification and functional characterization of the actual eQTL SNP even though the eQTL target gene is known. Through the integration of multi-layer genomics data in a functionally relevant human cell type and tissue and replication in the GTEx and TwinsUK cohorts, we show that the DHSs within the interacting chromosomal regions are enriched for tissue-specific TF motifs and explain a significant proportion of the heritability of gene expression in *cis*. Furthermore, we identified *LACTB*, *ACADS*, *ORMDL3*, and *MAP2K5* as obesity-related genes in humans and provide a set of 38 non-GWAS candidate genes for future studies in obesity.

Methods

Cell lines and culture reagents. We obtained and cultured the primary human white preadipocyte (HWP) cells as recommended by PromoCell (PromoCell C-12731, lot 395Z024) for preadipocyte growth and differentiation into adipocytes. Cell media (PromoCell) was supplemented with 1% penicillin-streptomycin. We maintained the cells at 37 °C in a humidified atmosphere at 5% CO_2. We serum-starved the primary human white adipocytes (HWA) for 16 h using 0.5% fetal calf serum (FCS) in supplemented adipocyte basal medium (PromoCell), prior to treatment with 0.23% fatty acid free bovine serum albumin (BSA, Sigma Aldrich A8806) in media containing 0.5% FCS for 24 h prior to fixation.

Adipocyte fixation and nuclei collection. We rinsed 10 M adherent HWA with serum-free media prior to fixation. We fixed the HWA directly in culture plates with 2% formaldehyde (EMD Millipore 344198) in serum-free adipocyte nutrition media (PromoCell). We incubated cells in fixation medium with rocking at room temperature for 1 min, and then quenched with 1 M ice-cold glycine for a final concentration of 125 mM. After 5 min of rocking incubation at room temperature, we rinsed fixed cells twice with ice-cold PBS. Then we incubated the cells with rocking on ice with ice-cold permeabilization buffer (10 mM Tris–HCl pH 8, 10 mM NaCl, 0.2% Igepal CA-630, Complete EDTA-free protease inhibitor cocktail [Roche])[36] for 30 min. We scraped cells from the culture plate and centrifuged at 2500 × *g* for 5 min at 4 °C to pellet nuclei. The supernatant was discarded and nuclei were flash frozen in liquid nitrogen and put at −80 °C.

Hi–C library preparation. We prepared the Hi–C library as described in Rao et al.[7] with modifications. We processed 10 M HWA nuclei in 5 M cell aliquots, closely following Rao et al.[7] protocol I.a.1 except we digested chromatin with 400U of *Hind*III (New England Biolabs R3104) at 37 °C overnight with shaking (950 rpm). After digestion, we pelleted nuclei with centrifugation at 2500 × *g* for 5 min at 4 °C. We then resuspended nuclei in 265 µl 1× NEBuffer 2 and removed 10% of the cells and kept on ice for a 3 C control[37]. We followed Rao et al.[7] protocol I.a.1 to end-fill and mark with biotin, perform in-nucleus ligation, degrade protein, and perform ethanol precipitation and purification, except we used biotin-14-dCTP (Invitrogen 19518-018) to incorporate biotin during the end-filling step. After quality control to examine Hi-C marking and ligation efficiency, we sheared 5 µg of DNA to 250–550 bp using the Covaris M220 instrument. We performed double size-selection using Agencourt AMPure XP beads (Beckman Coulter A63881) as described in Rao et al.[7] protocol I.a.1.

We immobilized the fragments containing biotin using DYNAL™ MyOne™ Dynabeads™ Streptavidin T1 (Invitrogen 65601) beads following Rao et al.[7] protocol I.a.1. After end-repair and attachment of dATP, we ligated Illumina paired-end adaptors to the bead-bound library following the SureSelect^{XT} user manual for Illumina Paired-End Multiplexed Sequencing (Agilent Technologies). After washing, we resuspended the Hi-C library in 20 µl of 1× Tris buffer and subsequently removed the Streptavidin beads from the DNA by heating at 98 °C for 10 min. We then amplified the adaptor-ligated library using 8 PCR cycles and purified using Agencourt AMPure XP beads, following the SureSelect^{XT} user manual.

Capture Hi-C. The RNA baits were designed in Mifsud et al.[6] for capturing *Hind*III fragments containing gene promoters (Dr. Cameron Osborne kindly shared the exact design). As described in Mifsud et al.[6], 120-mer RNA baits were designed to target both ends of *Hind*III fragments that contain annotated gene promoters (Ensembl promoters of protein-coding, noncoding, antisense, snRNA, miRNA and snoRNA transcripts). The bait sequence was deemed valid if GC content ranged from 25 to 65%, contained <3 consecutive Ns, and was within 330 bp of *Hind*III fragment ends. A total of 550 ng of the Hi–C library was hybridized to the biotinylated RNA baits, captured with DYNAL™ MyOne™ Dynabeads™ Streptavidin T1 beads, and amplified in a post-capture PCR to add indexes, using 12 PCR cycles. The library was sequenced on the Illumina HiSeq 4000 platform.

Capture Hi-C data processing and interaction calling. To ensure all downstream analysis was comparable, we first reduced the number of sequencing reads to match the number used in Mifsud et al.[6] analysis of their CHi-C data. We next processed the sequencing reads with the Hi–C User Pipeline (HiCUP) software[38], aligning reads to the human reference genome (GRCh37/hg19) and using all HiCUP default parameters. We called significant chromosomal interactions with the Capture Hi-C Analysis of Genome Organization (CHiCAGO) software[39], using default parameters, including the threshold of 5 for calling significant interactions. Briefly, the background is estimated by borrowing information across interactions on two separate components of the data: the interactions with baited fragments in the surrounding genomic region are used to model Brownian collisions, which are distance-dependent interactions, and interchromosomal interactions are used to model technical noise. CHiCAGO then employs a weighted *p*-value based on the expected number of interactions at a range of distances[39].

Adipocyte nuclear protein extraction. Nuclear protein was extracted from adipocytes after centrifugation of cells at 200 × *g* for 5 min using a nuclear protein extraction kit as recommended (Active Motif 40010). The quantity of protein extracted was measured with BCA protein assay kit (Pierce 23227).

Electrophoretic mobility shift assay. Oligonucleotide probes (15 bp flanking SNP site for reference or alternate allele) (Supplementary Table 10) with a biotin tag at the 5′ end of the sequence (Integrated DNA Technologies) were incubated with HWA nuclear protein and the working reagent from the Gelshift Chemiluminescent EMSA kit (Active Motif 37341). For competitor assays, an unlabeled probe of the same sequence was added to the reaction mixture at 100 × excess. The reaction was incubated for 30 min at room temperature, and then loaded on a 6% retardation gel (ThermoFisher Scientific EC6365BOX) that was run in 0.5 × TBE buffer. The contents of the gel were transferred to a nylon membrane, and visualized with the chemi-luminescent reagent as recommended. Similarly, we performed the EMSAs with 1 µg purified CTCF protein (Origene TP720882). Supershift assays were performed with 1 µg anti-CTCF antibodies (Santa Cruz sc-15914 and EMD Millipore 07–729) that were incubated on ice with nuclear protein from HWA for 30 min prior to addition of oligonucleotide probes and run on gel.

Study cohort. The study sample consisted of a subset of the participants of the Finnish Metabolic Syndrome in Men (METSIM; $n = 10{,}197$) cohort, described in detail previously[40,41]. The study was approved by the local ethics committee (Research Ethics Committee, Hospital Restrict of Northern Savo) and all participants gave a written informed consent. The METSIM participants are Finnish males recruited at the University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland. The median age of the METSIM participants is 57 years (range: 45–74 years). The biochemical lipid, glucose, and other clinical and metabolic phenotypes were measured, as described previously[40,41]. A random subset of the METSIM men underwent a subcutaneous abdominal adipose needle biopsy, with 335 unrelated individuals (IBD sharing estimated as <0.2 using a genetic relationship matrix calculated in PLINK[42]) analyzed here using RNA-seq.

Identification of *cis*-eQTL SNPs. We processed the METSIM RNA-seq dataset similarly as described in Rodriguez et al.[43]. Briefly, for the METSIM RNA-seq dataset, we isolated total RNA from abdominal subcutaneous adipose needle biopsy using the Qiagen miRNeasy kit. Polyadenylated mRNA was prepared using the Illumina TruSeq RNA Sample Preparation Kit v2 and sequenced using Illumina HiSeq 2000 platform generating paired-end, 50-bp reads. We used STAR[44] to align the reads to the hg19 reference genome, and assembled transcripts using Cufflinks v2.2.1[45]. We filtered genes for those with expression of FPKM > 0 in more than 90% of the samples. Additional details of this dataset have been previously described in Rodriguez et al.[43]. We inverse-normal transformed the FPKMs and adjusted for hidden confounding factors using PEER[46] by removing 22 PEER

factors based on a *cis*-eQTL analysis on chromosome 20 and choosing an optimal number of PEER factors without a loss of statistical power.

To decrease computation time, we prephased the METSIM genotype data, produced using the Illumina HumanOmniExpress BeadChip, by employing SHAPEIT2[47] with the phase 1 version 3 reference panel of the 1000 Genomes Project. We performed imputations with the same reference panel and IMPUTE2[48] with a cosmopolitan imputation approach, which included all populations from the 1000 Genomes Project, to ensure a high accuracy and maximize the number of imputed SNPs. Imputed data were filtered using the quality control inclusion criteria of info ≥0.8, MAF ≥5%, and Hardy–Weinberg equilibrium (HWE) $p > 0.00001$. The *cis*-eQTL analysis was performed using Matrix-eQTL[49]. We classified the variants as in *cis* if they were within 1 Mb of either end of a gene. The first three genetic principal components were included as covariates in the *cis*-eQTL analysis to account for population stratification.

Replication of *cis*-eQTL analysis in GTEx. To ensure robustness of the results, we filtered the *cis*-eQTL SNPs and their target genes detected in the METSIM cohort so that both the *cis*-eQTL SNP and its predicted target gene were replicated in the *cis*-eQTL data by the GTEx Consortium ($n = 277$) for subcutaneous adipose tissue, filtered using their permutation test for significance, which used the adaptive permutation scheme in FastQTL[50] and a permutation test *p*-value threshold equal to the empirical *p*-value of the gene closest to the FDR 5% threshold, as reported by GTEx[12]. Only replicated adipose *cis*-eQTLs and their target genes were used in our downstream analyses.

Heritability of *cis* expression in chromosomal interactions. To investigate the functional importance of open chromatin regions (i.e., DHSs) within chromosomal interactions in adipocytes to heritability of *cis* expression, we used LD-score regression[11]. More specifically, we generated an annotation for each region within 1 Mb of the TSS of every gene with at least 1 significant promoter interaction. Per gene, this annotation consists of marking the variants within a distal fragment within 1 MB of the TSS that interact with the fragment containing the promoter of the gene. We further refined these annotations to the open chromatin regions available for TF binding. Accordingly, we only marked those variants located in regions identified in the union of DNase I hypersensitivity sites (DHSs) from all tissues in the ENCODE and Roadmap Epigenomics project[51]. Since these chromosomal interaction annotations change on a per-gene basis, we could not use the genome-wide overlapping matrix in the original software, which treats the annotations as fixed genome-wide. In our analyses, we generated an average overlapping matrix aggregated across all the regions. Importantly, we tested that this weighted overlap matrix does not qualitatively change the overall enrichment of heritability of gene expression for fixed annotations, such as coding regions (Supplementary Figure 1). These changes amount to altering equations 7 and 8 from Liu et al.[11] as follows (Equation 1 and 2).

Equation 1: Modified equation 8 from Liu et al.[11] using a weighted overlap matrix instead of the genome-wide average.

$$\mathrm{prop}_{h_g^2(C)} = \frac{h_g^2(C)}{h_g^2(\mathrm{total})} = \frac{\sum_C \overline{\tau_C}\overline{M}_{C\cap C}}{\sum_C \overline{\tau_C}\overline{M_C}}$$

$$\text{Where } \overline{M} = \sum_{\text{gene } i}^{N} \frac{M_i}{NSNP_i}$$

In the equation above, C is a given annotation category; $\mathrm{prop}\,h_g^2(C)$ is the proportion of heritability for a given category; $\overline{\tau_c}$ is the regression coefficient for the category; $\overline{M}$ is the average overlap matrix for each local region;M_i is the overlap matrix for each gene in the local region; and $NSNP_i$ is the number of SNPs in each local region.

Equation 2: Modified equation 9 from Liu et al.[11] using the average proportion of SNPs instead of the genome-wide average.

$$\mathrm{enrichment}(C) = \frac{\mathrm{prop}_{h_g}^2(C)}{\mathrm{prop}_{SNPs}(\overline{M})}$$

$$\text{Where } \overline{M} = \sum_{\text{gene } i}^{N} \frac{M_i}{NSNP_i}$$

In the equation above, the variables are as in Equation 1, and $\mathrm{prop\_SNPs}(\overline{M})$ is the proportion of SNPs in the overlap matrix for a given category.

Transcription factor motif enrichment in adipocytes. We used Hypergeometric Optimization of Motif EnRichment (HOMER, v4.9) to investigate the enrichment of known TFs in the open chromatin regions (i.e., DHSs) within chromosomal interactions in adipocytes[9]. As input data, we used chromosomal interactions in adipocytes that interacted with a promoter fragment intersected with the union of all DHSs from ENCODE and Epigenomics Roadmap. We chose to use the DHSs in all cell types since there are no publicly available DHS data in adipocytes or adipose. Furthermore, since we were interested in the TF enrichments in adipocytes, we used $CD34^+$ chromosomal interactions intersected with the union of all DHSs as the background input file[6]. Any regions that were shared between the $CD34^+$ and adipocyte datasets were not considered in this analysis. We considered significant any TFs that were enriched in the DHSs within chromosomal interactions in adipocytes at an FDR of 5%. To ensure our background input file was not biasing the results, we also performed the same analysis with all DHSs not found in adipocyte chromosomal interactions as the background input.

We also assessed predicted differential TF binding using the tool deep learning–based sequence analyzer (DeepSEA)[19], which assesses differential histone modification, TF binding, and DHS profiles using a deep learning-based algorithmic approach and gives a functional significance score at the single nucleotide resolution.

Overlap of *cis*-eQTL SNPs and chromosomal interactions. To investigate functional *cis*-eQTL SNPs, we overlapped the imputed *cis*-eQTL SNPs and their target genes with Capture Hi-C chromosomal interactions by first overlapping the position other end of the looping interaction with the location of the *cis*-eQTL SNP. These were subsequently designated as regulatory element *cis*-eQTL SNPs. Simultaneously, we examined the identity of the predicted target gene for the *cis*-eQTL SNP and the gene involved in the looping interaction for a match. Only when both these criteria were fulfilled, was the *cis*-eQTL SNP defined as a looping *cis*-eQTL SNP and considered for further analyses.

Identification of LD proxies of GWAS SNPs. GWAS variants associated with BMI were obtained from Locke et al.[2], and with lipids and metabolites from Willer et al.[15] and Shin et al.[14]. We identified the *cis*-eQTL SNPs in tight LD ($r^2 > 0.80$) with GWAS variants in the METSIM adipose RNA-seq dataset using PLINK[42] and used them as the LD proxies for BMI, lipid, and metabolite GWAS SNPs. These sets of *cis*-eQTL SNPs were considered as BMI GWAS SNPs, lipid GWAS SNPs, and metabolite GWAS SNPs, respectively. These set of BMI, lipid, and metabolite GWAS SNPs were then overlapped with the looping *cis*-eQTL SNPs to identify all BMI, lipid, and metabolite GWAS SNPs involved in chromosomal interactions acting through distant regulatory elements.

Correlation of BMI with adipose gene expression. The BMI measurements in the METSIM cohort were first adjusted for age, age^2 and then the resulting residuals were inverse normal transformed to reduce the possible outlier effects. Next, we log transformed the FPKM values and then corrected them for 14 technical factors, including the RIN values, batch, percentage of coding reads, 5′ to 3′ bias, and percentage of uniquely mapped reads using Picard tools. The expression levels were correlated with the BMI measurements using Pearson correlation. The *p*-values were corrected for multiple testing for the number of eGenes using the Bonferroni correction (adjusted *p*-value < 0.05). To directly compare the effects sizes and *p*-values obtained for BMI associations in TwinsUK with those in METSIM, we also tested the 42 replicated genes using a linear regression model with BMI and age, age^2 and the 14 technical factors as covariates when compared to a null model without BMI in METSIM (Table 1 and Supplementary Table 6). These models were compared using an F-test.

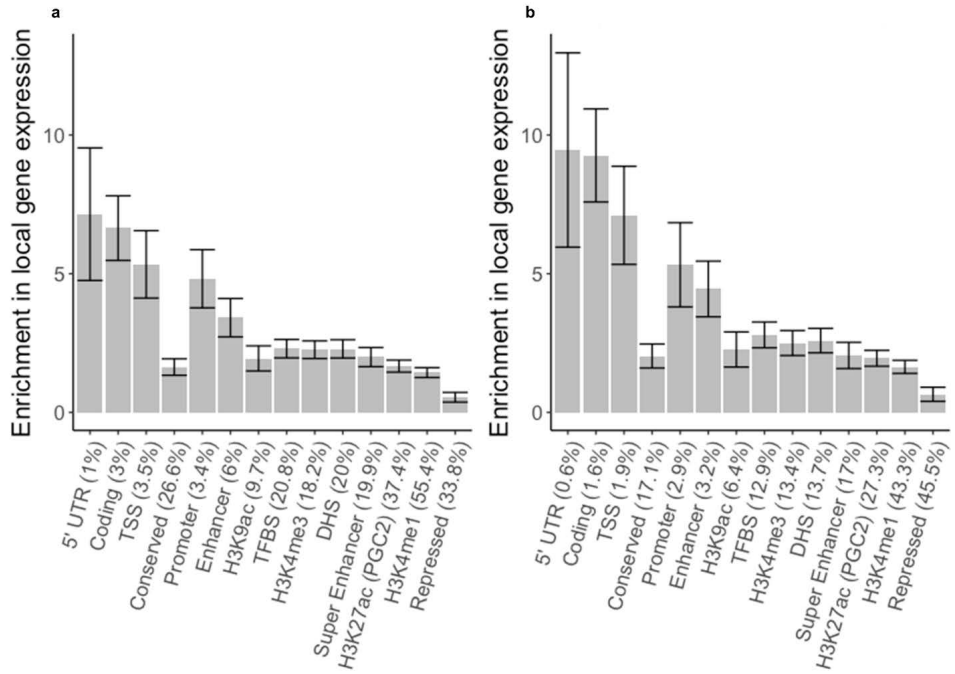
Replication of BMI-adipose gene expression correlation. Association analysis between BMI and adipose expression in the TwinsUK cohort was performed on 720 female twins. RNA-seq was generated as previously described[52] and gene level quantifications were generated to Gencode v19. Association between gene expression level and inversed normalized BMI was tested with a linear mixed effects model (LMEx) implemented using the lme4 package[53]. A full model including BMI was compared to a null model in which the same model was fitted, but with BMI omitted. These models were compared using an F-test. All known technical variables (batch, GC content, insert size mode, and primer index), age, age^2, and family structure were included as covariates in the models. All variables were centered and scaled to unit variance. Four genes were not present in the TwinsUK cohort dataset and we were thus unable to test them for replication, resulting in 54 genes tested for replication. Each replicated gene was examined to determine if effect size direction in TwinsUK and METSIM has the same direction. A Bonferroni corrected *p*-value (adjusted $p < 0.001$) with the same direction of effect as in METSIM was considered as statistical evidence for replication in the TwinsUK.

Data availability. The human primary adipocyte Capture Hi–C data are available at GEO (Accession ID: GSE110619)

Received: 12 September 2017 Accepted: 22 February 2018
Published online: 17 April 2018

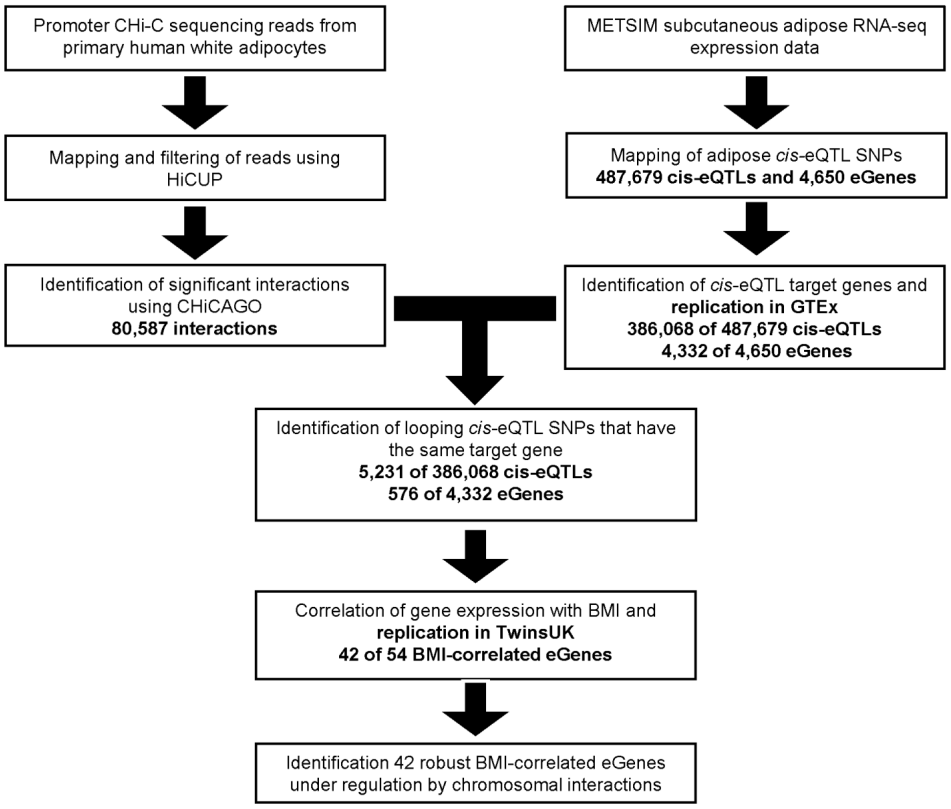
Integration of human adipocyte chromosomal interactions with adipose gene expression prioritizes obesity-related genes from GWAS

Pan et al.



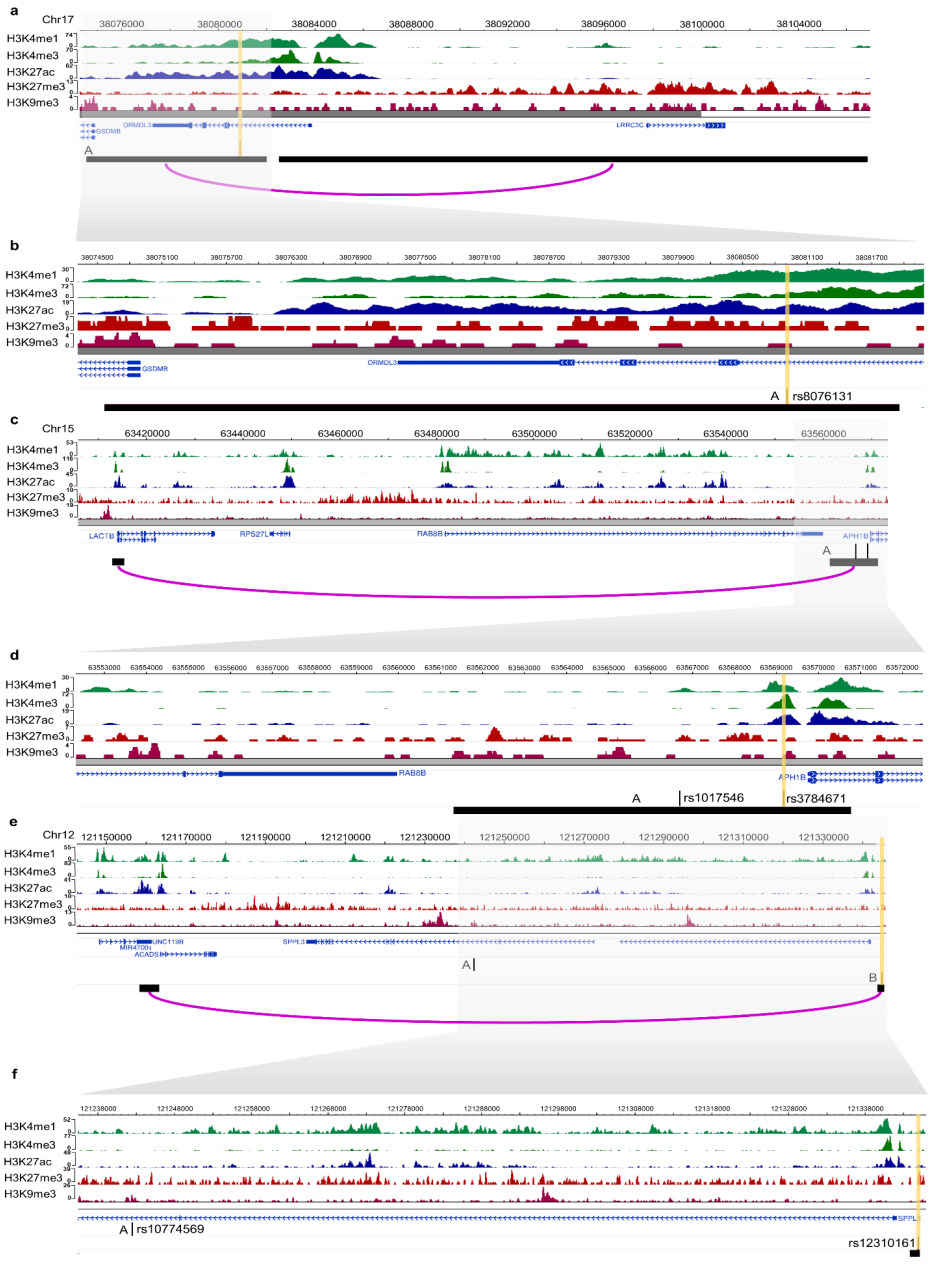
Supplementary Figure 1. Modification to LD Score regression software does not show significant changes when compared with the data obtained using the published version.

Enrichments in local gene expression with error bars for different categories using the LD score regression analysis. For the horizontal axis labels, the value in parentheses shows the percentage of SNPs contained within the respective annotation category that contributed to the enrichment calculation (for the full data on all 52 baseline annotation categories, see Supplementary Table 3-4). Error bars represent jackknife standard errors around the estimates of enrichment. (a) Enrichment in local gene expression for the modified LD Score regression software. (b) Enrichment in local gene expression for the original, unmodified LD Score regression software.



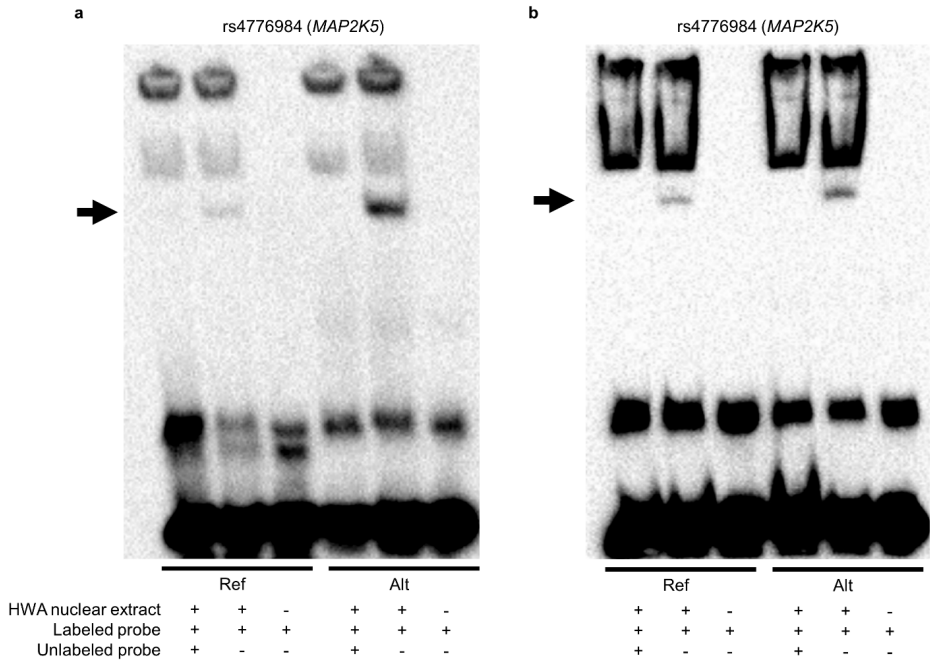
Supplementary Figure 2. Overview of the study design targeted to identify causal and reactive BMI-correlated genes.

Flow chart showing the data processing and analysis pipeline of the promoter Capture Hi-C in primary human white adipocytes (HWA) (the left side); adipose RNA-sequencing followed by *cis*-eQTL mapping (the right side); and the integration of these genomics data (in the middle) to identify eGenes correlated with BMI.


a
Chr17
38076000
38080000
38084000
38088000
38092000
38096000
38100000
38104000
H3K4me1
H3K4me3
H3K27ac
H3K27me3
H3K9me3
GSDMB
ORMDL3
LRRC3C
A
b
38074500
38075100
38075700
38076300
38076900
38077500
38078100
38078700
38079300
38079900
38080500
38081100
38081700
H3K4me1
H3K4me3
H3K27ac
H3K27me3
H3K9me3
GSDMB
ORMDL3
A
rs8076131
c
Chr15
63420000
63440000
63460000
63480000
63500000
63520000
63540000
63560000
H3K4me1
H3K4me3
H3K27ac
H3K27me3
H3K9me3
LACTB
RPS27L
RAB8B
APH1B
A
d
63553000
63554000
63555000
63556000
63557000
63558000
63559000
63560000
63561000
63562000
63563000
63564000
63565000
63566000
63567000
63568000
63569000
63570000
63571000
63572000
H3K4me1
H3K4me3
H3K27ac
H3K27me3
H3K9me3
RAB8B
APH1B
A
rs1017546
rs3784671
e
Chr12
121150000
121170000
121190000
121210000
121230000
121250000
121270000
121290000
121310000
121330000
H3K4me1
H3K4me3
H3K27ac
H3K27me3
H3K9me3
UNC119B
MIR4700
ACAD5
SPPL3
A
B
f
121238000
121248000
121258000
121268000
121278000
121288000
121298000
121308000
121318000
121328000
121338000
H3K4me1
H3K4me3
H3K27ac
H3K27me3
H3K9me3
SPPL
A
rs10774569
rs12310161

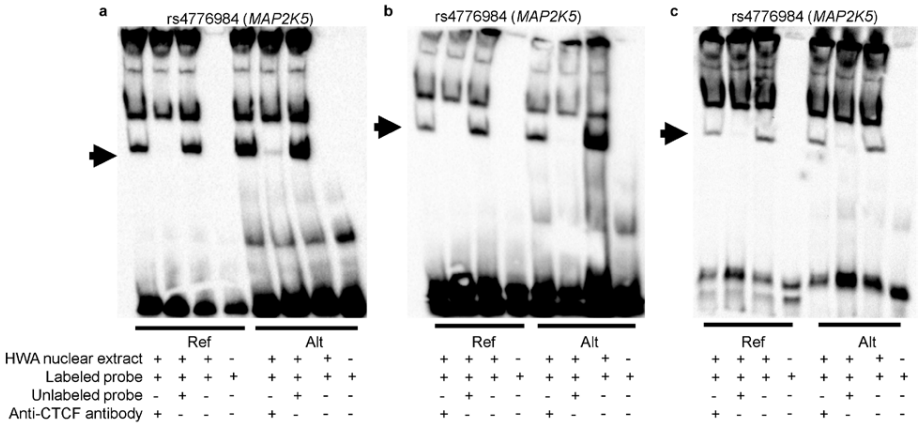
Supplementary Figure 3. Promoter Capture Hi-C enables refinement of the GWAS loci that colocalizes with *cis*-eQTLs interacting with the target gene promoter of *ORMDL3*, *LACTB*, and *ACADS*.

Genomic landscape of the lipid GWAS locus, *ORMDL3* (panels a, b), metabolite GWAS locus, *LACTB* (panels c, d), and metabolite GWAS locus, *ACADS* (panels e, f), modified from the WashU Genome Browser to show the histone mark calls from ChIP-seq data; gene transcripts; promoter and eQTL *Hin*dIII fragments that interact in primary human white adipocytes (HWA); and GWAS SNP (A, the rs number indicated in the magnified box) or their LD proxies if applicable (B, r^2>0.80) located in the interacting *Hin*dIII fragment. The vertical yellow band highlights the significantly influential variant (the rs number is indicated in the magnified box). (a) Genomic landscape containing *ORMDL3* and the interacting lipid GWAS SNP. (b) Magnification of the boxed region in (a). (c) Genomic landscape containing *LACTB* and the interacting metabolite GWAS SNPs. (d) Magnification of the boxed region in (c). (e) Genomic landscape containing *ACADS* and the interacting *cis*-eQTLs and corresponding metabolite GWAS SNP. (f) Magnification of the boxed region in (e).



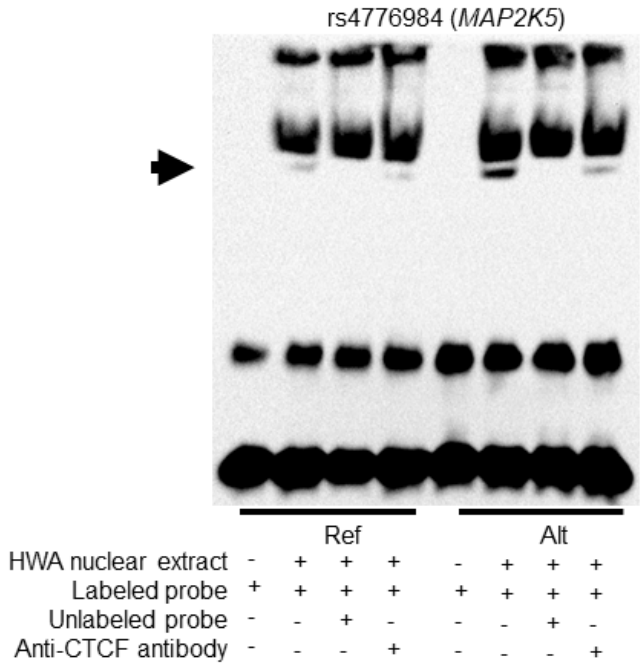
Supplementary Figure 4. Two independent replicates for the Electrophoretic mobility shift assay (EMSA) data show increased binding of nuclear protein extracted from primary human white adipocytes (HWA) to the alternate allele when compared to the reference allele of the *MAP2K5 cis*-eQTL SNP rs4776984.

Biotinylated (labeled probe) 31-bp oligonucleotide complexes with +/-15 bp flanking the reference or alternate allele for variant rs4776984 were incubated with nuclear protein extracted from primary HWA and resolved on a 6% polyacrylamide gel. Competitor assays were performed by incubating the reaction with 100X excess of unlabeled (no biotin) oligonucleotide complexes with identical sequence. Arrow denotes specific binding of HWA nuclear protein to reference (left) and alternate (right) allele. (a) First replicate of the EMSA for rs4776984. (b) Second replicate of the EMSA for rs4776984.



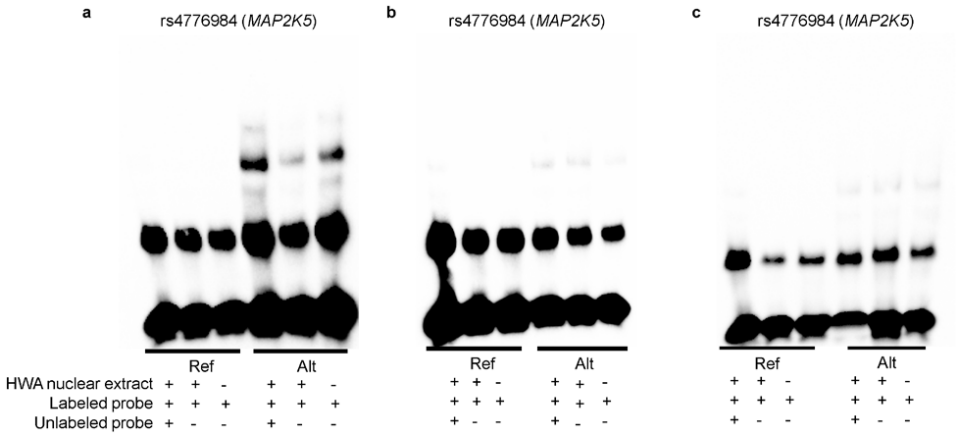
Supplementary Figure 5. Three independent replicates for the Electrophoretic mobility shift assay (EMSA) do not show a supershift when using antibody against CTCF and nuclear protein extracted from primary human white adipocytes (HWA) at the *MAP2K5* *cis*-eQTL SNP rs4776984.

Biotinylated (labeled probe) 31-bp oligonucleotide complexes with +/-15 bp flanking the reference or alternate allele for variant rs4776984 were incubated with nuclear protein extracted from primary HWA and resolved on a 6% polyacrylamide gel. Competitor assays were performed by incubating the reaction with 100X excess of unlabeled (no biotin) oligonucleotide complexes with identical sequence. Arrow denotes specific binding of HWA nuclear protein to reference (left) and alternate (right) allele. Supershift assays were performed with 1µg anti-CTCF antibodies (Santa Cruz sc-15914). (a) First replicate of the supershift EMSA for rs4776984. (b) Second replicate of the supershift EMSA for rs4776984. (c) Third replicate of the supershift EMSA for rs4776984.



Supplementary Figure 6. The Electrophoretic mobility shift assay (EMSA) does not show a supershift when using a different antibody against CTCF and nuclear protein extracted from primary human white adipocytes (HWA) at the *MAP2K5 cis*-eQTL SNP rs4776984.

Biotinylated (labeled probe) 31-bp oligonucleotide complexes with +/-15 bp flanking the reference or alternate allele for variant rs4776984 were incubated with nuclear protein extracted from primary HWA and resolved on a 6% polyacrylamide gel. Competitor assays were performed by incubating the reaction with 100X excess of unlabeled (no biotin) oligonucleotide complexes with identical sequence. Arrow denotes specific binding of HWA nuclear protein to reference (left) and alternate (right) allele. Supershift assays were performed with 1μg anti-CTCF antibodies (EMD Millipore 07-729).



Supplementary Figure 7. Three independent replicates for the Electrophoretic mobility shift assay (EMSA) do not show specific binding using purified CTCF protein at the *MAP2K5 cis*-eQTL SNP rs4776984.

Biotinylated (labeled probe) 31-bp oligonucleotide complexes with +/-15 bp flanking the reference or alternate allele for variant rs4776984 were incubated with purified CTCF protein (Origene TP720882) and resolved on a 6% polyacrylamide gel in our EMSA experiment. Competitor assays were performed by incubating the reaction with 100X excess of unlabeled (no biotin) oligonucleotide complexes with identical sequence. The reference allele is on the left and alternate allele on the right. (a) First replicate of the EMSA for rs4776984. (b) Second replicate of the EMSA for rs4776984. (c) Third replicate of the EMSA for rs4776984.

Supplementary Table 1. Parameters used for identification of novel *cis*-eQTL and looping interactions

cis-eQTL discovery	METSIM (n=335)
Type of genetic data	Illumina Omni Express
# *cis*-eQTLSNPs with the same target gene and beta direction replicated in subcutaneous adipose GTEx data	386,068
# PEER factors corrected	22
# Genetic principal components corrected	3
Minor allele frequency (MAF)	> 5%
Type of expression data	RNA-seq
Normalization technique	Inverse normal transformation of FPKMs
FDR significance threshold for cis-eQTL SNPs	< 5%
# of *cis*-eQTL target genes with looping interactions	4,332
Promoter capture Hi-C	Primary human white adipocytes
# reads from sequencing	138,217,259
# uniquely aligned paired reads	101,187,918
# valid pairs of reads after capture Hi-C specific filtering by HiCUP	88,583,089
# significant looping interaction pairs identified from CHiCAGO	80,567
# METSIM genes in looping interaction pairs	10,083

Supplementary Table 2. Histone mark enrichment in looping *Hin*dIII fragments in primary HWA

Histone mark	Base pairs of feature enrichment in looping *Hin*dIII fragments	Base pairs of feature enrichment in random *Hin*dII fragments*	Standard deviation	*p*-value†
H3K4me1	42181	39278.45	56.94	$<2.2x10^{-16}$
H3K4me3	42347	39502.39	55.86	$<2.2x10^{-16}$
H3K27ac	42095	39597.19	49.24	$<2.2x10^{-16}$
H3K27me3	42813	40529.33	43.78	$<2.2x10^{-16}$
H3K9me3	41222	39408.19	54.07	$<2.2x10^{-16}$
DHS	35578	30547.74	89.82	$<2.2x10^{-16}$

*Random *Hin*dIII fragments were controlled for distance away from the target promoter when selected.

†*p*-value computed from Pearson's chi-squared test.

Supplementary Table 3. Adipocyte chromosomal interactions are enriched for 30 transcription factors (adjusted *p*<0.05) when compared to CD34+ chromosomal interactions

Motif logo	Motif name	*p*-value	Adjusted *p*-value	Number of target sequences with motif (of 189013)	Percent of target sequences with motif	Number of background sequences with motif (of 173261)	Percent of background sequences with motif
	CTCF(Zf)/CD4+-CTCF-ChIP-Seq(Barski_et_al.)/Homer	$1.00x10^{-102}$	0	5746	3.04%	3915.7	2.26%
	BORIS(Zf)/K562-CTCFL-ChIP-Seq(GSE32465)/Homer	$1.00x10^{-53}$	0	6336	3.35%	4760.1	2.75%
	CEBP(bZIP)/ThioMac-CEBPb-ChIP-Seq(GSE21512)/Homer	$1.00x10^{-10}$	0	18925	10.01%	16547.8	9.56%
	Sp5(Zf)/mES-Sp5.Flag-ChIP-Seq(GSE72989)/Homer	$1.00x10^{-7}$	0	16995	8.99%	14950.8	8.64%
	Elk4(ETS)/Hela-Elk4-ChIP-Seq(GSE31477)/Homer	$1.00x10^{-5}$	0.00010	10491	5.55%	9180.3	5.31%
	YY1(Zf)/Promoter/Homer	$1.00x10^{-5}$	0.00030	817	0.43%	638.7	0.37%
	NRF1(NRF)/MCF7-NRF1-ChIP-Seq(Unpublished)/Homer	$1.00x10^{-5}$	0.00040	1163	0.61%	935.2	0.54%
	TEAD2(TEA)/Py2T-Tead2-ChIP-Seq(GSE55709)/Homer	$1.00x10^{-4}$	0.00050	12965	6.86%	11440.4	6.61%
	E2F3(E2F)/MEF-E2F3-ChIP-Seq(GSE71376)/Homer	$1.00x10^{-4}$	0.00070	7993	4.23%	6986.6	4.04%
	Erra(NR)/HepG2-Erra-ChIP-Seq(GSE31477)/Homer	$1.00x10^{-4}$	0.0023	40662	21.50%	36580.6	21.14%
	TEAD(TEA)/Fibroblast-PU.1-ChIP-Seq(Unpublished)/Homer	$1.00x10^{-4}$	0.0026	17798	9.41%	15848.7	9.16%
	Elk1(ETS)/Hela-Elk1-ChIP-Seq(GSE31477)/Homer	$1.00x10^{-4}$	0.0026	10671	5.64%	9422.4	5.45%
	TEAD4(TEA)/Tropoblast-Tead4-ChIP-Seq(GSE37350)/Homer	$1.00x10^{-4}$	0.0026	20464	10.82%	18265.1	10.56%
	GFY(?)/Promoter/Homer	$1.00x10^{-3}$	0.0031	1159	0.61%	950.2	0.55%
	E2F4(E2F)/K562-E2F4-ChIP-Seq(GSE31477)/Homer	$1.00x10^{-3}$	0.0069	5131	2.71%	4475.5	2.59%
	Sp1(Zf)/Promoter/Homer	$1.00x10^{-3}$	0.0084	3622	1.92%	3133.5	1.81%
	NFY(CCAAT)/Promoter/Homer	$1.00x10^{-3}$	0.0084	14237	7.53%	12676.1	7.33%
	Ronin(THAP)/ES-Thap11-ChIP-Seq(GSE51522)/Homer	$1.00x10^{-3}$	0.012	443	0.23%	346.4	0.20%
	Olig2(bHLH)/Neuron-Olig2-ChIP-Seq(GSE30882)/Homer	$1.00x10^{-3}$	0.014	45873	24.26%	41429.3	23.94%
	E2F6(E2F)/Hela-E2F6-ChIP-Seq(GSE31477)/Homer	$1.00x10^{-2}$	0.022	7047	3.73%	6221.3	3.60%
	ZNF143|STAF(Zf)/CUTLL-ZNF143-ChIP-Seq(GSE29600)/Homer	$1.00x10^{-2}$	0.024	5580	2.95%	4906.1	2.84%
	DUX4(Homeobox)/Myoblasts-DUX4.V5-ChIP-Seq(GSE75791)/Homer	$1.00x10^{-2}$	0.029	936	0.49%	777.7	0.45%
	NRF(NRF)/Promoter/Homer	$1.00x10^{-2}$	0.036	1622	0.86%	1382.4	0.80%
	PPARE(NR),DR1/3T3L1-Pparg-ChIP-Seq(GSE13511)/Homer	$1.00x10^{-2}$	0.039	17385	9.19%	15586.6	9.01%
	Pax7(Paired,Homeobox),longest/Myoblast-Pax7-ChIP-Seq(GSE25064)/Homer	$1.00x10^{-2}$	0.040	513	0.27%	415	0.24%
	NFAT(RHD)/Jurkat-NFATC1-ChIP-Seq(Jolma_et_al.)/Homer	$1.00x10^{-2}$	0.045	20039	10.60%	18002.5	10.40%

Supplementary Table 4. LD score enrichments, heritability estimates, and *p*-values using the published LD Score software[9]

Category	Prop. of SNPs	Prop. of h^2	Enrichment	SE	*p*-value
Coding_UCSC	0.02	0.15	9.26	0.9	6.66×10^{-23}
Coding_UCSC.extend.500	0.07	0.28	4.21	0.3	1.27×10^{-27}
Conserved_LindbladToh	0.03	0.15	5.32	0.8	1.26×10^{-8}
Conserved_LindbladToh.extend.500	0.34	0.56	1.64	0.1	4.45×10^{-7}
CTCF_Hoffman	0.02	0.06	2.52	0.5	4.92×10^{-3}
CTCF_Hoffman.extend.500	0.07	0.11	1.58	0.3	2.59×10^{-2}
DGF_ENCODE	0.14	0.34	2.46	0.2	6.87×10^{-10}
DGF_ENCODE.extend.500	0.54	0.77	1.42	0.1	2.60×10^{-5}
DHS_peaks_Trynka	0.11	0.27	2.37	0.3	3.92×10^{-6}
DHS_Trynka	0.17	0.35	2.05	0.2	8.32×10^{-5}
DHS_Trynka.extend.500	0.50	0.72	1.44	0.1	3.56×10^{-5}
Enhancer_Andersson	0.00	0.01	2.60	2.0	4.30×10^{-1}
Enhancer_Andersson.extend.500	0.02	0.03	1.70	0.7	2.87×10^{-1}
Enhancer_Hoffman	0.06	0.15	2.27	0.3	6.34×10^{-5}
Enhancer_Hoffman.extend.500	0.16	0.30	1.93	0.2	2.51×10^{-7}
FetalDHS_Trynka	0.09	0.28	3.25	0.3	3.78×10^{-12}
FetalDHS_Trynka.extend.500	0.29	0.48	1.69	0.2	2.28×10^{-8}
H3K27ac_Hnisz	0.39	0.65	1.64	0.1	2.77×10^{-8}
H3K27ac_Hnisz.extend.500	0.43	0.68	1.58	0.1	9.46×10^{-8}
H3K27ac_PGC2	0.27	0.53	1.95	0.1	2.57×10^{-11}
H3K27ac_PGC2.extend.500	0.34	0.61	1.80	0.1	8.44×10^{-1}
H3K4me1_peaks_Trynka	0.18	0.35	1.98	0.2	2.26×10^{-5}
H3K4me1_Trynka	0.43	0.71	1.64	0.1	5.61×10^{-8}
H3K4me1_Trynka.extend.500	0.61	0.86	1.41	0.1	6.55×10^{-6}
H3K4me3_peaks_Trynka	0.04	0.11	2.55	0.6	1.35×10^{-3}
H3K4me3_Trynka	0.14	0.35	2.59	0.2	5.48×10^{-13}
H3K4me3_Trynka.extend.500	0.26	0.49	1.88	0.1	4.52×10^{-10}
H3K9ac_peaks_Trynka	0.04	0.12	3.03	0.5	1.03×10^{-5}
H3K9ac_Trynka	0.13	0.36	2.79	0.2	1.07×10^{-14}
H3K9ac_Trynka.extend.500	0.23	0.50	2.15	0.2	1.18×10^{-14}
Intron_UCSC	0.39	0.40	1.00	0.1	9.63×10^{-1}
Intron_UCSC.extend.500	0.40	0.51	1.27	0.1	3.63×10^{-3}
PromoterFlanking_Hoffman	0.01	0.04	4.43	1.1	1.82×10^{-3}
PromoterFlanking_Hoffman.extend.500	0.03	0.13	3.70	0.4	3.39×10^{-12}
Promoter_UCSC	0.03	0.14	4.45	0.5	6.18×10^{-12}
Promoter_UCSC.extend.500	0.04	0.18	4.51	0.4	1.01×10^{-20}
Repressed_Hoffman	0.45	0.30	0.65	0.1	5.99×10^{-20}
Repressed_Hoffman.extend.500	0.71	0.47	0.65	0.1	8.43×10^{-6}
SuperEnhancer_Hnisz	0.17	0.35	2.03	0.2	1.79×10^{-5}
SuperEnhancer_Hnisz.extend.500	0.17	0.34	1.97	0.2	1.57×10^{-6}
TFBS_ENCODE	0.13	0.34	2.50	0.2	2.61×10^{-11}
TFBS_ENCODE.extend.500	0.35	0.49	1.42	0.1	6.98×10^{-4}
Transcribed_Hoffman	0.35	0.38	1.07	0.1	5.34×10^{-1}
Transcribed_Hoffman.extend.500	0.77	0.74	0.97	0.1	6.53×10^{-1}
TSS_Hoffman	0.02	0.13	7.10	0.9	5.15×10^{-12}
TSS_Hoffman.extend.500	0.04	0.19	5.23	0.5	6.73×10^{-15}
UTR_3_UCSC	0.01	0.08	6.88	0.9	4.11×10^{-10}
UTR_3_UCSC.extend.500	0.03	0.13	4.48	0.6	4.99×10^{-10}
UTR_5_UCSC	0.01	0.05	9.46	1.8	1.36×10^{-6}
UTR_5_UCSC.extend.500	0.03	0.15	5.10	0.5	1.19×10^{-14}
WeakEnhancer_Hoffman	0.02	0.04	1.92	0.6	1.02×10^{-1}
WeakEnhancer_Hoffman.extend.500	0.09	0.17	1.92	0.2	1.58×10^{-4}

Supplementary Table 5. LD score enrichments, heritability estimates, and *p*-values after modification of the LD score software

Category	Prop. of SNPs	Prop. of h^2	Enrichment	SE	*p*-value
Coding_UCSC	0.03	0.20	6.64	0.6	$3.44x10^{-22}$
Coding_UCSC.extend.500	0.12	0.39	3.19	0.2	$5.65x10^{-27}$
Conserved_LindbladToh	0.03	0.16	4.82	0.5	$3.25x10^{-13}$
Conserved_LindbladToh.extend.500	0.40	0.63	1.59	0.1	$3.09x10^{-10}$
CTCF_Hoffman	0.03	0.07	2.18	0.4	$1.54x10^{-3}$
CTCF_Hoffman.extend.500	0.10	0.14	1.47	0.2	$7.92x10^{-3}$
DGF_ENCODE	0.18	0.40	2.19	0.2	$3.73x10^{-12}$
DGF_ENCODE.extend.500	0.64	0.83	1.29	0.1	$1.23x10^{-4}$
DHS_peaks_Trynka	0.14	0.31	2.32	0.2	$2.94x10^{-9}$
DHS_Trynka	0.20	0.40	1.99	0.2	$7.84x10^{-9}$
DHS_Trynka.extend.500	0.56	0.76	1.34	0.1	$2.98x10^{-5}$
Enhancer_Andersson	0.01	0.01	2.21	1.4	$3.80x10^{-1}$
Enhancer_Andersson.extend.500	0.03	0.04	1.48	0.5	$2.89x10^{-1}$
Enhancer_Hoffman	0.10	0.19	1.95	0.2	$2.96x10^{-5}$
Enhancer_Hoffman.extend.500	0.23	0.38	1.66	0.1	$5.12x10^{-7}$
FetalDHS_Trynka	0.11	0.31	2.92	0.2	$4.30x10^{-15}$
FetalDHS_Trynka.extend.500	0.34	0.53	1.58	0.1	$1.68x10^{-7}$
H3K27ac_Hnisz	0.54	0.75	1.39	0.1	$7.15x10^{-6}$
H3K27ac_Hnisz.extend.500	0.57	0.77	1.35	0.1	$2.25x10^{-5}$
H3K27ac_PGC2	0.37	0.62	1.67	0.1	$7.02x10^{-10}$
H3K27ac_PGC2.extend.500	0.46	0.71	1.54	0.1	$6.89x10^{-9}$
H3K4me1_peaks_Trynka	0.24	0.41	1.74	0.2	$4.82x10^{-7}$
H3K4me1_Trynka	0.55	0.80	1.44	0.1	$9.82x10^{-7}$
H3K4me1_Trynka.extend.500	0.74	0.92	1.25	0.1	$3.15x10^{-4}$
H3K4me3_peaks_Trynka	0.06	0.15	2.36	0.3	$6.40x10^{-5}$
H3K4me3_Trynka	0.20	0.46	2.29	0.2	$4.79x10^{-15}$
H3K4me3_Trynka.extend.500	0.35	0.60	1.71	0.1	$2.44x10^{-11}$
H3K9ac_peaks_Trynka	0.07	0.17	2.50	0.3	$2.10x10^{-6}$
H3K9ac_Trynka	0.21	0.48	2.30	0.2	$5.70x10^{-15}$
H3K9ac_Trynka.extend.500	0.36	0.64	1.77	0.1	$4.00x10^{-13}$
Intron_UCSC	0.47	0.43	0.91	0.1	$2.60x10^{-1}$
Intron_UCSC.extend.500	0.49	0.60	1.21	0.1	$4.05x10^{-3}$
PromoterFlanking_Hoffman	0.01	0.04	3.30	0.8	$2.11x10^{-3}$
PromoterFlanking_Hoffman.extend.500	0.05	0.15	2.91	0.3	$3.11x10^{-12}$
Promoter_UCSC	0.06	0.20	3.41	0.4	$2.98x10^{-12}$
Promoter_UCSC.extend.500	0.07	0.25	3.43	0.3	$2.08x10^{-20}$
Repressed_Hoffman	0.34	0.18	0.55	0.1	$2.20x10^{-7}$
Repressed_Hoffman.extend.500	0.56	0.32	0.57	0.1	$1.10x10^{-15}$
SuperEnhancer_Hnisz	0.27	0.43	1.63	0.2	$1.92x10^{-5}$
SuperEnhancer_Hnisz.extend.500	0.27	0.43	1.59	0.1	$2.07x10^{-5}$
TFBS_ENCODE	0.18	0.41	2.26	0.2	$2.21x10^{-15}$
TFBS_ENCODE.extend.500	0.43	0.60	1.38	0.1	$7.13x10^{-5}$
Transcribed_Hoffman	0.43	0.43	1.02	0.1	$7.98x10^{-1}$
Transcribed_Hoffman.extend.500	0.76	0.72	0.94	0.1	$3.15x10^{-1}$
TSS_Hoffman	0.03	0.19	5.34	0.6	$9.97x10^{-13}$
TSS_Hoffman.extend.500	0.06	0.26	3.97	0.4	$1.84x10^{-15}$
UTR_3_UCSC	0.02	0.10	5.06	0.6	$2.30x10^{-10}$
UTR_3_UCSC.extend.500	0.05	0.16	3.41	0.4	$2.80x10^{-10}$
UTR_5_UCSC	0.01	0.07	7.15	1.2	$2.72x10^{-7}$
UTR_5_UCSC.extend.500	0.05	0.20	3.99	0.4	$4.37x10^{-16}$
WeakEnhancer_Hoffman	0.03	0.05	1.67	0.4	$7.95x10^{-2}$
WeakEnhancer_Hoffman.extend.500	0.13	0.22	1.62	0.2	$3.10x10^{-4}$

Supplementary Table 6. Fifty-four eGenes in METSIM, including the 42 genes replicated for correlation with BMI and effect direction in TwinsUK

		Pearson		Linear regression					
		METSIM[†]		METSIM[‡]			TwinsUK[§]		
Gene	Chr[#]	Effect size (r)	p-value	Effect size (β)	SE	p-value	Effect size (β)	SE	p-value
ADH1B	4	-0.45	$7.40x10^{-18}$	-0.21	0.02	$1.68x10^{-20}$	-0.58	0.03	$4.47x10^{-71}$
ORMDL3[*]	17	-0.45	$8.57x10^{-18}$	-0.16	0.02	$2.06x10^{-20}$	-0.58	0.03	$2.65x10^{-70}$
AKR1C3	10	0.33	$4.78x10^{-10}$	0.13	0.02	$2.95x10^{-11}$	0.49	0.03	$5.19x10^{-54}$
CMTM3	16	0.41	$4.32x10^{-15}$	0.087	0.01	$3.84x10^{-17}$	0.50	0.03	$6.64x10^{-52}$
LPIN1	2	-0.38	$1.49x10^{-13}$	-0.14	0.02	$2.27x10^{-15}$	-0.47	0.03	$2.38x10^{-44}$
RNF157	17	-0.29	$5.19x10^{-8}$	-0.096	0.02	$5.87x10^{-9}$	-0.47	0.03	$8.86x10^{-42}$
MYOF	10	0.32	$1.07x10^{-9}$	0.086	0.01	$7.37x10^{-11}$	0.46	0.03	$2.59x10^{-40}$
NAA40	11	0.28	$1.81x10^{-7}$	0.052	0.009	$2.67x10^{-8}$	0.46	0.03	$4.00x10^{-40}$
TMEM165	4	0.33	$2.45x10^{-9}$	0.045	0.007	$1.84x10^{-10}$	0.45	0.03	$3.52x10^{-37}$
RFFL	11	0.27	$1.02x10^{-6}$	0.035	0.006	$1.84x10^{-8}$	0.43	0.03	$5.67x10^{-37}$
TMCO6	5	-0.28	$9.23x10^{-8}$	-0.060	0.01	$1.18x10^{-8}$	-0.44	0.03	$5.04x10^{-35}$
SCRN2	17	-0.38	$2.23x10^{-13}$	-0.10	0.01	$3.79x10^{-15}$	-0.38	0.03	$5.32x10^{-35}$
CSGALNACT1	8	0.24	$1.00x10^{-5}$	0.047	0.01	$2.04x10^{-6}$	0.42	0.03	$1.41x10^{-31}$
TAPBP	6	0.25	$6.71x10^{-6}$	0.047	0.02	$1.60x10^{-6}$	0.32	0.03	$1.52x10^{-29}$
CLN8	8	0.32	$4.50x10^{-9}$	0.044	0.007	$3.67x10^{-10}$	0.36	0.03	$4.41x10^{-29}$
DRAM1	12	0.30	$1.87x10^{-8}$	0.050	0.008	$1.80x10^{-9}$	0.40	0.03	$5.94x10^{-29}$
WNT2B	1	0.25	$2.44x10^{-6}$	0.026	0.005	$4.90x10^{-7}$	0.38	0.03	$1.41x10^{-27}$
S100A1	1	-0.27	$2.52x10^{-7}$	-0.20	0.04	$3.59x10^{-8}$	-0.38	0.03	$3.69x10^{-26}$
RPS6KL1	14	0.26	$2.54x10^{-6}$	0.060	0.01	$5.25x10^{-7}$	0.34	0.03	$3.27x10^{-25}$
SLC16A7	12	-0.26	$3.47x10^{-6}$	-0.068	0.01	$7.60x10^{-7}$	-0.30	0.03	$2.08x10^{-23}$
ZNF592	15	-0.27	$8.26x10^{-7}$	-0.037	0.007	$1.40x10^{-7}$	-0.33	0.03	$2.10x10^{-23}$
MFSD1	3	0.31	$8.31x10^{-9}$	0.069	0.01	$6.70x10^{-10}$	0.35	0.04	$2.82x10^{-22}$
HYI	1	-0.31	$6.45x10^{-9}$	-0.11	0.02	$5.52x10^{-10}$	-0.29	0.03	$5.95x10^{-22}$
ANXA4	2	0.24	$1.04x10^{-5}$	0.045	0.009	$2.52x10^{-6}$	0.35	0.04	$1.20x10^{-21}$
RAB30	11	0.24	$8.19x10^{-6}$	0.040	0.008	$1.98x10^{-6}$	0.31	0.03	$1.16x10^{-20}$
PLD1	3	-0.28	$2.26x10^{-7}$	-0.050	0.009	$3.24x10^{-8}$	-0.32	0.03	$7.95x10^{-20}$
MYO5A	15	0.30	$3.20x10^{-8}$	0.049	0.008	$3.41x10^{-9}$	0.32	0.04	$4.61x10^{-19}$
ACADS[*]	12	-0.37	$2.91x10^{-12}$	-0.085	0.01	$7.12x10^{-14}$	-0.24	0.03	$6.65x10^{-19}$
SCAI	9	-0.28	$1.81x10^{-7}$	-0.034	0.006	$2.50x10^{-8}$	-0.27	0.03	$1.42x10^{-18}$
HLA-DRB1	6	0.25	$3.53x10^{-6}$	0.14	0.03	$7.83x10^{-7}$	0.31	0.03	$2.09x10^{-18}$
LACTB[*]	15	0.30	$1.67x10^{-8}$	0.069	0.01	$1.40x10^{-9}$	0.32	0.04	$4.94x10^{-18}$
GPHN	14	-0.43	$7.51x10^{-17}$	-0.11	0.01	$3.20x10^{-19}$	-0.29	0.03	$4.28x10^{-17}$
MPHOSPH8	13	-0.24	$8.25x10^{-6}$	-0.033	0.007	$2.02x10^{-6}$	-0.23	0.04	$3.97x10^{-11}$
MAP2K5[*]	15	-0.25	$7.83x10^{-6}$	-0.039	0.008	$1.90x10^{-6}$	-0.21	0.03	$3.81x10^{-10}$
RRNAD1	1	-0.24	$1.05x10^{-5}$	-0.032	0.007	$2.30x10^{-6}$	-0.19	0.03	$3.14x10^{-9}$
CCDC50	3	-0.33	$1.16x10^{-9}$	-0.059	0.009	$7.24x10^{-11}$	-0.18	0.03	$9.93x10^{-9}$
RAD54L2	3	-0.25	$2.32x10^{-6}$	-0.030	0.006	$4.70x10^{-7}$	-0.20	0.04	$2.78x10^{-8}$
SCMH1	1	-0.32	$1.11x10^{-9}$	-0.047	0.007	$7.55x10^{-11}$	-0.19	0.03	$3.85x10^{-8}$
ATP7B	13	-0.26	$6.30x10^{-7}$	-0.040	0.007	$1.11x10^{-7}$	-0.20	0.04	$7.22x10^{-8}$
CYP7B1	8	0.24	$6.90x10^{-6}$	0.047	0.01	$1.64x10^{-6}$	0.19	0.03	$1.07x10^{-7}$
RERE	1	-0.24	$1.02x10^{-5}$	-0.031	0.006	$2.61x10^{-6}$	-0.17	0.04	5.39x10-6
RPAP1	15	-0.35	$1.86x10^{-10}$	-0.042	0.006	$8.82x10^{-12}$	-0.14	0.03	$9.58x10^{-6}$
ARHGEF7	13	-0.35	$4.12x10^{-11}$	-0.050	0.007	$1.60x10^{-12}$	0.022	0.04	NS[‖]
NCKIPSD	3	-0.34	$3.28x10^{-10}$	-0.067	0.01	$1.51x10^{-11}$	-0.042	0.03	NS[‖]
NDUFS2	1	-0.24	$9.38x10^{-6}$	-0.029	0.006	$2.17x10^{-6}$	-0.048	0.03	NS[‖]
REEP1	2	-0.24	$6.38x10^{-6}$	-0.033	0.007	$1.45x10^{-6}$	0.022	0.04	NS[‖]
RGCC	13	-0.25	$2.81x10^{-6}$	-0.076	0.01	$4.91x10^{-7}$	0.087	0.03	NS[‖]
SETD6	16	-0.27	$3.99x10^{-7}$	-0.041	0.007	$6.30x10^{-8}$	-0.047	0.04	NS[‖]
SLC35A3	12	-0.26	$1.13x10^{-6}$	-0.024	0.004	$2.15x10^{-7}$	0.043	0.03	NS[‖]
SPAG7	17	-0.26	$1.21x10^{-6}$	-0.035	0.007	$2.27x10^{-7}$	-0.076	0.03	NS[‖]
NUDCD3	7	-0.34	$7.00x10^{-10}$	-0.032	0.005	$4.17x10^{-11}$	NA[¶]	NA[¶]	NA[¶]
RP11-387H17.4	17	-0.40	$4.40x10^{-14}$	-0.26	0.03	$4.74x10^{-16}$	NA[¶]	NA[¶]	NA[¶]
RSBN1L-AS1	7	-0.36	$1.65x10^{-11}$	-0.056	0.007	$5.73x10^{-13}$	NA[¶]	NA[¶]	NA[¶]
TUBB2B	6	0.34	$1.01x10^{-10}$	0.14	0.02	$4.51x10^{-12}$	NA[¶]	NA[¶]	NA[¶]

[*]GWAS gene.

[†]Effect size (r, Pearson rho) and *p*-value calculated from Pearson correlation between gene expression and BMI (see Methods).

[‡]Effect size, standard error (SE), and *p*-value calculated using a linear regression model with BMI and age, age^2 and the 14 technical factors as co-variates when compared to a null model without BMI. These models were compared using an F-test (see Methods).

[§]Effect size, standard error (SE), and *p*-value calculated from linear mixed effects model. A full model including BMI was compared to a null model in which the same model was fitted, but with the phenotype (BMI) omitted. These models were compared using an F-test (see Methods).

[‖]Adjusted p-value > $9.26x10^{-4}$.

[¶]Value not applicable due to inability to test for replication in TwinsUK cohort.

[#]Chr indicates chromosome.

Supplementary Table 7. The 42 replicated BMI-correlated eGenes show significant enrichment for metabolic and inflammatory pathways using KEGG pathway analysis as implemented in WebGestalt[13]

KEGG Pathway Name	Ratio of Enrichment	Number of Genes	Genes in Pathway	*p*-value	Adjusted *p*-value*
Fatty acid metabolism	18.76	2	*ACADS ADH1B*	0.0051	0.010
Metabolism of xenobiotics by cytochrome P450	21.78	2	*AKR1C3 ADH1B*	0.0038	0.010
Steroid hormone biosynthesis	30.69	2	*AKR1C3 CYP7B1*	0.0019	0.010
Antigen processing and presentation	11.85	2	*HLA-DRB1 TAPBP*	0.012	0.019

**p*-value adjusted using Benjamini-Hochberg correction for multiple testing.

Supplementary Table 8. DeepSEA analysis of the variants in the MAP2K5 locus supports the functionality of the looping *cis*-eQTL SNP rs4776984.

SNP ID	Chr	Position	Ref	Alt	DeepSEA score
rs4776984	chr15	68118194	A	C	2.36×10^{-3}
rs4776982	chr15	68114974	A	G	3.90×10^{-2}
rs4492996	chr15	68113240	A	G	7.16×10^{-2}
rs4776990	chr15	68137364	C	T	1.09×10^{-1}
rs28742003	chr15	68127769	C	T	1.30×10^{-1}
rs28427879	chr15	68124256	G	T	1.98×10^{-1}

Supplementary Table 9. Significant CHiCAGO interaction and replication scores from a separate HWA Capture Hi-C experiment verify the looping *cis*-eQTLs for the four identified obesity-related loci.

Other End	Baited Fragment	Target Gene	Looping *cis*-eQTL	CHiCAGO score	Replication score
chr15,67834655,67840760	chr15,68111739,68138337	MAP2K5	rs4476984	5.05	6.15
chr17,38082534,38106859	chr17,38074576,38081958	ORMDL3	rs8076131	6.35	6.73
chr15,63413071,63415370	chr15,63561331,63570763	LACTB	rs3784671	6.65	13.92
chr12,121158545,121162946	chr12,121343847,121345146	ACADS	rs10774569	5.29	6.62

Supplementary Table 10. DNA oligonucleotides used for electrophoretic mobility shift assay.

DNA oligonucleotides	Sequence (5' -> 3') for positive and negative strand
Reference allele – A (positive) biotinylated probe*	GCGCGCCCAACTCGGAGCGCCCTGCTGGGCG
Reference allele – A (negative) biotinylated probe	CGCCCAGCAGGGCGCTCCGAGTTGGGCGCGC
Alternate allele – C (positive) biotinylated probe*	GCGCGCCCAACTCGGCGCGCCCTGCTGGGCG
Alternate allele – C (negative) biotinylated probe	CGCCCAGCAGGGCGCGCCGAGTTGGGCGCGC

*Biotinylated probes were created by adding biotin to the 5' end of positive strand probes.

References

1. Gregg, E. W. & Shaw, J. E. Health effects of overweight and obesity in 195 countries over 25 years. *N. Engl. J. Med.* **377**, 13–27 (2017).
2. Locke, A., Kahali, B., Berndt, S., Justice, A. & Pers, T. Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
3. Claussnitzer, M. et al. FTO obesity variant circuitry and adipocyte browning in humans. *N. Engl. J. Med.* **373**, 895–907 (2015).
4. Hnisz, D. et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**, 1454–1458 (2016).
5. Lupianez, D. G. et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012–1025 (2015).
6. Mifsud, B. et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* **47**, 598–606 (2015).
7. Rao, S. S. P. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
8. Trynka, G. & Raychaudhuri, S. Using chromatin marks to interpret and localize genetic associations to complex human traits and diseases. *Curr. Opin. Genet. Dev.* **23**, 635-641 (2013).
9. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
10. Moseti, D., Regassa, A. & Kim, W. K. Molecular regulation of adipogenesis and potential anti-adipogenic bioactive molecules. *Int. J. Mol. Sci.* **17**, 124 (2016).
11. Liu, X. et al. Functional architectures of local and distal regulation of gene expression in multiple human tissues. *Am. J. Hum. Genet.* **100**, 605–616 (2017).
12. Ardlie, K. G. et al. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
13. Wang, J., Duncan, D., Shi, Z. & Zhang, B. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Res.* **41**, W77–W83 (2013).
14. Shin, S.-Y. et al. An atlas of genetic influences on human blood metabolites. *Nat. Genet.* **46**, 543–550 (2014).
15. Willer, C. J. et al. Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1283 (2013).
16. Fuchsberger, C. et al. The genetic architecture of type 2 diabetes. *Nature* **536**, 41–47 (2016).
17. Shungin, D. et al. New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187–196 (2015).
18. Kheradpour, P. & Kellis, M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res.* **42**, 2976–2987 (2014).
19. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).
20. Roman, T. S. et al. A type 2 diabetes-associated functional regulatory variant in a pancreatic islet enhancer at the ADCY5 locus. *Diabetes* **66**, 2521–2530 (2017).
21. Russo, S. B., Ross, J. S. & Cowart, L. A. Sphingolipids in obesity, type 2 diabetes, and metabolic disease. *Handb. Exp. Pharmacol.* **216**, 373–401 (2013).
22. Kang, S. C., Kim, B. R., Lee, S. Y. & Park, T. S. Sphingolipid metabolism and obesity-induced inflammation. *Front. Endocrinol.* **4**, 67 (2013).
23. Suhre, K. et al. Human metabolic individuality in biomedical and pharmaceutical research. *Nature* **477**, 714–715 (2011).
24. Yang, X. et al. Validation of candidate causal genes for obesity that affect shared metabolic pathways and networks. *Nat. Genet.* **41**, 415–423 (2009).
25. Lefterova, M., Zhang, Y. & Steger, D. PPARgamm and C/EBP factors orchestrate adipocyte biology via adjacent binding on a genome-wide scale. *Genes Dev.* **22**, 2941–2952 (2008).
26. Jin, F. et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290–294 (2013).
27. Zhu, H. et al. Role of extracellular signal-regulated kinase 5 in adipocyte signaling. *J. Biol. Chem.* **289**, 6311-6322 (2014).
28. Kato, Y. et al. BMK1/ERK5 regulates serum-induced early gene expression through transcription factor MEF2C. *EMBO J.* **16**, 7054–7066 (1997).
29. Schmiedel, B. J. et al. 17q21 asthma-risk variants switch CTCF binding and regulate IL-2 production by T cells. *Nat. Commun.* **7**, 13426 (2016).
30. Chakraborti, C. K. New-found link between microbiota and obesity. *World J. Gastrointest. Pathophysiol.* **6**, 110–119 (2015).
31. Chen, Y. et al. Variations in DNA elucidate molecular networks that cause disease. *Nature* **452**, 429–435 (2008).
32. Henegar, C. et al. Adipose tissue transcriptomic signature highlights the pathological relevance of extracellular matrix in human obesity. *Genome Biol.* **9**, R14 (2008).
33. Kaneko, K. J. & DePamphilis, M. L. TEAD4 establishes the energy homeostasis essential for blastocoel formation. *Development* **140**, 3680–3690 (2013).
34. Phan, J. & Reue, K. Lipin, a lipodystrophy and obesity gene. *Cell Metab.* **1**, 73–83 (2005).
35. O'Reilly, M. W. et al. AKR1C3-mediated adipose androgen generation drives lipotoxicity in women with polycystic ovary syndrome. *J. Clin. Endocrinol. Metab.* **102**, 3327-3339 (2017).
36. Nagano, T. et al. Comparison of Hi-C results using in-solution versus in-nucleus ligation. *Genome Biol.* **16**, 175 (2015).
37. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Sci. (80-.).* **326**, 289–293 (2009).
38. Wingett, S. W. et al. HiCUP: pipeline for mapping and processing Hi-C data. *F1000Research* **4**, (2015).
39. Cairns, J. et al. CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. *Genome Biol.* **17**, 127 (2016).
40. Stančáková, A. et al. Hyperglycemia and a common variant of GCKR are associated with the levels of eight amino acids in 9,369 finnish men. *Diabetes* **61**, 1895–1902 (2012).
41. Laakso, M. et al. METabolic Syndrome In Men (METSIM) Study: a resource for studies of metabolic and cardiovascular diseases. *J. Lipid Res.* **58**, 481-493 (2017).
42. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
43. Rodriguez, A. et al. Molecular characterization of the lipid genome-wide association study signal on chromosome 18q11.2 implicates HNF4A-mediated regulation of the TMEM241 gene. *Arterioscler. Thromb. Vasc. Biol.* **36**, 1350–1355 (2016).
44. Dobin, A. et al. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
45. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
46. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).
47. Delaneau, O. et al. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nat. Commun.* **5**, 3934 (2014).
48. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
49. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).
50. Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479–1485 (2016).
51. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
52. Buil, A. et al. Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. *Nat. Genet.* **47**, 88–91 (2014).
53. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).

Acknowledgements

We thank the individuals who participated in the METSIM and GTEx studies. We also thank the sequencing core at UCLA for performing the RNA sequencing. In addition, we thank Cameron Osborne for his advice with the CHi-C protocol. We thank Xuanyao Liu for his assistance with the LD Score software. Francis Collins is thanked for providing the METSIM genotype 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 on 03/23/17. This study was funded by National Institutes of Health (NIH) grants HL-095056, HL-28481, U01 DK105561, R00 HL121172, and DK093757. D.Z.P. was supported by the NIH-NCI National Cancer Institute grant T32LM012424, M.A. was supported by the NIH grant T32HG002536, and A.K. by NIH grant F31HL127921. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the article. Genotyping for the METSIM cohort were supported by NIH grants DK072193, DK093757, DK062370, and Z01HG000024 and provided by the Center for Inherited Disease Research (CIDR). CIDR is fully funded through a federal contract from the NIH to The Johns Hopkins University, contract number HHSN268201200008I.

Author contributions

D.Z.P., K.M.G., P.P., and A.K designed the study. D.Z.P., K.M.G., J.B., P.P., A.K., R.M.C., J.S.S, and Z.M. performed methods development and statistical analysis. D.Z.P., K.M.G., M.A., Z.M., and J.B. performed computation analysis of the data. K.M.G. and Y.V.B performed the experiments. A.K., E.N., M.A., K.L.M., C.R., and P.P. performed RNA-sequencing and quality control. M.L. performed phenotyping. M.C., A.J.L., M.L., E.N., K.

L.M., M.B., and P.P performed data collection and METSIM genotyping. C.A.G. and K.S. S performed the replication analysis (TwinsUK). D.Z.P., K.M.G., A.K., and P.P. wrote the manuscript and all authors read, reviewed, and/or edited the manuscript.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-03554-9.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Chapter 3

Reverse gene-environment interaction approach to identify variants influencing body-mass index in humans

ARTICLES
https://doi.org/10.1038/s42255-019-0071-6
nature metabolism

Reverse gene–environment interaction approach to identify variants influencing body-mass index in humans

Kristina M. Garske[1], David Z. Pan[1,2], Zong Miao[1,2], Yash V. Bhagat[1], Caroline Comenho[1], Christopher R. Robles[3], Jihane N. Benhammou[1,4], Marcus Alvarez[1], Arthur Ko[5], Chun Jimmie Ye[6], Joseph R. Pisegna[1,4], Karen L. Mohlke[7], Janet S. Sinsheimer[1,8], Markku Laakso[9] and Päivi Pajukanta[1,2,9*]

Identifying gene–environment (G×E) interactions contributing to human cardiometabolic disorders is challenging. Here we apply a reverse G×E candidate search by deriving candidate variants from promoter–enhancer interactions that respond to dietary fatty acid challenge through altered chromatin accessibility in primary human adipocytes. We then test all variants residing in lipid-responsive open chromatin sites in adipocyte promoter–enhancer contacts for interaction effects between genotype and dietary saturated fat intake on body-mass index (BMI) in the UK Biobank. We discover 14 new G×E variants in 12 lipid-responsive promoters, including in well-known lipid-related genes (*LIPE*, *CARM1* and *PLIN2*) and newly associated genes, such as *LDB3*, for which we provide further functional and integrative genomic evidence. We further identify 24 G×E variants in enhancers, for a total of 38 new G×E variants for BMI in the UK Biobank, demonstrating that molecular genomics data produced in physiologically relevant contexts can be applied to discover new functional G×E mechanisms in humans.

Cardiometabolic disorders develop as a result of genetic predisposition, environmental factors and their interactions[1,2]. Genome-wide association studies (GWAS) have detected additive genetic effects for these traits, but the biological mechanisms explaining how genetic variation is involved in the increasing prevalence of obesogenic cardiometabolic disorders have yet to be identified. Some examples of G×E interactions are emerging, including, for instance, the highly replicated BMI risk variant rs9939609 in an intron of *FTO* that exhibits a significant interaction with physical activity for effect on BMI[3]. However, overall, there are few replicated G×E signals for cardiometabolic disorders in humans[1]. It has remained challenging to identify these signals, owing to small cohort sizes and poorly standardized definitions for human environmental phenotypes. Even with large cohorts such as the UK Biobank[4], the statistical power to detect G×E interactions by using a genome-wide agnostic search is limited owing to the small effect sizes of G×E interactions and heavy multiple-testing penalties. Furthermore, once G×E signals have been detected, the mechanisms underlying the associations remain unclear, warranting further fine-mapping studies.

To systematically identify genes involved in G×E interactions, we set out to quantify molecular genomic responses to saturated and monounsaturated fatty acid challenge in primary human adipocytes, as a cellular model of dietary fat intake in this key adipose tissue cell type. We measured differences in chromatin accessibility and searched the whole genome for chromosomal interactions between lipid-responsive gene promoters and enhancers to shed new light on the genomic molecular mechanisms relevant for lipid responses in human adipocytes. We hypothesized that these genomic responses would provide targeted regions harbouring candidate genetic variants for analysis of G×E interactions in the large UK Biobank cohort[4]. Using these targeted regions should restrict the multiple-testing burden hampering the typical agnostic genome-wide G×E analysis and expand knowledge of the true environmental exposures responsible for G×E signals, thereby revealing the underlying functional mechanisms. Thus, integrating context-specific molecular genomics with environmental phenotypes and clinical outcome data in the UK Biobank should help elucidate molecular mechanisms occurring in response to obesogenic cellular context that contribute to cardiometabolic traits in humans.

Results

Adipocyte accessible chromatin identifies regulatory regions. To obtain primary human adipocytes for study of the effects of lipids on chromatin dynamics, we first differentiated primary human white preadipocytes into adipocytes in vitro (Fig. 1a). We performed assay for transposase-accessible chromatin using sequencing (ATAC-seq)[5] on three biological replicates of the preadipocytes and adipocytes to identify open chromatin regions that were differentially accessible in the two cell types (Fig. 1b and Supplementary Tables 1 and 2).

[1]Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA. [2]Bioinformatics Interdepartmental Program, UCLA, Los Angeles, CA, USA. [3]Department of Computer Science, UCLA, Los Angeles, CA, USA. [4]Vache and Tamar Manoukian Division of Digestive Diseases, UCLA, Los Angeles, CA, USA. [5]Department of Medicine, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA. [6]Institute for Human Genetics, Department of Epidemiology and Biostatistics and Department of Bioengineering and Therapeutic Sciences, UCSF, San Francisco, CA, USA. [7]Department of Genetics, University of North Carolina, Chapel Hill, NC, USA. [8]Department of Biomathematics, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA. [9]Internal Medicine, Institute of Clinical Medicine, University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland.
*e-mail: ppajukanta@mednet.ucla.edu

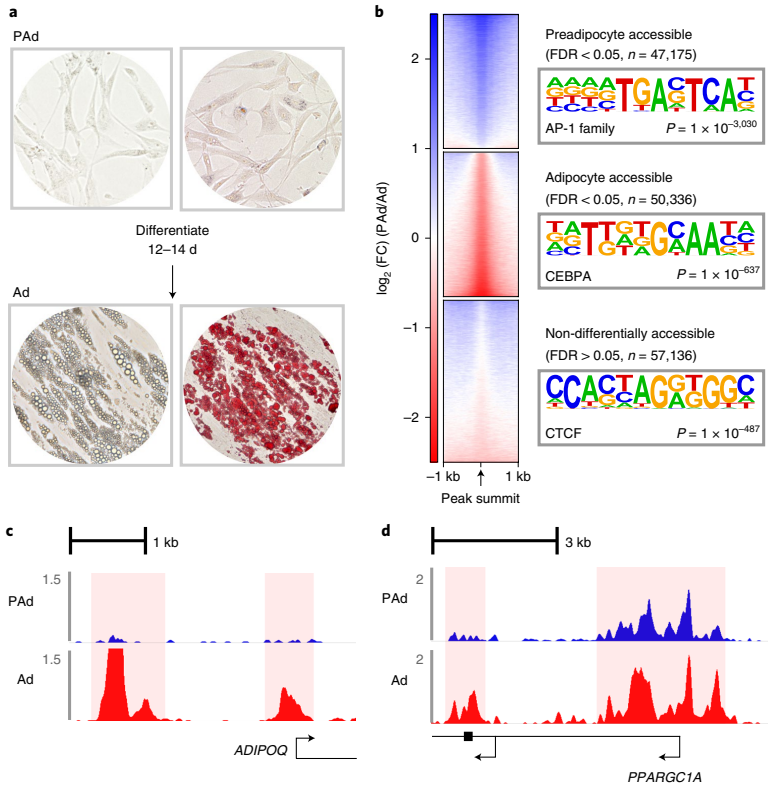


Fig. 1 | ATAC-seq analysis comparing primary human preadipocytes and adipocytes indicates successful adipocyte differentiation and widespread changes in chromatin accessibility. a, Bright-field images of preadipocytes (top; PAd) and in vitro-differentiated adipocytes (bottom; Ad) for unstained cells (left) and cells stained with Oil Red O (right). Images are representative examples from two independent experiments. **b**, Heat maps showing $\log_2$-transformed fold change ($\log_2$ (FC)) in bins per million mapped reads (BPM) for preadipocytes as compared to adipocytes in the three indicated peak sets. FDR was calculated (adjusting for $n = 154{,}647$ ATAC-seq peaks) from the *P* values of the quasi-likelihood (QL) *F* test (see Methods) for differential accessibility between preadipocytes and adipocytes using ATAC-seq libraries from $n = 3$ replicates per cell type. The most enriched TF motif for the indicated peak set is listed to the right. Enrichment *P* values were derived from the hypergeometric enrichment test of the proportion of the given top de novo-identified[6] TF motifs in the three indicated peak sets as compared with the background set (see Methods). **c**,**d**, Read coverage (BPM) in one representative ($n = 3$ replicates per cell type) preadipocyte (blue) and adipocyte (red) ATAC-seq library at the adipocyte accessible ATAC peaks in the promoters of the adipocyte hormone gene adiponectin (*ADIPOQ*) (**c**) and the adipocyte-specific marker peroxisome proliferator-activated receptor gamma coactivator 1-alpha (*PPARGC1A*) (**d**).

The 50,336 ATAC-seq peaks that were more accessible in primary human adipocytes ('adipocyte accessible') included the promoters of the *ADIPOQ* and *PPARGC1A* genes with known adipocyte-specific expression (Fig. 1c,d), providing evidence that we successfully differentiated adipocytes in vitro. To explore whether the adipocyte accessible peaks harboured transcription factor (TF) motifs relevant for adipocyte biology, we performed TF motif enrichment analyses with HOMER[6]. We found that the most enriched motif corresponded to the motif for the CCAAT-enhancer-binding protein alpha (C/EBPα) TF (Fig. 1b), an important regulator of the later stages of adipogenesis[7]. We then classified the adipocyte accessible peaks into functional genomic annotations[8] and observed that the adipocyte accessible peaks fell more often in adipocyte enhancers and less frequently in quiescent regions when compared with the full peak set or preadipocyte accessible peaks (Supplementary Fig. 1). Taken together, these results provide evidence that in vitro differentiation of adipocytes leads to an increase in chromatin accessibility in regions important for genomic regulation in adipocytes.

Genomic responses to dietary lipids in human adipocytes. We next searched for genomic regions harbouring regulatory elements that mediate adipocyte responses to the intake of different dietary lipids, by treating the adipocytes with the saturated fatty acid (SFA) palmitic acid (C16:0) or the monounsaturated fatty acid (MUFA) oleic acid (C18:1 *cis*-9) and then performing ATAC-seq on three biological replicates per condition (Fig. 2a). We found that treatment with either of these fatty acids resulted in increased staining with Oil Red O, which incorporates into neutral lipids, indicating that the lipid challenge resulted in increased storage of fatty acids in the lipid droplets of the cells (Supplementary Fig. 2). We identified 1,653 ATAC-seq peaks that were differentially accessible in the lipid-challenged primary human adipocytes in

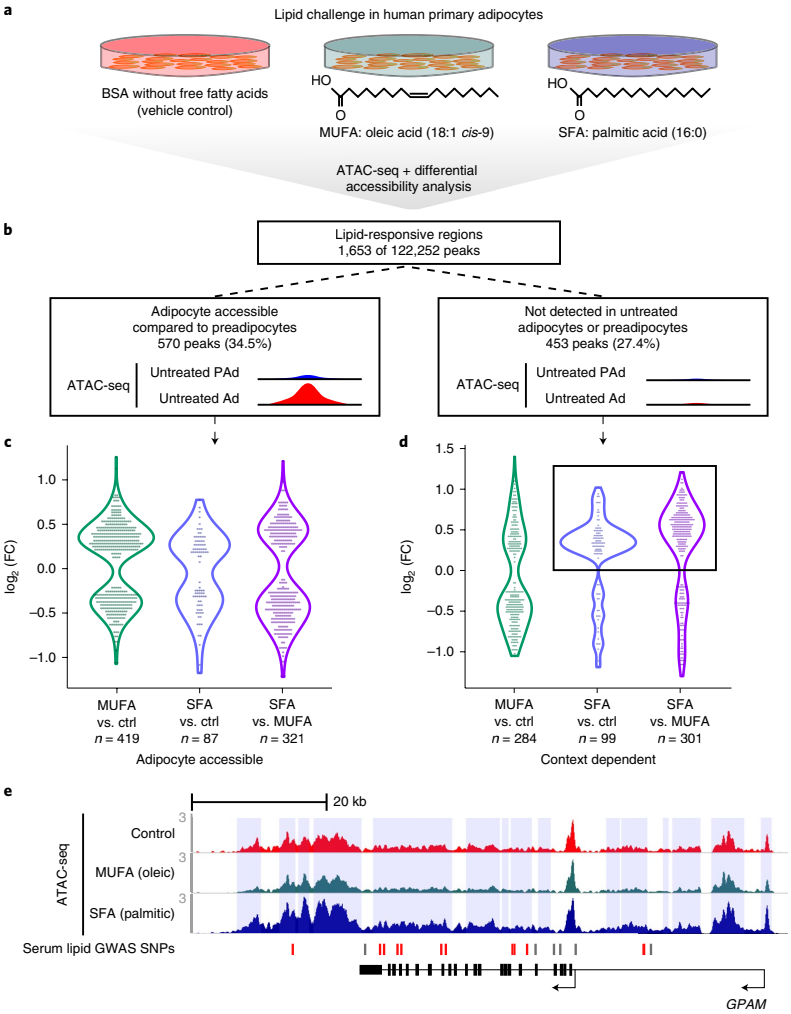


Fig. 2 | Lipid-responsive regions fall within adipocyte accessible regions of the genome, as well as within context-dependent regions that are not present in untreated adipocytes. a, Schematic overview of the lipid challenge experiment in human primary adipocytes. Treatments were performed in three replicates per condition. **b**, Schematic overview indicating the two categories of lipid-responsive peaks used for all downstream analyses. Peaks were considered differentially accessible at a cutoff of FDR < 0.05. FDR was calculated (adjusting for $n = 122{,}252$ ATAC-seq peaks) from the *P* values of the QL *F* test (see Methods) in one-way analysis of variance (ANOVA). Significant lipid-responsive peaks categorized as adipocyte accessible ($n = 570$) or context dependent (not identified in untreated preadipocytes or adipocytes; $n = 453$) were used in all downstream analyses. **c**,**d**, Violin plots showing the $\log_2$-transformed fold change in differentially accessible peaks in the indicated comparisons, stratified by whether the peak was adipocyte accessible (**c**) or context dependent (**d**). In **c**, the violin plot characteristics are as follows: MUFA versus control (ctrl) ($n = 419$): range, −1.07 to 1.25; median, 0.26; 25th percentile, −0.34; 75th percentile, 0.42; SFA versus control ($n = 87$): range, −1.17 to 0.77; median, 0.15; 25th percentile, −0.34; 75th percentile, 0.27; SFA versus MUFA ($n = 321$): range, −1.22 to 1.21; median, −0.26; 25th percentile, −0.48; 75th percentile, 0.43. In **d**, the violin plot characteristics are as follows: MUFA versus control ($n = 284$): range, −1.05 to 1.40; median, −0.33; 25th percentile, −0.53; 75th percentile, 0.37; SFA versus control ($n = 99$): range, −1.19 to 1.02; median, 0.34; 25th percentile, 0.20; 75th percentile, 0.47; SFA versus MUFA ($n = 301$): range, −1.30 to 1.21; median, 0.49; 25th percentile, −0.19; 75th percentile, 0.65. The box in **d** indicates a shift towards increased accessibility in SFA-treated cells, observed especially in the context-dependent peaks. **e**, Human genome browser snapshot (WashU) of the *GPAM* locus, which harbours 15 lipid-responsive peaks in a ~50-kb region (highlighted in blue vertical rectangles). Fourteen of these peaks are SFA responsive. Read coverage (BPM) is shown from one representative ATAC-seq library ($n = 3$ replicates per condition) for control (red), MUFA (green) and SFA (blue) treatment. GWAS SNPs for serum lipid traits are categorized as being in a lipid-responsive peak (red) or outside of a lipid-responsive peak (grey).

comparison to control adipocytes, referred to as lipid-responsive peaks (Fig. 2b, Supplementary Fig. 3 and Supplementary Tables 3 and 4). We cross-referenced these lipid-responsive peaks against the adipocyte accessible peaks (Fig. 1) and found that the 570 adipocyte accessible, lipid-responsive peaks fell mostly into enhancer and promoter annotations in adipocytes[8], in line with their likely importance in environmental responses and regulation of gene expression in adipocytes (Fig. 2b and Supplementary Fig. 4).

Notably, we found that 453 of the 1,653 lipid-responsive peaks were not detected as open chromatin in the initial ATAC-seq data created in untreated adipocytes and preadipocytes (Fig. 2b). When compared to the adipocyte accessible open chromatin, these context-dependent open chromatin regions fell into a higher percentage of quiescent annotations imputed from data created in unchallenged adipocytes[8] (Supplementary Fig. 4). This indicates that genomic regions that are not open and accessible for TF binding in untreated, steady-state adipocytes or preadipocytes become activated in adipocytes under lipid challenge conditions.

When we stratified the SFA and MUFA treatment responses by adipocyte accessible or context-dependent open chromatin regions, we found that the effects of SFA treatment in context-dependent open chromatin regions were shifted towards increased accessibility (Fig. 2c,d). This was in contrast to the MUFA responses, which were generally evenly distributed between peaks with decreased and increased accessibility (Fig. 2c,d). This suggests that different fatty acids can result in distinct signalling effects on genome-level responses to lipid intake in adipocytes, and, in particular, SFA intake seems to activate regions of the genome that are normally inactive in untreated adipocytes.

On closer examination, we found 14 context-dependent peaks, exhibiting increased accessibility in SFA-treated adipocytes, which fell into a ~50-kb region on chromosome 10 (Fig. 2e). The locus contained a total of 15 lipid-responsive peaks, which spanned the entirety of the gene encoding glycerol-3-phosphate acyltransferase, mitochondrial (*GPAM*) (Fig. 2e). The GPAM enzyme prefers saturated fatty acid substrates, and the *GPAM* locus has been associated with serum lipid traits[9–11] (triglycerides (TGs), high-density lipoprotein cholesterol (HDL), low-density lipoprotein cholesterol (LDL) and total cholesterol (TC)) and serum alanine aminotransferase (ALT), which is a biomarker for liver health[12], in previous GWAS[13]. The earlier GWAS associations at this locus, in combination with the strong genomic response to SFA treatment in adipocytes observed here, suggest that dysregulation of the important lipogenic pathway mediated by GPAM in adipocytes could contribute to obesogenic cardiometabolic disorders such as dyslipidaemias and non-alcoholic fatty liver disease (NAFLD).

Lipid-responsive gene promoters in chromosomal interactions. To identify genes under transcriptional regulation via chromosomal interactions, we performed promoter capture Hi-C (pCHi-C)[14] on the lipid-challenged human adipocytes with two biological replicates per condition (Supplementary Table 5). We identified 264 lipid-responsive ATAC-seq peaks that fell within adipocyte chromosomal interactions. To test whether these interacting, lipid-responsive regions of the genome harbour motifs for TFs that are important for lipid metabolism, we performed TF motif enrichment analysis[6] comparing the lipid-responsive peaks to non-lipid-responsive peaks within the chromosomal interactions. We found that motifs for peroxisome proliferator-activated receptor gamma (PPARγ), an important TF in adipogenesis and lipid uptake, and its cofactor retinoid X receptor (RXR) were among the ten most enriched motifs (Fig. 3a and Supplementary Table 6). This indicates that the lipid-responsive sites in adipocyte promoter–enhancer contacts represent genomic regions that are important for mediating cellular responses to lipid uptake.

To identify the target genes of the adipocyte responses to lipid challenge, we first focused on the interacting promoters from adipocyte pCHi-C (Fig. 3b), as promoters are more highly enriched for single-nucleotide polymorphisms (SNPs) that contribute to the heritability of local gene expression than enhancers[15,16]. The 86 interacting pCHi-C baits represented 154 gene promoters, given that the resolution of pCHi-C interaction data depends on the frequency of restriction sites in the genome (Fig. 3b and Supplementary Table 7). We performed a Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis[17] on the set of 154 interacting, lipid-responsive target genes, which identified two significantly enriched pathways for amino acid metabolism (false-discovery rate (FDR) < 0.05) (Supplementary Table 8).

As energy homeostasis is important for survival, we hypothesized that the 154 gene regions responsible for mediating the effects of lipid uptake in adipocytes might exhibit differences in the level of conservation when compared to other genes in the genome. We therefore obtained an average conservation score for the 114 protein-coding genes among the 154 genes (gene body ± 500 kb) by using PhastCons[18] and found that the lipid-responsive protein-coding genes had higher conservation scores across placental mammals than all other protein-coding genes in the human genome ($P = 0.020$) (Fig. 3c).

We further investigated whether these lipid-responsive genes exhibited constraints on genetic mutation, by using the probability of each gene being intolerant to loss-of-function mutation (pLI), defined in Lek et al.[19] as a high unlikelihood of protein-truncating mutations in humans. We found that, of the genes for which pLI scores were available ($n = 104$)[19], 27 genes (26.0%) were considered LoF intolerant. Given that 17.7% of all genes are considered LoF intolerant, the pLI for lipid-responsive genes is higher than expected by chance alone ($P = 0.022$) (Fig. 3d). Taking these findings together, we identified 154 genes with lipid-responsive promoters in chromosomal interactions that are less tolerant of LoF variants and reside within genomic regions that are more conserved than expected by chance alone.

Genes that interact with lipid-responsive enhancers. We next tested whether the genes that interacted with lipid-responsive enhancers exhibited similar characteristics to those of the genes that had lipid-responsive promoters. We first found that 169 lipid-responsive enhancers interacted with 223 promoter-containing HindIII baits in the adipocyte pCHi-C analysis (Supplementary Fig. 5). Given that multiple gene promoters can be captured within a single HindIII fragment, these 223 baits represented 323 gene promoters (Supplementary Fig. 5 and Supplementary Table 9). When we tested whether these 323 genes were enriched in any KEGG pathways, we did not find any functional pathways passing multiple-testing correction. This may be due to the fact that, on average, each lipid-responsive enhancer interacted with approximately two promoters (Supplementary Fig. 5), thus leading to ambiguities regarding which gene might be the true target of the lipid signalling response.

We further determined whether the genes that interacted with lipid-responsive enhancers exhibited mutational constraints by determining whether the *cis* regions (gene body ± 500 kb) of the protein-coding genes in this set ($n = 217$) had higher average conservation scores than all other protein-coding genes in the genome. In contrast to what we observed for genes with lipid-responsive promoters (Fig. 3c), we did not observe a significant difference in the conservation scores for genes that interacted with lipid-responsive enhancers (Supplementary Fig. 5). Interestingly, of the enhancer-interacting genes that had a pLI score ($n = 207$)[19], 50 (24.2%) were LoF intolerant, which is significantly higher than would be expected by chance alone ($P = 0.014$) (Supplementary Fig. 5). Taken together, these results are consistent with more moderate functional signifi-

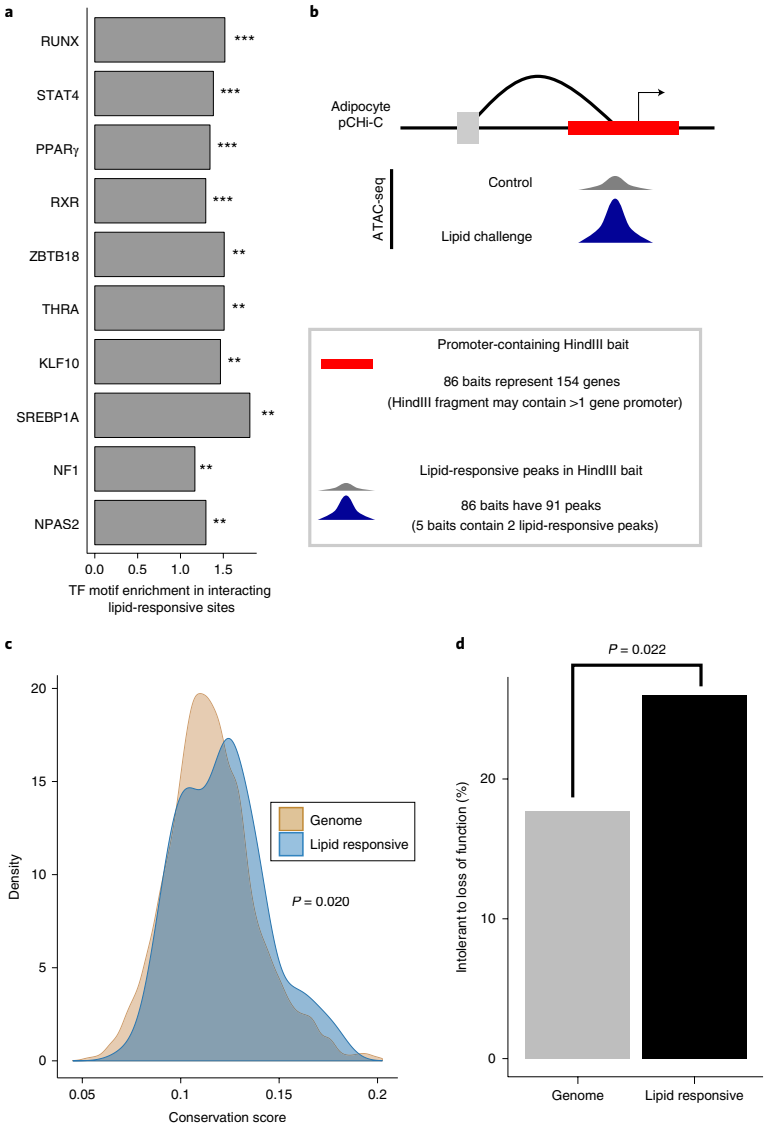


Fig. 3 | The 154 genes with lipid-responsive promoters within chromosomal interactions exhibit cross-species conservation and constraints on loss-of-function mutations, in line with their potential importance for energy homeostasis and survival. a, The top ten most enriched TF motifs in lipid-responsive open chromatin regions in adipocyte chromosomal interactions ($n = 264$) include motifs for key TFs in lipid metabolism, such as the cofactors PPARγ and RXR. **FDR < 0.001, ***FDR < 0.0001. Enrichment *P* values were derived from a hypergeometric enrichment test of the proportion of the given TF motif in the peak set as compared with the proportion in the background set of peaks (all non-lipid-responsive peaks in adipocyte chromosomal interactions), adjusted by the Benjamini–Hochberg method for the number of known motifs tested ($n = 364$)[6]. **b**, Schematic of the data we integrated to identify 154 genes with lipid-responsive promoters in adipocyte chromosomal interactions. **c**, Density plot showing the distribution of per-gene average conservation scores across placental mammals[18] for all protein-coding genes in the genome ($n = 19,316$) as compared with all protein-coding genes in the set of 154 genes with lipid-responsive promoters ($n = 114$). The *P* value was obtained from a two-sided Wilcoxon signed-rank test. **d**, Bar graph showing the proportion of protein-coding genes that are intolerant to loss-of-function mutation (unlikely to have protein-truncating variants in humans)[19] in the whole genome ($n = 3,204/18,122$; 17.7%) as compared to the protein-coding genes among the 154 genes ($n = 27/104$; 26.0%). The *P* value was obtained from a hypergeometric enrichment test.

Table 1 | Five lipid-responsive ATAC-seq peaks in interacting promoters overlap with GWAS SNPs for serum lipid traits

Peak chr.	Peak start	Peak end	Gene[a]	SNP(s) in peak	MAF[b]	Associated trait[c] (from ref. [11])	*P* value[c] (from ref. [11])	Index SNP (from ref. [11])	LD with index SNP[d] (r^2)
11	61,594,652	61,596,828	*FADS2–FADS1*	rs99780	0.37	TG/HDL/**LDL**/TC	$2.32\times10^{-16}/5.52\times10^{-9}$ **$/2.39\times10^{-21}$**$/8.62\times10^{-18}$	rs174546	0.93
				rs968567	0.16		3.4×10^{-9}/NS/8.91×10^{-11} $/2.27\times10^{-9}$		0.35
				rs191508698	0		$1\times10^{-16}/2.37\times10^{-8}$/ $1.24\times10^{-20}/3.1\times10^{-17}$		-
7	73,036,880	73,038,991	*MLXIPL*	rs55747707	0.2	TG	3.55×10^{-44}	rs17145738	0.47
				rs34060476	0.13		9.73×10^{-46}		0.77
2	27,432,323	27,432,971	*SLC5A6–ATRAID*	rs2580759	0.22	TG	2.18×10^{-17}	rs1260326	-
				rs1275530	0.72		1.88×10^{-17}		-
16	68,115,758	68,116,375	*NFATC3*	rs2107269	0.015	HDL	5.07×10^{-45}	rs16942887	0.66
19	10,981,139	10,983,631	*CARM1*	rs12460421	0.44	LDL	4.32×10^{-11}	rs6511720	-

Lipid-responsive ATAC-seq peaks that fell within promoters in adipocyte chromosomal interactions ($n=91$) were assessed for whether they contained GWAS SNPs for serum lipid traits from the meta-GWAS performed in Willer et al.[11]. Chr., chromosome; NS, not significant. [a]The gene listed corresponds to the promoter in the baited HindIII fragment with a lipid-responsive ATAC-seq peak. [b]MAF is the European frequency from the 1000 Genomes Project. [c]The most significant association is in bold when a SNP is associated with more than one serum lipid trait. [d]LD was calculated on the basis of Europeans in the 1000 Genomes Project; LD calculations > 0.2 are reported.

cance of the genes that interact with lipid-responsive enhancers when compared to genes with lipid-responsive ATAC-seq peaks in their promoters.

Lipid responses contribute to heritability of serum lipid traits. We hypothesized that the genes we were able to identify through lipid-responsive promoter–enhancer interactions might highlight important genomic regions that contribute to the heritability of cardiometabolic traits. We found that five lipid-responsive gene promoters and three lipid-responsive enhancers within adipocyte chromosomal interactions contained SNPs with genome-wide-significant ($P<5\times10^{-8}$) associations with serum lipid traits, identified in a meta-GWAS of ~180,000 individuals[11] (Table 1 and Supplementary Table 10). One of the lipid-responsive GWAS loci was the well-known nutritional response locus containing the fatty acid desaturase (*FADS1–FADS2–FADS3*) gene cluster on chromosome 11 (Fig. 4a), which harbours SNPs that have been associated with multiple cardiometabolic traits[20] and intermediate phenotypes[21,22]. In line with the observed pleiotropy among the serum-lipid-associated SNPs, the lipid-responsive peak in the *FADS2* promoter contained GWAS SNPs for all tested serum lipid traits[11] (LDL, HDL, TC and TG), with the strongest signal coming from rs99780 for LDL ($P=2.39\times10^{-21}$; Table 1). Notably, the observed open chromatin peak in *FADS2* was more accessible with palmitic acid treatment than with oleic acid treatment (Fig. 4b; FDR = 0.0021), corresponding to the fact that one of the substrates of FADS2 is palmitic acid[23].

This response at a GWAS locus for serum lipids was reminiscent of the strong SFA response at the *GPAM* locus (Fig. 2e); in fact, all five GWAS SNPs for serum lipids in lipid-responsive gene promoters within chromosomal interactions exhibited increased chromatin accessibility specifically in response to SFA intake in adipocytes (Supplementary Table 11). The lipid-responsive enhancers that interacted with gene promoters in adipocyte pCHi-C exhibited a similar trend (Supplementary Table 12). These results suggest that environmental responses, particularly to saturated fat intake, explain functional mechanisms at these lipid GWAS loci.

Because signals that do not reach genome-wide significance probably also contribute to the heritability of cardiometabolic disorders, particularly with the added effect of relevant environmental stimuli, we wanted to test the combined effect of all variants in our lipid-responsive regions while still accounting for the linkage disequilibrium (LD) between them. We therefore tested whether genetic variants in the *cis* region (gene body ± 500 kb) of all 154 lipid-responsive, interacting promoters contributed significantly to the heritability of serum lipid levels. For these analyses, we used the LD score-partitioned heritability method[24] and GWAS summary statistics from the high-powered meta-GWAS for serum lipid traits[11]. We found that 2.9% of all variants resided within the *cis* regions of the 154 genes, and these variants contributed significantly to the heritability of the four lipid traits ($0.0088 \leq P \leq 0.045$, with an average enrichment of 2.915; Supplementary Table 13). In contrast, 5.5% of all variants resided within the *cis* regions of genes with promoters interacting with lipid-responsive enhancers. These SNPs contributed significantly to the heritability of HDL, but not to that of the other lipid traits (Supplementary Table 14). This is consistent with the more diffuse overall functional characterization of the genes that interacted with lipid-responsive enhancers when compared to the genes with lipid-responsive open chromatin in their promoters. Overall, these results indicate that adipocyte lipid-responsive, interacting loci are important in modulating serum lipid levels in humans and provide evidence that variants in these regions might have a role in G×E interactions in humans.

Lipid responses identify new G×E interactions for BMI in UK Biobank. The large, deeply phenotyped UK Biobank[4] cohort can provide a valuable resource for G×E studies, particularly because the participants' environmental phenotypes have been characterized in a systematic manner. Saturated fat intake has known adverse effects in the context of cardiometabolic disorders[25–27], and we present evidence here in human adipocytes of an enhanced effect of human adipocyte genomic responses to SFA intake on cardiometabolic traits in comparison to genomic responses from MUFA intake (Fig. 2c–e and Supplementary Tables 11 and 12). To maximize the number of individuals for whom phenotypes were available, as well as to aim for the most relevant environment and cardiometabolic outcome, we used dietary intake of saturated fat (24-h recall) as the environmental variable and BMI as the outcome for our G×E analysis.

We first tested whether there were any genome-wide-significant signals for G×E interactions by using 167,908 individuals in the UK Biobank. We corrected the BMI measurements for array type,

sex (inferred), age (when the participant attended an assessment centre), age^2, the assessment centre ID and genetic principal components 1–20, as done previously[28]. We then inverse normal transformed the residuals to account for mean–variance relationships in the phenotype, which have been shown previously to impact G×E signals[29]. In the quantile–quantile plot from the genome-wide scan for G×E interactions, there was no evidence of genomic inflation in the G×E linear model (Supplementary Fig. 6). Furthermore, the fact that we were not able to detect any genome-wide-significant signals in this genome-wide G×E analysis (see equation (1) in the Methods; Supplementary Fig. 6) supports the feasibility of our reverse G×E candidate search approach, which identifies functional candidates for G×E analyses from molecular genomics data produced under biologically relevant conditions.

We have provided evidence that lipid challenge in adipocytes highlights important regions of the genome that respond to environmental cues and contribute to the heritability of cardiometabolic traits. Thus, these regions represent strong candidates for G×E interactions in humans. The 154 promoters in chromosomal interactions contained 91 lipid-responsive open chromatin sites (Fig. 3b), and we determined that 75 of these 91 candidate regions contained variants with minor allele frequency (MAF) > 0.05 in the set of 167,908 individuals for whom we had both dietary saturated fat intake and BMI phenotypes available (Supplementary Table 15).

We performed G×E analysis by incorporating all SNPs residing in the open chromatin, lipid-responsive promoter regions (n = 290; Supplementary Table 15) into a multivariable linear model (see equation (2) in the Methods). This resulted in the identification of 14 significant nonredundant G×E SNPs (LD $r^2 < 0.2$) in 12 interacting promoters, including new G×E SNPs in the promoters of the well-known lipid-associated genes encoding hormone-sensitive lipase (*LIPE*), coactivator-associated arginine methyltransferase 1 (*CARM1*) and perilipin 2 (*PLIN2*) (Table 2 and Supplementary Table 16).

We next performed a similar G×E analysis on all SNPs at the lipid-responsive enhancers that interacted with gene promoters in human adipocytes. Of the 173 lipid-responsive regions within the interacting enhancers (Supplementary Fig. 5), 142 contained SNPs with MAF > 0.05 in the 167,908 individuals in the UK Biobank (Supplementary Table 17). We used the same multivariable linear model approach to test these SNPs (n = 410) for an interaction with the effect of saturated fat intake on BMI and found 24 nonredundant (LD $r^2 < 0.2$) significant G×E SNPs (Supplementary Table 18). Given that enhancer fragments can interact with more than one promoter-containing pCHi-C bait, these 24 nonredundant SNPs interacted with a total of 27 promoter baits in human adipocytes (Supplementary Table 18).

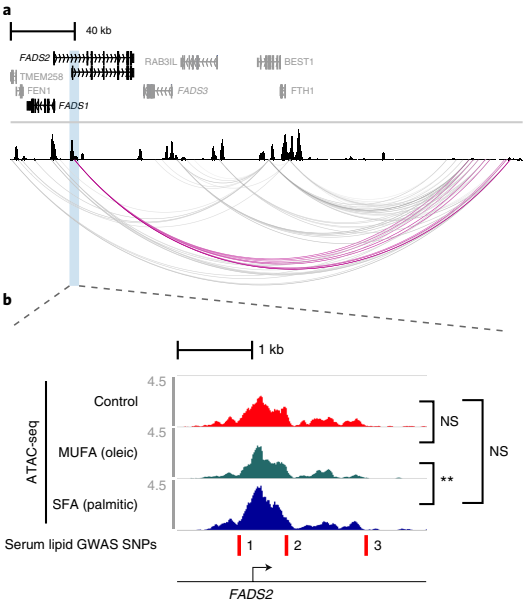


Fig. 4 | A lipid-responsive open chromatin region in human primary adipocytes at the 11q12.2 *FADS1–FADS2–FADS3* locus harbours GWAS SNPs for serum lipid traits. a, Genome browser snapshot showing the *FADS1–FADS2–FADS3* locus with data from adipocyte baseline ATAC-seq (one representative example from n = 3 vehicle control (BSA) ATAC-seq libraries) and pCHi-C (interactions identified in at least one condition from the adipocyte lipid challenge pCHi-C analysis were included; see Methods). Chromosomal interactions of the *FADS2* promoter are highlighted in magenta. **b**, Read coverage (BPM) in one representative ATAC-seq library (n = 3 replicates per condition) from vehicle control (red), MUFA (green) and SFA (blue) treatment. The lipid-responsive peak in one of the *FADS2* promoters is more accessible in SFA-treated than in MUFA-treated adipocytes and contains three independent GWAS SNPs for serum lipid traits (1, rs191508698; 2, rs968567; 3, rs99780). FDR was calculated (adjusting for n = 122,252 ATAC-seq peaks) from the P values of the QL F test (see Methods) in one-way ANOVA. For the post hoc test to determine which comparison was significant after one-way ANOVA (MUFA vs. control, SFA vs. control or SFA vs. MUFA), we determined the least significant difference; **FDR = 0.0021; NS, not significant.

Identifying altered chromatin states at G×E SNP sites. The differential chromatin accessibility in response to lipid challenge in adipocytes probably stems from altered chromatin states, such as TF binding or histone modifications. This idea is supported by our finding that the lipid-responsive regions within chromosomal interactions are enriched for the motifs of TFs important in lipid metabolism (Fig. 3a and Supplementary Table 6). To determine the predicted allelic effect of the G×E SNPs on chromatin features, we used the DeepSEA tool[30], which applies a deep learning algorithm to publicly available molecular genomics data to predict chromatin features on the basis of genomic sequence in silico. Notably, 11 of the 20 (55%) G×E SNPs in lipid-responsive promoters had a functional significance score of less than 0.05, and the predicted impacts of the G×E SNPs included differential binding of RXRA (Supplementary Table 19). Conversely, only 5 of the 26 (19%) G×E SNPs in lipid-responsive enhancers had functional significance scores of less than 0.05 (Supplementary Table 20). It is worth noting that the publicly available data used to train the DeepSEA neural network do not include molecular genomics data for adipocytes or adipose tissue. Therefore, it is possible that the G×E SNPs fall into cell-type-specific regulatory elements and disrupt chromatin features that cannot be predicted with this tool.

To obtain further evidence for the function of the G×E SNPs in adipose tissue, we examined whether the G×E SNPs affected local gene expression in human adipose tissue as *cis* expression quantitative trait loci (*cis*-eQTLs) and whether the *cis*-eQTL target gene was the same gene that harboured the lipid-responsive promoter and G×E signal. We found that 3 of the 12 genes with lipid-responsive promoters harbouring G×E signals were also regulated in *cis* by their G×E SNP at the genome-wide significance level in subcutaneous adipose RNA-seq data (n = 335) from the Finnish METSIM cohort[15,31] (Table 2 and Supplementary Table 16). These genes were *GLTSCR2* (encoding glioma tumour-suppressor candidate region gene 2 protein), *PLIN2* and *LDB3* (LIM domain binding 3). Additionally, 2 of the 27 genes interacting with lipid-responsive

Table 2 | Significant G×E interactions affecting BMI from a multivariable linear model for 290 promoter SNPs in lipid-responsive ATAC-seq peaks

SNP	P_G	P_{GE}	β_G	β_{GE}	Genes in bait	*Cis*-eQTL FDR[a] (from ref. [15])	Target gene (from ref. [15])	$\log_2$ (FC) (from ref. [15])
rs1974817[b]	0.0089	0.0010	2.3	−0.089	*GLTSCR2–SNORD23*	2.4×10^{-31}	*SEPW1*	0.73
rs58631862	0.032	0.0031	0.085	−0.0035	*RGMB*	-	-	-
rs74249860	0.0013	0.0043	0.081	−0.0021	*SH3GL3*	0.021	*GOLGA6L4*	0.90
rs112438892[b]	0.0017	0.0050	−0.50	0.012	*CARM1*	2.7×10^{-5} 0.015	*SMARCA4* *ICAM4*	−0.28 0.38
rs17625418	0.015	0.0054	−2.2	0.073	*GLTSCR2–SNORD23*	1.3×10^{-13} 0.0038	*SEPW1* *GLTSCR2*	−0.76 −0.3
rs3848589	0.045	0.014	0.073	−0.0027	*HOOK2–JUNB*	0.0058	*CACNA1A*	−0.6
rs882881	0.034	0.016	−0.051	0.0017	*PLIN2*	-	-	-
rs35213231	0.56	0.020	0.66	−0.038	*RNU2-10P*	0.019	*PTPRG*	0.29
rs41322049	0.29	0.021	0.032	−0.0021	*BLVRB–SPTBN4*	-	-	-
rs35678764[b]	0.028	0.025	−5.7	0.19	*RDH8–COL5A3*	-	-	-
rs10788522[b]	0.013	0.027	3.9	−0.14	*LDB3*	6.1×10^{-4}	*LDB3*	0.42
rs10422283	0.045	0.029	−0.072	0.0023	*LIPE–LIPE-AS1*	-	-	-
rs867773[b]	0.31	0.033	−2.0	0.11	*PLIN2*	0.048	*PLIN2*	−0.39
1:12245360_CCTTTTT_C	0.047	0.034	0.64	−0.023	*TNFRSF1B–MIR4632*	-	-	-

The reported *P* values are from the β values in the multivariable linear model (see equation (2) in the Methods), where *g* is the number of minor alleles of the genotype and *e* is saturated fat intake. Here P_G indicates the *P* value for the genotype effect and P_{GE} indicates the *P* value for the G×E effect; β values follow the same notation. For the multivariable linear model, there were a total of 290 SNPs and 38,394 individuals with no missing data available for study. [a]*Cis*-eQTLs were identified in adipose tissue from the METSIM cohort[15,31]. [b]When more than one non-independent SNP (LD $r^2 > 0.2$) has a significant G×E *P* value for the lipid-responsive region, only the top SNP is reported; the SNPs in LD with the top SNP are listed in Supplementary Table 16.

enhancers were regulated in *cis* by their interacting G×E SNP at the genome-wide significance level in human subcutaneous adipose tissue (Supplementary Table 18): *THBS2* (thrombospondin 2) and *CITED4* (Cbp/p300-interacting transactivator with Glu/Asp-rich C-terminal domain 4).

We further examined whether the imputed *cis* expression values for these five *cis*-eQTL target genes (eGenes) were correlated with BMI or other obesogenic cardiometabolic phenotypes, as determined through transcriptome-wide association analysis (TWAS)[32,33]. The *LDB3* adipose expression model was strongly associated with BMI, arm fat percentage (genome-wide-significant TWAS score > 5.0) and other related body fat distribution phenotypes (TWAS score > 4.0)[32]. Furthermore, the tibial artery expression model for *LDB3* was also significantly associated with high blood pressure and cardiovascular disease (genome-wide-significant TWAS score < −5.0)[32]. None of the other eGene expression models was associated with cardiometabolic phenotypes at the genome-wide significance level.

Because the adipose expression models of *LDB3* were significantly associated with BMI in TWAS[32], we followed up on the most significant G×E SNP in the *LDB3* lipid-responsive peak (Fig. 5a), rs10788522, which was also an adipose *cis*-eQTL for *LDB3* in the METSIM cohort (Table 2). As evidenced by the ATAC-seq reads intersecting with SNP rs10788522 (Fig. 5b), we found, by electrophoretic mobility shift assay (EMSA), that adipocyte nuclear protein bound this G×E SNP (Fig. 5c and Supplementary Table 21). Whereas *LDB3* was expressed in subcutaneous adipose tissue from the METSIM cohort, as well as the Gene–Tissue Expression (GTEx) Project[34], we could not reliably detect *LDB3* expression by quantitative PCR (qPCR) in the lipid-challenged adipocytes (data not shown). We therefore determined whether any publicly available datasets analysed the transcriptome of sorted cells from human adipose tissue. In previously published microarray datasets (GSE80654 and GSE100795) from human adipose biopsies in which adipocytes were collected and the remaining cell types were separated by fluorescence-activated cell sorting (FACS), *LDB3* was expressed in human adipocytes at a level comparable to that in the other adipose cell types[35,36]. Taken together with our finding that the *LDB3* promoter responds to lipid uptake in adipocytes, these data support the conclusion that the G×E SNP rs10788522 regulates *LDB3* expression in adipocytes in response to dietary saturated fat and that the interaction has a protective (BMI-lowering) effect (Table 2). Altogether, we provide a mechanistic interpretation and fine-mapping of a causal G×E SNP, rs10788522, narrowing it to the promoter of *LDB3* that exhibits differential open chromatin in response to lipid challenge in human adipocytes (Figs. 5 and 6).

Discussion

It is well established that environment has a major role in the development of cardiometabolic disorders. However, G×E interactions have been challenging to detect owing to both the lack of extensive study cohorts with sufficient statistical power to detect the small G×E effects and the complexity of environmental exposures that are difficult to measure in a standardized way in humans[1]. Systematic identification of the effects of defined environmental contributions to cardiometabolic disorders is thus necessary to effectively move towards the promise of precision medicine. Through our integration of context-specific molecular genomics data with human epidemiological and clinical outcome data in the UK Biobank, we provide much-needed information on how the chromatin landscape of human adipocytes responds to external environmental signals and identify the molecular basis of new G×E interactions in humans (Fig. 6).

We reversed the typical approach of selecting candidate G×E interactions among GWAS SNPs, by first scanning the genome for molecular responses to controlled environmental stimuli, apply-

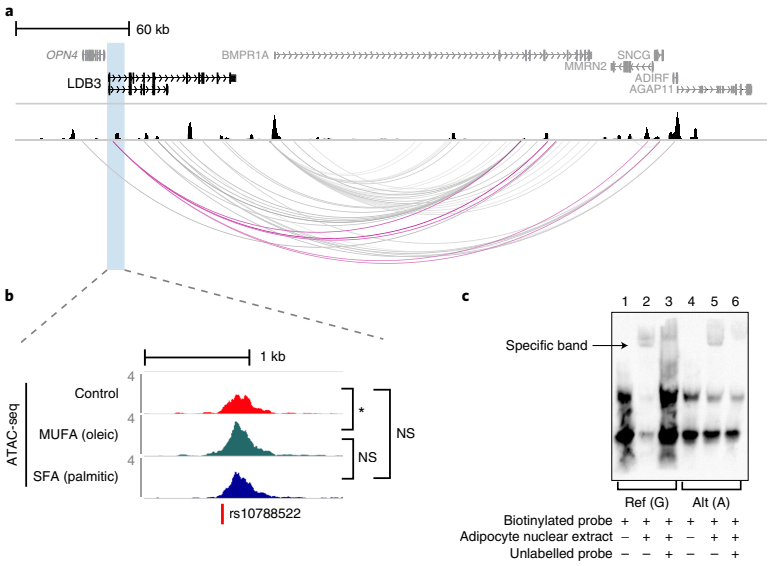


Fig. 5 | Fine-mapping of the gene–diet interaction for BMI in the *LDB3* promoter region. a, Genome browser snapshot of the *LDB3* locus with adipocyte baseline ATAC-seq (one representative example from $n=3$ vehicle control (BSA) ATAC-seq libraries) and pCHi-C data (interactions identified in at least one condition from the adipocyte lipid challenge pCHi-C analysis are included; see Methods). *LDB3* promoter interactions are highlighted in magenta. **b**, Read coverage (BPM) in one representative ATAC-seq library ($n=3$ replicates per condition) for control (red), MUFA (green) and SFA (blue) treatment. The lipid-responsive ATAC-seq peak harbouring the significant G×E SNP rs10788522 resides in the *LDB3* gene. FDR was calculated (adjusting for $n=122{,}252$ ATAC-seq peaks) from the *P* values of the QL *F* test (see Methods) in one-way ANOVA. For the post hoc test to determine which comparison was significant after one-way ANOVA (MUFA vs. control, SFA vs. control or SFA vs. MUFA), we determined the least significant difference; *FDR = 0.024. The *P* value for β_{GEi} at rs10788522 was derived from the multivariable linear model testing for G×E interactions in the UK Biobank; $P=0.027$. **c**, EMSA showing binding of adipocyte nuclear protein to the G×E SNP rs10788522. The specific band is competed away for both alleles (lanes 3 and 6). Ref, reference allele; alt, alternate allele.

ing a cellular model of the effects of saturated or monounsaturated fat intake on chromatin dynamics in primary human adipocytes. Through our integration of chromatin accessibility and chromosomal interactions in lipid-challenged adipocytes, we identified lipid-responsive open chromatin within promoter–enhancer contacts, effectively identifying candidate G×E interaction genes with strong evidence of genomic regulation in response to fatty acid uptake and processing in human adipocytes. This systematic approach culminated in testing a total of 700 SNPs in the accessible, lipid-responsive chromatin regions for interactions with dietary saturated fat intake affecting BMI in the UK Biobank[4]. This led to the identification of 14 significant, nonredundant G×E SNPs in 12 gene promoter regions and 24 nonredundant G×E SNPs in 20 enhancers, representing new gene–diet interactions affecting BMI (Fig. 6).

We observed that the *LDB3* gene is regulated in *cis* by its promoter G×E SNP in human adipose tissue, and previous TWAS analyses have shown that the imputed local adipose expression of *LDB3* is significantly associated with BMI and related cardiometabolic phenotypes[32,33]. Notably, individuals with nonsynonymous mutations in exon 6 of *LDB3* have been shown to exhibit autosomal dominant myofibrillar myopathy characterized by fatty degeneration (steatosis) of the muscle that progresses with age[37,38]. This is suggestive of the role of lipid metabolism in the pathophysiology of these variants. Here we show that accessibility of the promoter region is increased in response to lipid challenge in human adipocytes, and adipocyte nuclear protein binds to the G×E SNP site. Although we did not detect *LDB3* expression in our cultured adipocytes, we found that the gene was expressed in mature adipocytes isolated from human adipose biopsies, in vivo, suggesting that future studies to understand the role of LDB3 in adipocytes may require in vivo mouse models. In line with this, LDB3 is known to bind α-actinin isoforms that are not muscle specific, and actin cytoskeleton organization is critically important in maintaining proper tissue functions. Taken together with the adipose *cis*-eQTL and TWAS results, our lipid challenge findings provide a functional mechanism for the G×E effect in human adipocytes, adding to the knowledge of environmental response to diet and the consequent effects on genetic predisposition to cardiometabolic traits in humans.

The 154 lipid-responsive promoters within adipocyte chromosomal interactions provide a set of biologically important genes for studies in adipose tissue. These gene regions have a higher conservation score and the 154 genes are more likely to be intolerant to LoF mutation than expected by chance, suggesting that there are evolutionary constraints to maintain their proper function. The LoF-intolerant genes are widely and highly expressed[19], and they may exhibit pleiotropy. Nonetheless, the lipid-responsive mechanism of genomic regulation identified for the LoF-intolerant genes in this study provides evidence that these genes may be important in maintaining energy homeostasis, which is critical for survival.

It is known that dietary saturated fat intake is correlated with various adverse cardiometabolic outcomes[25,27], and a genetic risk score (GRS) for obesity-related traits was previously shown to interact with saturated fat intake to affect BMI[26]. However, as the authors of this study note, the underlying mechanisms for the G×E

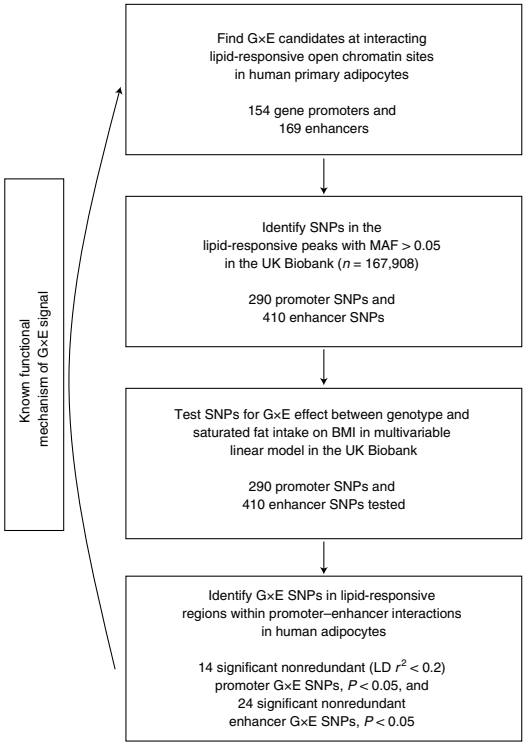


Fig. 6 | Analytical approach. Flowchart of our approach to integrate molecular genomics data created in human adipocytes in physiologically relevant contexts that, when combined with human cohort molecular and phenotype data, enable the detection of G×E signals.

interactions remain elusive, particularly when the effect is estimated across tens of SNPs[26]. Here we bridge this knowledge gap and show through a genome-wide scan of regulatory open chromatin responses to saturated versus monounsaturated fat uptake that a subset of these responses are probably specific to saturated fat intake and, when dysregulated at the genetic level, could underlie GWAS and G×E signals for cardiometabolic traits. We note that we assessed the effects of exogenous fatty acids and did not quantify the consequent cellular fatty acid processing, which could include desaturation of palmitic acid (C16:0) to the MUFA palmitoleic acid (C16:1), by stearoyl-CoA desaturase (SCD). Thus, we do not know whether the effects we observed at the DNA level were directly due to SFA signalling or resulted from downstream signalling mechanisms of the SCD pathway, affecting MUFA concentrations inside the cell.

Our finding that SNPs within the *cis* regions of the 154 lipid-responsive gene promoters (gene body ± 500 kb) contributed significantly to the heritability of serum lipid levels suggests that responses to lipid uptake in adipocytes are associated with cellular programs that can modulate serum lipid levels. Correspondingly, we identified five adipocyte lipid-responsive, interacting gene promoters and three lipid-responsive enhancers that harboured genome-wide-significant signals for serum lipid traits, including in the well-known *FADS1–FADS2–FADS3* gene cluster involved in nutrient sensing. Specific lipid-associated SNPs within this locus have undergone positive selection in Inuits, which is thought to have been in response to the polyunsaturated fatty acids (PUFAs) in diets with high amounts of marine mammalian fat[39]. Furthermore, a gene–diet interaction for PUFA intake has been identified in Europeans at the *FADS1–FADS2–FADS3* locus[40], and, while many studies of the effects of *FADS1–FADS2* polymorphisms on PUFA metabolism have shown a clear role for this locus in modulating serum fatty acid levels, an understanding of the mechanistic effects of these SNPs, associated with cardiometabolic disorders, has been less conclusive[41,42]. Our results suggest that the underlying mechanisms may derive from the effects of saturated fats at this locus. This additional role for a well-established, yet-inconclusive locus supports the applicability of our approach to identify G×E interactions through characterization of molecular genomic responses to relevant environmental stimuli.

Although there has been a strong international effort by the scientific community to characterize genomic regulatory elements in various cell types and tissues, many of the publicly available datasets and corresponding genomic annotations have been created in cells at steady state or under unstimulated baseline conditions[43–46]. Context-specific molecular genomics studies have mainly been performed in immune cell types[47–49], while similar studies in other cell types are scarce. In line with the importance of studying molecular genomic phenotypes under physiologically relevant conditions, we found here that, whereas most lipid-responsive regions in human adipocytes reside in known adipocyte enhancers, a subset emerge from regions that were not identified as open chromatin and were annotated as quiescent[8] in unchallenged adipocytes, indicating that quiescent regions of the human genome are activated under specific environmental contexts. Thus, this genome-wide scan for response to fatty acid uptake in adipocytes adds to the currently incomplete understanding of genomic regulation in contexts that are expected to confer complex cardiometabolic disease states.

In conclusion, this study highlights the value of performing genome-wide functional genomics experiments in a context-specific manner to advance understanding of environmental epigenomic responses underlying complex traits. We performed a global assessment of the genomic responses of primary human adipocytes to dietary fatty acid uptake, through incorporating open chromatin and chromosomal interaction data that we followed for G×E interactions in UK Biobank. Overall, our study helped discover candidate functional mechanisms at 38 new gene–diet interactions on BMI, identified over 100 genes important for lipid uptake that may contribute to variance in cardiometabolic traits and uncovered a new set of interacting open chromatin elements responding to lipid challenge in a primary human cell type relevant for lipid synthesis and storage.

Methods

Cell lines and culture reagents. We obtained and cultured primary human white preadipocyte cells as recommended by PromoCell (PromoCell, C-12731, lot 395Z024) for preadipocyte growth and differentiation into adipocytes. Cell medium (PromoCell) was supplemented with 1% penicillin-streptomycin. We maintained the cells at 37 °C in a humidified atmosphere at 5% CO_2. For the lipid challenge experiments in adipocytes, we serum starved cells for 16 h with 0.5% FCS in supplemented adipocyte basal medium (PromoCell), before treatment with 200 µM palmitic acid:BSA complex, 200 µM oleic acid:BSA complex (Sigma-Aldrich, O3008) or 0.23% fatty-acid-free BSA (Sigma-Aldrich, A8806) as a vehicle control, in medium containing 0.5% FCS for 24 h before performing experiments.

Palmitic acid conjugation to BSA. We dissolved 25.6 mg of palmitic acid (Sigma-Aldrich, P5585) into 1 ml of 0.15 M NaCl at 70 °C in a shaking heat block to make 100 mM palmitic acid solution. We added the palmitic acid solution dropwise into 10% (wt/vol) BSA in 0.15 M NaCl at 37 °C to generate palmitic acid:BSA conjugate at 8 mM stock for palmitic solution.

Oil Red O staining and quantification. We prepared Oil Red O stock by making 0.3% Oil Red O solution in >99% isopropanol and filtering through a 0.45-µm filter. This solution was diluted 3:2 in water, incubated at room temperature for 10 min and then filtered through a 0.22-µm filter. We fixed cells for 30–60 min in 10%

formalin, rinsed with distilled water and incubated for 2–5 min at room temperature with 60% isopropanol. We stained with Oil Red O for 15 min, rinsed well with water and collected images for quantification. Cells were photographed with a Keyence bright-field light microscope under ×10 magnification for publication images and ×20 magnification for lipid droplet quantification. Lipid droplet numbers were determined for ~20 cells per condition (untreated, BSA, palmitic acid and oleic acid). The total area of Oil Red O staining was quantified with ImageJ[50].

ATAC-seq. We performed the ATAC-seq protocol in untreated primary human preadipocytes and adipocytes for 300,000 nuclei in three biological replicates per cell type, similarly to the protocol developed in Buenrostro et al.[5]. Specifically, we lysed cells in ice-cold lysis buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$) plus 0.03–0.1% Tween-20 for 10 min on ice. We centrifuged at 500*g* for 10 min at 4 °C and then resuspended the nuclear pellet in 50 µl of transposition master mix (25 µl of 2× TDE1 buffer, 2.5 µl of transposase, 22.5 µl of nuclease-free water; Illumina, FC-121-1030). We incubated samples at 37 °C for 30 min and then purified the DNA with the Qiagen MinElute kit (Qiagen, 28204). Libraries were amplified for six cycles and sequenced on an Illumina HiSeq 4000 to produce an average of 23,376,290 (±3,337,206) reads.

For the ATAC-seq analysis in primary human adipocytes that underwent lipid challenge, we performed omni-ATAC as developed in Corces et al.[51] for 300,000 nuclei in three biological replicates per condition. Specifically, we treated the cells with DNase I (Worthington; 200 U ml^{-1}) at 37 °C for 30 min, rinsed the cells with ice-cold PBS, scraped the cells gently to the side of the plate, resuspended them in 50 µl of ice-cold lysis buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% Igepal CA-630, 0.1% Tween-20, 0.01% digitonin) and incubated them on ice for 3 min. We washed with 1 ml of ice-cold lysis quench (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% Tween-20) and centrifuged at 500*g* for 10 min at 4 °C. The nuclear pellet was resuspended in 50 µl of transposition master mix (25 µl of 2× TDE1 buffer and 2.5 µl of transposase, 16.5 µl of PBS, 0.5 µl of 1% digitonin, 0.5 µl of 10% Tween-20, 5 µl of nuclease-free water). We incubated samples at 37 °C for 30 min with mixing at 1,000 r.p.m. and then purified the DNA with the Qiagen MinElute kit. Libraries were amplified for 6–7 cycles and sequenced on an Illumina HiSeq 4000 to produce an average of 40,315,572 (±14,577,770) reads.

ATAC-seq data processing and peak calling. We processed the sequencing reads and performed quality control by using the ENCODE ATAC-seq Data Standards and Prototype Processing Pipeline. Briefly, we aligned reads to the human reference genome (GRCh37/hg19) with Bowtie2 (ref. [52]) v2.2.9 (with parameters -k 4 -X 2000 --local) and filtered out unpaired mapped reads and reads with MAPQ < 30 (SAMtools[53]) as well as duplicates (marked with Picard Tools). Only reads from the autosomes and X chromosome were retained for downstream analyses.

Identification of differentially accessible ATAC-seq peaks. Read alignments from all untreated human preadipocyte and adipocyte libraries (three biological replicates per cell type) were merged before peak calling. Peaks were called with MACS2 (ref. [54]) v2.1.1 (by using the BEDPE function), and peaks with FDR < 0.05 were retained. We filtered out peaks in blacklisted regions and peaks that did not replicate in two of the three biological replicates in at least one condition. For differential accessibility analyses, we retained peaks with counts per million (c.p.m.) ≥ 1 in at least three libraries. We then input aligned read counts for each peak into cqn[55] v1.20.0 and normalized the counts for G+C content, peak length and library size, before inputting the counts into edgeR[56] v3.16.5 to detect differentially accessible peaks between preadipocytes and adipocytes with the generalized linear model (GLM) functionality and QL *F* test, applying an FDR threshold of 0.05. Libraries for adipocyte lipid challenge ATAC-seq were processed identically until the differential accessibility analysis. To detect open chromatin regions that exhibited differential accessibility in lipid-challenged adipocytes, we performed one-way ANOVA with the GLM functionality and QL *F* test functionality in edgeR[56], applying an FDR threshold of 0.05. For the post hoc test to determine which comparisons were significant after one-way ANOVA (oleic acid versus BSA, palmitic acid versus BSA or oleic acid versus palmitic acid), we determined the least significant difference.

Transcription factor motif enrichment in ATAC peaks. We used HOMER (v4.9)[6] to investigate the enrichment of motif sequences in open chromatin regions. For enrichment in differentially accessible peaks between untreated preadipocytes and adipocytes, we used the consensus peak set (all peaks that were called in both the preadipocyte and adipocyte data) as the background. We utilized the de novo motif enrichment functionality. To ensure that our background input file was not biasing the results, we performed the same analysis with the genome as the background input, which produced largely the same results, with smaller *P* values (data not shown). For enrichment in lipid-responsive peaks in adipocyte promoter–enhancer contacts, we used non-differentially accessible peaks within the promoter–enhancer contacts as the background. Owing to the small number of peaks ($n = 264$), we used the known TF motif enrichment functionality.

Hi-C library preparation. We prepared the Hi-C libraries for the primary human adipocyte lipid challenge experiment as described in Pan et al.[15] in two biological replicates per condition (BSA, oleic acid and palmitic acid). These methods were adapted by closely following the in-nucleus Hi-C methods in Rao et al.[57] and Nagano et al.[58]. Specifically, we fixed 7–10 million adherent cells directly in the culture plate in 2% formaldehyde and quenched with glycine to a final concentration of 125 mM. Cells were lysed in ice-cold lysis buffer (10 nM Tris-HCl pH 8.0, 10 mM NaCl, 0.2% Igepal CA-630, 1× protease inhibitors: cOmplete, EDTA-free Protease Inhibitor cocktail) on ice for 30 min with occasional agitation. We split lysates into aliquots with 5 million nuclei and centrifuged at 2,500*g* for 5 min at 4 °C. The nuclear pellets were resuspended in 50 µl of 0.5% SDS in 1× NEBuffer 2 (New England Biolabs) and incubated at 62 °C for 10 min. We then added 145 µl of water and 25 µl of 10% Triton X-100 and incubated at 37 °C for 15 min. We digested chromatin by adding 25 µl of 10× NEBuffer 2 and 400 U of HindIII (New England Biolabs), incubating at 37 °C overnight with shaking (950 r.p.m.).

The next day, we marked the DNA ends with biotin (1.5 µl of 10 mM dATP, 1.5 µl of 10 mM dGTP, 1.5 µl of 10 mM dTTP, 37.5 µl of 0.4 mM biotin-14-dCTP (Invitrogen), 8 µl of 5 U $µl^{-1}$ Klenow (New England Biolabs)), incubating for 60 min at 37 °C; we then added 895 µl of ligation mix (663 µl of water, 120 µl of 10× NEB T4 DNA ligase buffer, 100 µl of 10% Triton X-100, 12 µl of 10 mg ml^{-1} BSA, 5 µl of 400 U $µl^{-1}$ T4 DNA ligase (New England Biolabs)). Ligation was performed at room temperature for 4 h with slow rotation, and 50 µl of 20 mg ml^{-1} proteinase K and 120 µl of 10% SDS were added with incubation at 55 °C for 30 min. We added 130 µl of 5 M NaCl and incubated at 68 °C overnight. We then performed an ethanol precipitation and sheared the purified DNA to 250–550 bp in size with a Covaris M220 instrument. Double size selection was performed with SPRI select agent (Beckman Coulter) by adding 0.55 volumes and then 0.15 volumes according to the manufacturer's instructions, eluting the final DNA in 300 µl of 10 mM Tris pH 8.0.

Biotin pulldown was performed with 150 µl of 10 mg ml^{-1} DYNAL MyOne Dynabeads Streptavidin T1 (Invitrogen, 65601) per sample. First, the beads were washed twice with 400 µl of 1× Tween wash buffer (1× TWB: 5 mM Tris-HCl pH 7.5, 0.5 mM EDTA, 1 M NaCl, 0.05% Tween-20) and resuspended in 300 µl of 2× binding buffer (2× BB: 10 mM Tris-HCl pH 7.5, 1 mM EDTA, 2 M NaCl). Beads were then added to 300 µl of sheared and size-selected DNA. We incubated at room temperature for 15 min with rotation to bind biotinylated DNA to the streptavidin beads. We washed twice with mixing at 55 °C by adding 600 µl of 1× TWB and then washed beads in 100 µl of 1× NEB T4 DNA ligase buffer. We repaired the ends of the DNA by resuspending beads in 100 µl of master mix (88 µl of 1× NEB T4 DNA ligase buffer with 10 mM ATP, 2 µl of 25 mM dNTP mix, 5 µl of 10 U $µl^{-1}$ NEB T4 PNK, 4 µl of 3 U $µl^{-1}$ NEB T4 DNA polymerase I, 1 µl of 5 U $µl^{-1}$ NEB DNA polymerase I, large (Klenow) fragment) and incubated samples at room temperature for 30 min. We washed twice with 1× TWB, washed once with 1× NEBuffer 2 and then resuspended samples in 100 µl of dATP attachment master mix (90 µl of 1× NEBuffer 2, 5 µl of 10 mM dATP, 5 µl of 5 U $µl^{-1}$ NEB Klenow exo minus) and incubated samples for 30 min at 37 °C. We washed twice with 1× TWB and resuspended beads in 100 µl of 1× T4 DNA ligase buffer. We followed the manufacturer's instructions for the Agilent SureSelect to ligate the paired-end adaptors. The beads were then washed twice with 1× TWB and resuspended in 32 µl of 1× Tris buffer. DNA was removed from the streptavidin beads by heating at 98 °C for 10 min. We followed the manufacturer's instructions for the Agilent SureSelect for precapture PCR, carried out for eight cycles.

Promoter capture Hi-C library preparation. RNA baits were designed in Mifsud et al.[14] for capturing HindIII fragments containing gene promoters (C. Osborne (Department of Medical & Molecular Genetics, King's College London, London, UK.) kindly shared the exact design). As described in Mifsud et al.[14], 120-mer RNA baits were designed to target both ends of HindIII fragments that contained annotated gene promoters (Ensembl promoters of protein-coding, noncoding, antisense, snRNA, miRNA and snoRNA transcripts). A bait sequence was deemed valid if the sequence had a G+C content of 25–65%, contained fewer than three consecutive Ns and was within 330 bp of the ends of the HindIII fragment. Hi-C library hybridization to the capture library was performed according to the manufacturer's instructions for the Agilent SureSelect. A total of 550 ng of the Hi-C library was hybridized to the biotinylated RNA baits, captured with DYNAL MyOne Dynabeads Streptavidin T1 and amplified in post-capture PCR to add index sequences, for 12 PCR cycles. The library was sequenced on the Illumina HiSeq 4000 platform. All six libraries were sequenced together across two lanes of the Illumina HiSeq 4000 to produce an average of 127,069,374 (±16,855,586) sequencing reads per library.

Capture Hi-C data processing and interaction calling. We processed the sequencing reads as described in Pan et al.[15], by using the Hi-C User Pipeline (HiCUP) v0.5.9 (ref. [59]) with default settings except that the insert size restrictions for the filtering step were set to 200–600 bp. We called significant interactions for each library separately with CHiCAGO software v1.1.1 (ref. [60]). We used the default threshold of 5 for calling significant interactions. To create a stringent set of interactions, we merged all pCHi-C final alignments and called interactions with CHiCAGO, again by using a threshold of 5; we then filtered these interactions to include only those that were called in both biological replicates in at least one condition (BSA, oleic acid or palmitic acid).

Cross-species conservation analysis. For the cross-species conservation analysis, we used the PhastCons score[18] available on the UCSC ENCODE database. Briefly, the PhastCons score uses a phylo-HMM to predict per-base conservation across species. We used the PhastCons scores for placental mammal alignment and calculated the mean score for each of the lipid-responsive regions for the protein-coding lipid responsive genes on autosomes (gene body ± 500 kb). To create a background set for this comparison, we calculated the mean score for all other protein-coding genes annotated by Ensembl in GRCh37 and their surrounding regions of ±500 kb. We performed a non-parametric two-sided Wilcoxon signed-rank test to compare the lipid-responsive regions to the background set.

LD score analysis. We used LD score regression[24] to estimate the heritability explained by the lipid-responsive regions. More specifically, we generated an annotation for each lipid-responsive region consisting of the lipid-responsive gene (gene body ± 500 kb) and used the summary statistics from a lipid GWAS[11] to estimate the proportion of heritability explained by the 154 lipid-responsive promoters or 323 promoters that interacted with lipid-responsive enhancers in adipocytes, for the four lipid GWAS traits: serum TG, HDL, LDL and TC.

Genotype and phenotype data from the UK Biobank cohort. We downloaded imputed genotype data from the UK Biobank cohort[4]. We identified all SNPs in the lipid-responsive gene promoters involved in chromosomal interactions. For the G×E interaction test, we filtered out SNPs that had a genotype missing rate of greater than 5% or a MAF of less than 5%. We used the BMI value collected from the initial UK Biobank assessment visit at which participants were recruited. The data for 24-h recall of saturated fat consumption in diet was collected at five different time points, including during the initial assessment and four online cycle collections. If an individual had 24-h recall of saturated fat consumption collected at multiple time points, we used the value closest to the initial assessment. We then selected unrelated individuals of European ancestry from the UK Biobank participants who had data on both BMI and saturated fat diet collected for the G×E analysis. We corrected BMI for the following covariates and performed inverse normal transformation to ensure that BMI was normally distributed: array type, sex (inferred), age (when the participant attended the assessment centre), age^2, the assessment centre ID and genetic principal components 1–20.

Genome-wide G×E scan. To verify that our significant G×E interactions were not caused by overall inflation, we fitted the linear G×E interaction model on all SNPs across the human genome in the UK Biobank data. We first selected SNPs that were not in the same LD block ($r^2 < 0.2$) and then used the following linear model to detect the G×E interaction between each LD-pruned SNP and saturated fat intake on BMI

$$Y = \alpha + \beta_G g + \beta_E e + \beta_{GE} ge + \varepsilon \quad (1)$$

where Y is a vector of inverse normal transformed BMI values and g represents the vector of the number of minor alleles in the genotypes of the target SNP for the individuals in the study sample. We used e for the vector of saturated fat intake levels as the environmental covariate, and ε represents a vector of random errors, in which each entry is independently and normally distributed. α and β are the estimated parameters. The test for an interaction is based on the coefficient β_{GE}. A non-zero β_{GE} value indicates that there is an interaction between the genotype and environmental factor for the outcome phenotype. We constructed a quantile–quantile plot to compare the P values of β_{GE} and the expected P values based on multiple testing.

Testing for G×E interaction by multivariable linear model. We used the following multivariable linear model to detect the G×E interactions between SNPs and saturated fat intake for BMI

$$Y = \alpha + \beta_E e + \sum_{i=1}^{N} (\beta_{Gi} \times g_i) + \sum_{i=1}^{N} (\beta_{GEi} \times g_i e) + \varepsilon \quad (2)$$

where Y is a vector of inverse normal transformed BMI values and g_i represents a vector of the number of minor alleles in the genotypes of the target SNP i, where $i = 1, \ldots, N$ SNPs, for the individuals in the study sample. We use e for the vector of saturated fat intake levels as the environmental covariate, and ε represents a vector of random errors, in which each entry is independently and normally distributed. α and β are the estimated parameters. The significance of the interaction is given for the coefficient β_{GEi}. A non-zero β_{GEi} value indicates that there is an interaction between the genotype i and the environmental factor (24-h saturated fat recall) for the outcome phenotype (BMI). Individual β_{GEi} values are estimated with conditioning on the effects of the other genotypes and β_{Gi} values from the multivariable linear model. The P values given for the individual β_{GEi} values are calculated by t test.

Electrophoretic mobility shift assays. Nuclear protein was extracted from adipocytes with a nuclear protein extraction kit (Active Motif, 40010) following the manufacturer's instructions. Oligonucleotide probes (corresponding to the 15 bp flanking a SNP site for the reference or alternate allele; Supplementary Table 21) with a biotin tag at the 5′ end of the forward sequence (Integrated DNA Technologies) were incubated with human adipocyte nuclear protein and the working reagent from the Gelshift Chemiluminescent EMSA kit (Active Motif, 37341). For competitor assays, an unlabelled probe of the same sequence was added to the reaction mixture at 100× excess. The reaction was incubated for 30 min at room temperature and then loaded on a 6% retardation gel (Thermo Fisher Scientific, EC6365BOX) that was run in 0.5× TBE buffer. We transferred the contents of the gel to a nylon membrane and visualized signal with chemiluminescent reagent as recommended.

***Cis*-eQTLs in the METSIM cohort.** We obtained subcutaneous adipose *cis*-eQTL variants identified in RNA-seq data ($n = 335$) from the Finnish METSIM cohort[15,31].

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The ATAC-seq data for primary human preadipocytes and adipocytes (untreated and lipid-challenged cells) and the pCHi-C data for primary human adipocytes under lipid-challenge conditions have been deposited in the Gene Expression Omnibus under accession GSE129574 and are available upon request from the corresponding author.

Received: 29 November 2018; Accepted: 30 April 2019;
Published online: 14 June 2019

References

1. Joseph, P. G., Pare, G. & Anand, S. S. Exploring gene–environment relationships in cardiovascular disease. *Can. J. Cardiol.* **29**, 37–45 (2013).
2. Heianza, Y. & Qi, L. Gene–diet interaction and precision nutrition in obesity. *Int. J. Mol. Sci.* **18**, E787 (2017).
3. Kilpeläinen, T. O. et al. Physical activity attenuates the influence of *FTO* variants on obesity risk: a meta-analysis of 218,166 adults and 19,268 children. *PLoS Med.* **8**, e100116 (2011).
4. Sudlow, C. et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
5. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
6. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
7. Lefterova, M. I. & Lazar, M. A. New developments in adipogenesis. *Trends Endocrinol. Metab.* **20**, 107–114 (2009).
8. Ernst, J. & Kellis, M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat. Biotechnol.* **33**, 364–376 (2015).
9. Surakka, I. et al. The impact of low-frequency and rare variants on lipid levels. *Nat. Genet.* **47**, 589–597 (2015).
10. Teslovich, T. M. et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
11. Willer, C. J. et al. Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1283 (2013).
12. Kanai, M. et al. Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat. Genet.* **50**, 390–400 (2018).
13. MacArthur, J. et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).
14. Mifsud, B. et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* **47**, 598–606 (2015).
15. Pan, D. Z. et al. Integration of human adipocyte chromosomal interactions with adipose gene expression prioritizes obesity-related genes from GWAS. *Nat. Commun.* **9**, 1512 (2018).
16. Liu, X. et al. Functional architectures of local and distal regulation of gene expression in multiple human tissues. *Am. J. Hum. Genet.* **100**, 605–616 (2017).
17. Wang, J., Duncan, D., Shi, Z. & Zhang, B. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Res.* **41**, W77–W83 (2013).
18. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
19. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
20. Kathiresan, S. et al. Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat. Genet.* **41**, 56–65 (2009).

21. Shin, S.-Y. et al. An atlas of genetic influences on human blood metabolites. *Nat. Genet.* **46**, 543–550 (2014).
22. Kettunen, J. et al. Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nat. Commun.* **7**, 11122 (2016).
23. Vaittinen, M. et al. *FADS2* genotype regulates delta-6 desaturase activity and inflammation in human adipose tissue. *J. Lipid Res.* **57**, 56–65 (2016).
24. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
25. Nettleton, J. A., Brouwer, I. A., Geleijnse, J. M. & Hornstra, G. Saturated fat consumption and risk of coronary heart disease and ischemic stroke: a science update. *Ann. Nutr. Metab.* **70**, 26–33 (2017).
26. Casas-Agustench, P. et al. Saturated fat intake modulates the association between an obesity genetic risk score and body mass index in two US populations. *J. Acad. Nutr. Diet.* **114**, 1954–1966 (2014).
27. Luukkonen, P. K. et al. Saturated fat is more metabolically harmful for the human liver than unsaturated fat or simple sugars. *Diabetes Care* **41**, 1732–1739 (2018).
28. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
29. Robinson, M. R. et al. Genotype–covariate interaction effects and the heritability of adult body mass index. *Nat. Genet.* **49**, 1174–1181 (2017).
30. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat. Methods* **12**, 931–934 (2015).
31. Laakso, M. et al. METabolic Syndrome In Men (METSIM) Study: a resource for studies of metabolic and cardiovascular diseases. *J. Lipid Res.* **58**, 481–493 (2017).
32. Mancuso, N. et al. Integrating gene expression with summary association statistics to identify genes associated with 30 complex traits. *Am. J. Hum. Genet.* **100**, 473–487 (2017).
33. Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
34. Ardlie, K. G. et al. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
35. Ehrlund, A. et al. The cell-type specific transcriptome in human adipose tissue and influence of obesity on adipocyte progenitors. *Sci. Data* **4**, 170164 (2017).
36. Acosta, J. R. et al. Single cell transcriptomics suggest that human adipocyte progenitor cells constitute a homogeneous cell population. *Stem Cell Res. Ther.* **8**, 250 (2017).
37. Zheng, J., Chen, S., Chen, Y., Zhu, M. & Hong, D. A novel mutation in the PDZ-like motif of ZASP causes distal ZASP-related myofibrillar myopathy. *Neuropathology* **37**, 45–51 (2017).
38. Griggs, R. et al. Zaspopathy in a large classic late-onset distal myopathy family. *Brain* **130**, 1477–1484 (2007).
39. Fumagalli, M. et al. Greenlandic Inuit show genetic signatures of diet and climate adaptation. *Science* **349**, 1343–1347 (2015).
40. Lu, Y. et al. Dietary n-3 and n-6 polyunsaturated fatty acid intake interacts with *FADS1* genetic variation to affect total and HDL-cholesterol concentrations in the Doetinchem Cohort Study. *Am. J. Clin. Nutr.* **92**, 258–265 (2010).
41. He, Z. et al. *FADS1–FADS2* genetic polymorphisms are associated with fatty acid metabolism through changes in DNA methylation and gene expression. *Clin. Epigenetics* **10**, 113 (2018).
42. Vernekar, M. & Amarapurkar, D. Diet–gene interplay: an insight into the association of diet and *FADS* gene polymorphisms. *J. Nutr. Food Sci.* **6**, 503 (2016).
43. Dunham, I. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
44. Bujold, D. et al. The International Human Epigenome Consortium Data Portal. *Cell Syst.* **3**, 496–499 (2016).
45. Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
46. Bernstein, B. E. et al. The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* **28**, 1045–1048 (2010).
47. Simeonov, D. R. et al. Discovery of stimulation-responsive immune enhancers with CRISPR activation. *Nature* **549**, 111–115 (2017).
48. Gate, R. E. et al. Genetic determinants of co-accessible chromatin regions in activated T cells across humans. *Nat. Genet.* **50**, 1140–1150 (2018).
49. Phan, A. T., Goldrath, A. W. & Glass, C. K. Metabolic and epigenetic coordination of T cell and macrophage immunity. *Immunity* **46**, 714–729 (2017).
50. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).
51. Corces, M. R. et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14**, 959–962 (2017).
52. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
53. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
54. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
55. Hansen, K. D., Irizarry, R. A. & Wu, Z. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics* **13**, 204–216 (2012).
56. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
57. Rao, S. S. P. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
58. Nagano, T. et al. Comparison of Hi-C results using in-solution versus in-nucleus ligation. *Genome Biol.* **16**, 175 (2015).
59. Wingett, S. W. et al. HiCUP: pipeline for mapping and processing Hi-C data. *F1000Res* **4**, 1310 (2015).
60. Cairns, J. et al. CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. *Genome Biol.* **17**, 127 (2016).

Acknowledgements

This research has been conducted using the UK Biobank Resource under application number 33934. We thank the individuals who participated in the METSIM and UK Biobank studies. We also thank the UNGC sequencing core at UCLA for performing the DNA and RNA sequencing. This study was funded by National Institutes of Health (NIH) grants HL-095056, HL-28481 and U01DK105561. K.M.G. was supported by NIH-NHLBI grant 1F31HL142180, M.A. was supported by an HHMI Gilliam grant, D.Z.P. was supported by NIH-NCI grant T32LM012424 and NIH-NIDDK grant F31DK118865, and A.K. was supported by NIH grant F31HL127921. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the article.

Author contributions

K.M.G. and P.P. designed the study. K.M.G., D.Z.P., Z.M., J.R.P., C.J.Y., J.S.S. and P.P. performed methods development and statistical analysis. K.M.G., D.Z.P., Z.M., M.A. and C.R.R. performed computational analysis of the data. K.M.G., Y.V.B., M.A., C.C. and J.N.B. performed the experiments. M.L., K.M. and P.P. produced the METSIM RNA-seq data. A.K. performed quality control of the METSIM RNA-seq data. K.M.G. and P.P. wrote the manuscript and all authors read, reviewed and/or edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s42255-019-0071-6.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to P.P.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature metabolism

SUPPLEMENTARY INFORMATION

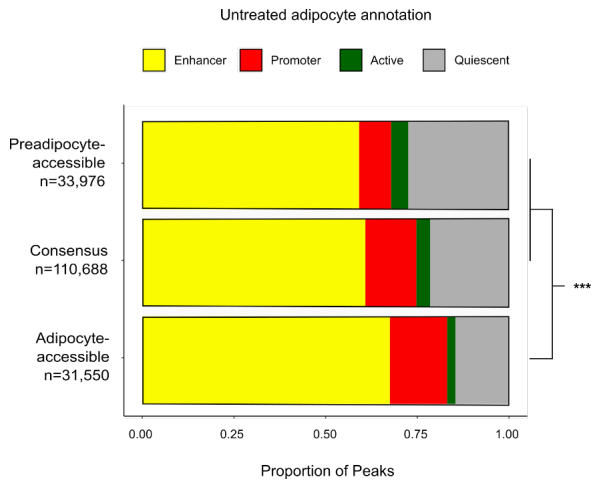
https://doi.org/10.1038/s42255-019-0071-6

In the format provided by the authors and unedited.

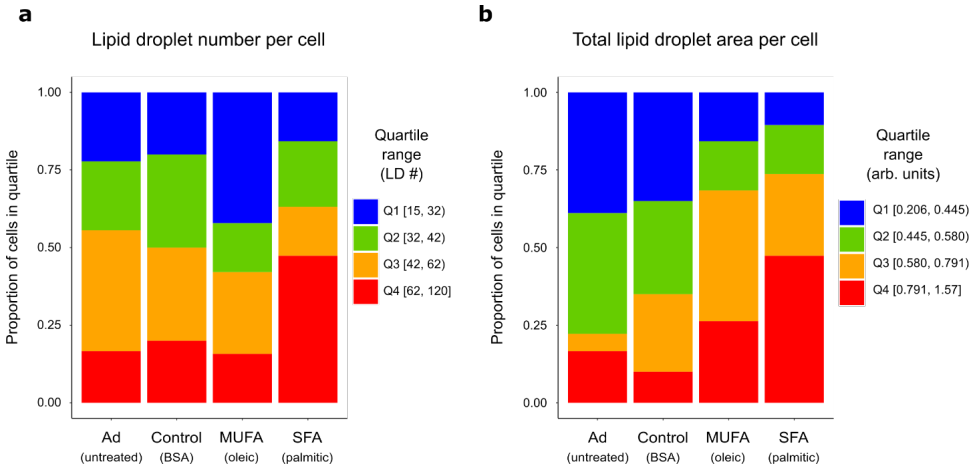
Reverse gene–environment interaction approach to identify variants influencing body-mass index in humans

Kristina M. Garske[1], David Z. Pan[1,2], Zong Miao[1,2], Yash V. Bhagat[1], Caroline Comenho[1], Christopher R. Robles[3], Jihane N. Benhammou[1,4], Marcus Alvarez[1], Arthur Ko[5], Chun Jimmie Ye[6], Joseph R. Pisegna[1,4], Karen L. Mohlke[7], Janet S. Sinsheimer[1,8], Markku Laakso[9] and Päivi Pajukanta[1,2,9]*

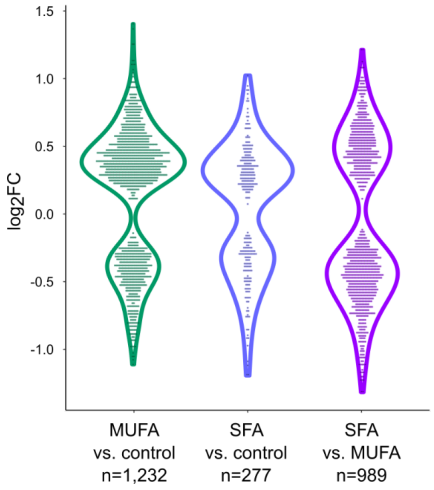
[1]Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA. [2]Bioinformatics Interdepartmental Program, UCLA, Los Angeles, CA, USA. [3]Department of Computer Science, UCLA, Los Angeles, CA, USA. [4]Vache and Tamar Manoukian Division of Digestive Diseases, UCLA, Los Angeles, CA, USA. [5]Department of Medicine, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA. [6]Institute for Human Genetics, Department of Epidemiology and Biostatistics and Department of Bioengineering and Therapeutic Sciences, UCSF, San Francisco, CA, USA. [7]Department of Genetics, University of North Carolina, Chapel Hill, NC, USA. [8]Department of Biomathematics, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA. [9]Internal Medicine, Institute of Clinical Medicine, University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland.
*e-mail: ppajukanta@mednet.ucla.edu



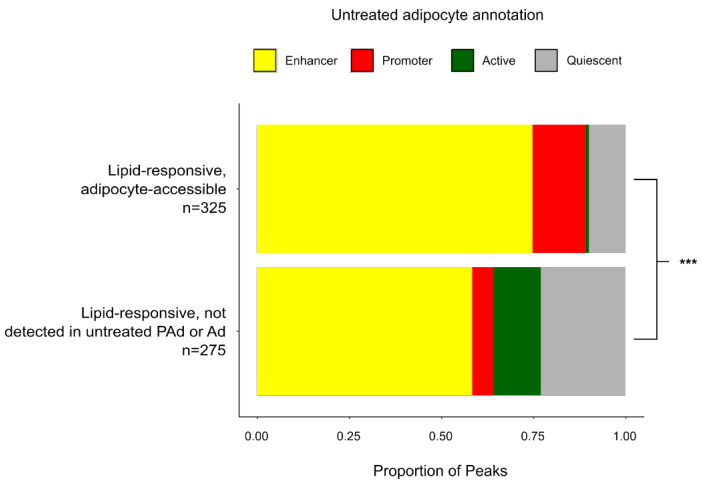
Supplementary Figure 1. Adipocyte-accessible peaks fall more into adipocyte enhancers and promoters than the preadipocyte-accessible peaks or the full peak set. ATAC-seq peaks from the indicated peak sets on the y-axis are distributed among four subsets of functional annotations from the 25-state imputed chromHMM[1] annotations from mesenchymal stem cell derived cultured adipocytes. Note that not all peaks were categorized into one of these 4 categories due to minimum peak proportion overlap (>50%) requirement not being met. ***depicts the *p*-value ($p < 1x10^{-5}$) for the chi-square test for independence between the distributions of peaks in the indicated annotations. Related to Figure 1.



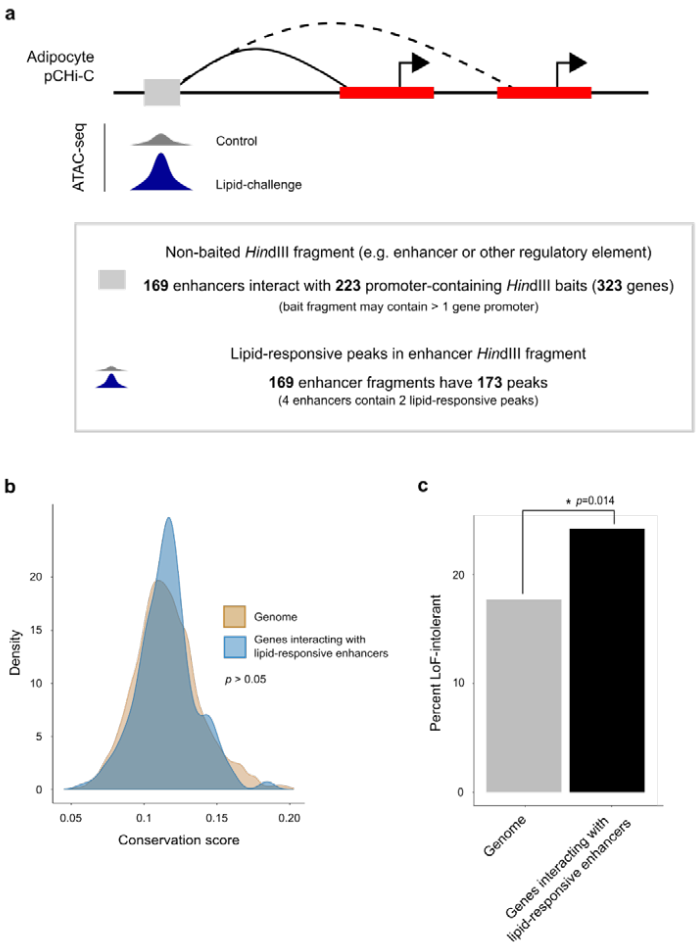
Supplementary Figure 2. Fatty acid lipid challenge in human adipocytes leads to increased storage of lipids in lipid droplets. (**a**,**b**) The proportion of cells in each of the indicated quartiles are reported for (**a**) lipid droplet (LD) number per cell, and (**b**) total LD area per cell, quantified from oil red o staining. Treatment with monounsaturated fatty acid (MUFA) leads to increased total area of LD but fewer total LDs (e.g. large LDs). Treatment with saturated fatty acid (SFA) leads to increased LD number and size. Data presented are from one representative experiment out of two independent experiments with similar results. Related to Figure 2.



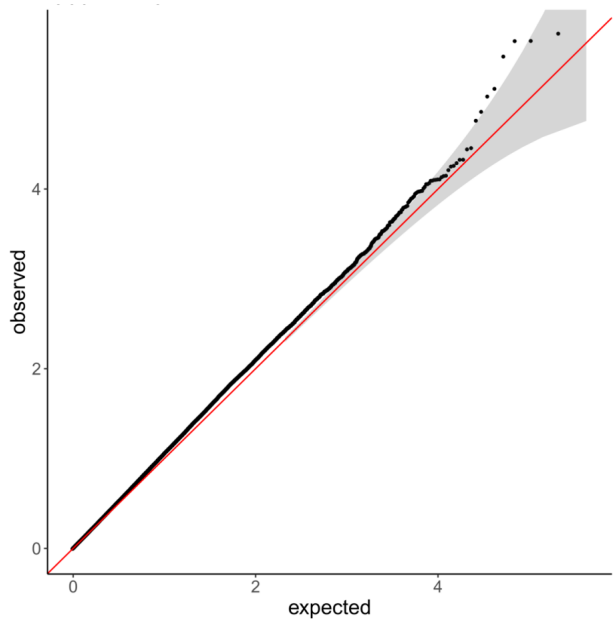
Supplementary Figure 3. Violin plots show the distribution of log2 fold-change (log2FC) for all differentially accessible peaks from the lipid challenge in adipocytes. Peaks were considered differentially accessible at a cutoff of FDR < 0.05. FDR was calculated (adjusting for n=122,252 ATAC-seq peaks) from the *p*-values of the QL F-test (see Methods) in the one-way ANOVA. For the *post hoc* test to determine which comparison was significant after the one-way ANOVA (OA vs. BSA, PA vs. BSA, or OA vs. PA), we determined the least significant difference. The violin plot characteristics are as follows: MUFA vs. ctrl (n=1,232) range: -1.11 – 1.40; median: 0.32; 25th percentile: -0.32; and 75th percentile: 0.47. SFA vs. ctrl (n=277) range: -1.19 – 1.02; median: 0.22; 25th percentile: -0.30; and 75th percentile: 0.37. SFA vs. MUFA (n=989) range: -1.31 – 1.21; median: -0.27; 25th percentile: -0.51; and 75th percentile: 0.47. MUFA indicates monounsaturated fatty acid; SFA, saturated fatty acid. Related to Figure 2.



Supplementary Figure 4. Lipid-responsive peaks in adipocyte-accessible regions fall more into adipocyte enhancers and promoters than lipid-responsive peaks in context-dependent regions. Lipid-responsive ATAC-seq peaks from the indicated peak sets on the y-axis are distributed among four subsets of functional annotations from the 25-state imputed chromHMM[1] annotations from mesenchymal stem cell derived culture adipocytes. Note that not all peaks were categorized into one of these 4 categories due to minimum peak proportion overlap (>50%) requirement not being met. ***depicts the *p*-value ($p < 1x10^{-5}$) for the chi-square test for independence between the distributions of peaks in the indicated annotations. Related to Figure 2.



Supplementary Figure 5. The 323 genes with promoters that interact with lipid-responsive enhancers exhibit constraints on loss-of-function mutations. (**a**) Schematic overview of the lipid-responsive sites in non-baited *Hin*dIII fragments from the adipocyte pCHi-C interactions. These data were integrated to identify the 323 gene promoters that interact with lipid-responsive enhancers in adipocytes. (**b**) Density plot shows the distribution of per-gene average conservation scores across placental mammals[2] for all protein-coding genes in the genome compared to all protein-coding genes in the set of 323 genes whose promoters interact with lipid-responsive enhancers. The two-sided Wilcoxon signed-rank test returned a non-significant *p*-value > 0.05. (**c**) Bar graph shows the proportion of protein-coding genes that are loss-of-function intolerant (i.e. are unlikely to have protein-truncating variants in humans)[3] in the whole genome (n=3,204/18,122; 17.7%) compared to the protein-coding genes among the 323 genes (n=50/207; 24.2%). LoF indicates loss-of-function; *depicts the *p*-value for the hypergeometric enrichment test. Compare with Figure 3 results for 154 genes with lipid-responsive promoters.



Supplementary Figure 6. Testing all SNPs genome-wide for gene-by-saturated fat intake effect on BMI does not show inflation or result in significant GxEs at the genome-wide significance threshold. We tested the SNPs that are not in the same LD block ($r^2 < 0.2$) genome-wide for a GxE between each SNP and saturated fat intake effect on BMI. There were a total of 211,187 SNPs and 167,908 individuals with no missing data available for study (see Equation 1 in the Methods). The Q-Q plot shows the observed *p*-values of β_{GE}, the expected p-values (red line) based on the multiple testing, and the 95% confidence interval (shaded area).

Supplementary Table 1. Sequencing, read processing, and QC metrics for untreated preadipocyte and adipocyte ATAC-seq.

	Reads	Uniquely aligned	Fraction uniquely aligned	Paired and filtered	De-duplicated	Fraction duplicates	Final Reads	Fraction mtDNA	Fraction reads in peaks
PAd rep1	23,240,008	16,960,464	0.73	20,403,898	16,707,418	0.18	15,810,810	0.019	0.48
PAd rep2	25,705,628	18,552,651	0.72	22,516,812	17,829,509	0.21	16,938,123	0.016	0.46
PAd rep3	27,548,038	19,008,502	0.69	23,815,866	19,142,102	0.20	17,984,628	0.025	0.47
Ad rep1	24,407,273	16,893,473	0.69	21,049,924	16,045,835	0.24	15,038,768	0.033	0.58
Ad rep2	18,190,931	12,109,272	0.67	15,537,582	11,344,211	0.27	10,459,715	0.049	0.53
Ad rep3	21,165,864	13,976,310	0.66	17,956,825	12,512,132	0.30	11,528,741	0.050	0.70

PAd indicates preadipocyte; Ad, adipocyte; mtDNA, mitochondrial DNA.

Supplementary Table 3. Sequencing, read processing, and QC metrics for adipocyte lipid-challenge ATAC-seq.

	Reads	Uniquely aligned	Fraction uniquely aligned	Paired and filtered	De-duplicated	Fraction duplicates	Final Reads	Fraction mtDNA	Fraction reads in peaks
BSA rep1	64,706,941	48,768,335	0.76	56,012,019	45,623,026	0.185	44,927,691	0.012	0.66
BSA rep2	24,133,180	18,644,383	0.77	21,149,557	17,972,138	0.150	17,655,287	0.015	0.66
BSA rep3	23,981,457	17,668,336	0.74	20,677,446	17,553,162	0.151	17,214,846	0.016	0.59
OA rep1	46,775,100	35,874,412	0.77	40,975,870	29,562,701	0.279	29,007,522	0.016	0.70
OA rep2	57,372,688	44,971,629	0.79	51,073,123	39,246,428	0.232	38,742,718	0.010	0.71
OA rep3	33,118,575	25,516,115	0.77	29,143,404	23,401,224	0.197	22,995,249	0.014	0.67
PA rep1	27,688,462	21,034,204	0.76	24,089,236	18,199,958	0.244	17,822,393	0.018	0.64
PA rep2	45,904,086	35,705,224	0.78	40,386,681	31,318,153	0.225	30,874,420	0.011	0.68
PA rep3	39,159,661	29,810,602	0.76	34,234,012	26,610,078	0.223	26,074,118	0.017	0.68

BSA indicates bovine serum albumin; OA, oleic acid; PA, palmitic acid; mtDNA, mitochondrial DNA.

Supplementary Table 5. Sequencing and read processing metrics for adipocyte lipid-challenge pCHi-C.

	Reads	Uniquely mapped and paired	Unique ditags	Cis-close	Cis-far	Trans
BSA rep1	156,781,294	115,665,713	69,664,503	10,560,429	49,905,501	9,198,573
BSA rep2	126,772,704	94,717,631	51,631,011	8,335,003	36,631,083	6,664,925
OA rep1	120,985,559	90,005,610	54,155,183	7,477,013	38,909,160	7,769,010
OA rep2	118,632,574	89,101,812	49,035,437	8,495,017	34,445,736	6,094,684
PA rep1	132,242,479	98,448,728	54,891,280	8,392,648	39,146,001	7,352,631
PA rep2	107,001,633	80,596,383	44,405,325	6,636,844	31,576,330	6,192,151

BSA indicates bovine serum albumin; OA, oleic acid; PA, palmitic acid.

Supplementary Table 6. The top 10 TF motifs enriched in adipocyte lipid-responsive open chromatin regions in chromosomal interactions.

Motif logo	Motif name	*p*-value	Adjusted *p*-value	Number of target sequences with motif (of 264)	Percent of target sequences with motif	Number of background sequences with motif (of 30,704)	Percent of background sequences with motif
	RUNX-AML(Runt)/CD4+-PolII-ChIP-Seq(Barski_et_al.)/Homer	$1.0x10^{-8}$	0	134.0	50.76%	10267.5	33.44%
	STAT4(Stat)/CD4-Stat4-ChIP-Seq(GSE22104)/Homer	$1.0x10^{-8}$	0	166.0	62.88%	13952.6	45.44%
	PPARE(NR),DR1/3T3L1-Pparg-ChIP-Seq(GSE13511)/Homer	$1.0x10^{-6}$	0	162.0	61.36%	14012.1	45.64%
	RXR(NR),DR1/3T3L1-RXR-ChIP-Seq(GSE13511)/Homer	$1.0x10^{-6}$	0	173.0	65.53%	15537.1	50.60%
	ZBTB18(Zf)/HEK293-ZBTB18.GFP-ChIP-Seq(GSE58341)/Homer	$1.0x10^{-6}$	$1.0x10^{-4}$	108.0	40.91%	8317.5	27.09%
	THRa(NR)/C17.2-THRa-ChIP-Seq(GSE38347)/Homer	$1.0x10^{-5}$	$1.0x10^{-4}$	105.0	39.77%	8090.0	26.35%
	KLF10(Zf)/HEK293-KLF10.GFP-ChIP-Seq(GSE58341)/Homer	$1.0x10^{-5}$	$1.0x10^{-4}$	108.0	40.91%	8567.6	27.90%
	Srebp1a(bHLH)/HepG2-Srebp1a-ChIP-Seq(GSE31477)/Homer	$1.0x10^{-5}$	$1.0x10^{-4}$	59.0	22.35%	3792.7	12.35%
	NF1-halfsite(CTF)/LNCaP-NF1-ChIP-Seq(Unpublished)/Homer	$1.0x10^{-4}$	$3.0x10^{-4}$	215.0	81.44%	21406.9	69.72%
	NPAS2(bHLH)/Liver-NPAS2-ChIP-Seq(GSE39860)/Homer	$1.0x10^{-4}$	$3.0x10^{-4}$	152.0	57.58%	13632.6	44.40%

Enrichment *p*-values were derived from the hypergeometric enrichment test of proportion of the given TF motif in the peak set [lipid-responsive open chromatin regions in adipocyte chromosomal interactions (n=264)] compared with the background set of peaks [all non-lipid-responsive peaks in adipocyte chromosomal interactions (n=30,704)], adjusted (Benjamini-Hochberg) for the number of known motifs tested (n=364)[4]. The top 10 enriched TF motifs in include key TFs in lipid metabolism, such as the co-factors PPARG and RXR. Related to Figure 3.

Supplementary Table 8. KEGG pathway enrichment analysis of 154 genes with lipid-responsive promoters.

KEGG pathway	Ratio of enrichment	Number of genes	Genes in pathway	FDR
Glycine, serine and threonine metabolism	13.96	5	*AGXT2* *AOC2* *AOC3* *GLYCTK* *MAOA*	0.0072
Phenylalanine metabolism	23.46	3	*AOC2* *AOC3* *MAOA*	0.036

The 154 genes with lipid-responsive promoters in adipocyte chromosomal interactions were tested for KEGG pathway enrichment using WebGetstalt[5], using all genes that were involved in adipocyte chromosomal interactions (n=17,052) as the background set. The FDR is calculated from the *p*-values of the hypergeometric test, adjusted for the number of pathways tested through WebGestalt. Related to Figure 3.

Supplementary Table 10. Three lipid-responsive ATAC-peaks in interacting enhancers overlap with GWAS SNPs[6] for serum lipid traits.

peakChr	peakStart	peakEnd	intBaitGene[¶]	SNP in peak	MAF[†]	Associated trait[6]	*p*-value[6]	Index SNP[6]	LD with index SNP[†] (r^2)
7	73015109	73016308	Bait1: *WBSCR22* Bait2: *STX1A*	rs34346326	0.2	TG	1.31e-44	rs17145738	0.5
15	58591111	58592050	Bait1: *ADAM10* Bait2: *RP11-30K9.7/ U3.10*	rs12899879	0.14	HDL	3.55e-09	rs1532085	-
10	113902081	113908608	*ADRA2A*	rs2792744	0.28	TC	2.73e-09	rs2255141	0.77

[¶]The genes listed are the promoters in the baited *Hin*dIII fragment with which the lipid-responsive enhances interact. More than one bait is listed when the lipid-responsive enhancer is interacting with more than one bait in the adipocyte pCHi-C; [†]Minor allele frequency (MAF) is the European frequency from the 1000 Genomes Project. Linkage disequilibrium (LD) is calculated based on Europeans in the 1000 Genomes Project; LD calculations > 0.2 are reported. Lipid-responsive ATAC-seq peaks that land in enhancers within adipocyte chromosomal interactions (n=173) were assessed for whether they contain GWAS SNPs for serum lipid traits from the meta-GWAS performed in Willer et al.[6]

Supplementary Table 11. Lipid-responsive gene promoters with GWAS SNPs respond to SFA treatment.

peakChr	peakStart	peakEnd	gene-pCHi-C bait	log_2FC (MUFA/ctrl)	log_2FC (SFA/ctrl)	log_2FC (SFA/MUFA)
11	61594652	61596828	*FADS2/ FADS1*	n.s	n.s	**0.37**
7	73036880	73038991	*MLXIPL*	n.s.	n.s	**0.29**
2	27432323	27432971	*SLC5A6/ ATRAID*	n.s	n.s	**0.33**
16	68115758	68116375	*NFATC3*	n.s	n.s	**0.48**
19	10981139	10983631	*CARM1*	n.s	**0.24**	n.s

MUFA indicates monounsaturated fatty acid; SFA, saturated fatty acid; ctrl, control; n.s., non-significant (based on the *post hoc* test of the one-way ANOVA, see below). Lipid-responsive ATAC-seq peaks that land in promoters within adipocyte chromosomal interactions (n=91) were assessed for whether they contain GWAS SNPs for serum lipid traits from the meta-GWAS performed in Willer et al.[6] The direction of the ATAC-seq differential accessibility effect was then assessed based on the quality (e.g. SFA or MUFA) of the fatty acid. Differential accessibility was evaluated at an FDR cutoff of 0.05. FDR was calculated (adjusting for n=122,252 ATAC-seq peaks) from the *p*-values of the QL F-test (see Methods) in the one-way ANOVA. For the *post hoc* test to determine which comparison was significant after the one-way ANOVA (MUFA vs. ctrl, SFA vs. ctrl, or SFA vs. MUFA), we determined the least significant difference. The lipid-responsive gene promoters in chromosomal interactions that contain GWAS SNPs exhibit increased accessibility in palmitic acid (saturated fatty acid) lipid challenge. Related to Table 1.

Supplementary Table 12. Lipid-responsive enhancers with GWAS SNPs stratified by quality of fatty acid.

peakChr	peakStart	peakEnd	intBaitGene[¶]	log_2FC (MUFA/ctrl)	log_2FC (SFA/ctrl)	log_2FC (SFA/MUFA)
7	73015109	73016308	Bait1: *WBSCR22* Bait2: *STX1A*	n.s	**0.45**	**0.64**
15	58591111	58592050	Bait1: *ADAM10* Bait2: *RP11-30K9.7/ U3.101*	**0.45**	n.s	**-0.28**
10	113902081	113908608	*ADRA2A*	n.s	**0.50**	n.s.

[¶]The genes listed are the promoters in the baited *Hin*dIII fragment with which the lipid-responsive enhancers interact. More than one bait is listed when the lipid-responsive enhancer is interacting with more than one bait in the adipocyte pCHi-C; MUFA indicates monounsaturated fatty acid; SFA, saturated fatty acid; ctrl, control; n.s., non-significant (based on the *post hoc* test of the one-way ANOVA, see below). Lipid-responsive ATAC-seq peaks that land in enhancers within adipocyte chromosomal interactions (n=173) were assessed for whether they contain GWAS SNPs for serum lipid traits from the meta-GWAS performed in Willer et al.[6] The direction of the effect was then assessed based on the quality (e.g. SFA or MUFA) of the fatty acid. Differential accessibility was evaluated at an FDR cutoff of 0.05. FDR was calculated (adjusting for n=122,252 ATAC-seq peaks) from the *p*-values of the QL F-test (see Methods) in the one-way ANOVA. For the *post hoc* test to determine which comparison was significant after the one-way ANOVA (MUFA vs. ctrl, SFA vs. ctrl, or SFA vs. MUFA), we determined the least significant difference. The lipid-responsive enhancers in chromosomal interactions that contain GWAS SNPs are more often differentially accessible in palmitic acid (saturated fatty acid) lipid challenge. Related to Supplementary Table 10.

Supplementary Table 13. LDSC analysis[7] of SNPs in *cis* regions of the 154 lipid-responsive promoters.

Category	Proportion of SNPs	Proportion of h^2	Proportion of h^2 SE	Enrichment	Enrichment SE	Enrichment *p*-value
TC	0.029	0.086	0.021	2.95	0.74	0.0088
LDL-C	0.029	0.084	0.026	2.91	0.91	0.038
HDL-C	0.029	0.083	0.027	2.87	0.92	0.042
Serum TG	0.029	0.085	0.026	2.93	0.90	0.045

SE indicates standard error; h^2, heritability; TC, total cholesterol; LDL-C, low-density lipoprotein cholesterol; HDL-C, high-density lipoprotein cholesterol; TG, triglycerides. LD score regression (LDSC)[7] was performed using the SNPs in the *cis* regions (gene body +/- 500 kb) of the 154 genes with lipid-responsive promoters in adipocyte chromosomal interactions; and the serum lipid trait summary statistics from the meta-GWAS performed in Willer et al.[6] The enrichment *p*-value and the SE for the proportion of h^2 and enrichment were calculated from block jackknife resampling used in the LDSC method. The *p*-value reported is not adjusted for multiple tests as these serum lipid traits are highly correlated. The *cis* regions of the 154 genes with lipid-responsive promoters in adipocyte chromosomal interactions contribute significantly to the heritability of serum lipid traits in humans.

Supplementary Table 14. LDSC analysis[7] of SNPs in *cis* regions of genes with lipid-responsive enhancers.

Category	Proportion of SNPs	Proportion of h^2	Proportion of h^2 SE	Enrichment	Enrichment SE	Enrichment *p*-value
TC	0.055	0.078	0.015	1.41	0.28	0.12
LDL-C	0.055	0.057	0.0089	1.03	0.16	0.86
HDL-C	0.055	0.11	0.020	1.94	0.36	0.011
Serum TG	0.055	0.12	0.041	2.10	0.74	0.15

SE indicates standard error; h^2, heritability; TC, total cholesterol; LDL-C, low-density lipoprotein cholesterol; HDL-C, high-density lipoprotein cholesterol; TG, triglycerides. LD score regression (LDSC)[7] was performed using the SNPs in the *cis* regions (gene body +/- 500 kb) of the 323 genes with promoters that interact with lipid-responsive enhancers in adipocyte chromosomal interactions; and serum lipid trait summary statistics from the meta-GWAS performed in Willer et al.[6] The enrichment *p*-value and the SE for the proportion of h^2 and enrichment were calculated from block jackknife resampling used in the LDSC method. The *p*-value reported is not adjusted for multiple tests as these serum lipid traits are highly correlated. The *cis* regions of the 323 genes that interact with lipid-responsive enhancers in adipocytes contribute significantly to the heritability of HDL, but not the other serum lipid traits, in humans.

Supplementary Table 16. Significant GxE promoter SNPs with LD proxies.

SNP	*p*-g	*p*-g*e	β-g	β-g*e	Genes in Bait	*cis*-eQTL FDR[8†]	Target Gene[8]	log_2FC[8]
rs1974817	0.0089	0.0010	2.3	-0.089	*GLTSCR2/*	2.4E-31	*SEPW1*	0.73
rs2334290	0.22	0.0011	-1.65	0.088	*SNORD23*	1.2E-08	*SEPW1*	0.46
rs112438892[¶]	0.0017	0.0050	-0.50	0.012	*CARM1*	2.7E-05/ 0.015	*SMARCA4/ ICAM4*	-0.28/ 0.38
rs117569851[¶]	0.0037	0.016	0.461	-0.011		0.0014/ 0.046	*SMARCA4/ KRI1*	-0.21/ 0.18
rs35678764	0.028	0.025	-5.7	0.19	*RDH8/COL5A3*	0.012	*OLFM2*	-0.67
rs66516040[±]	0.034	0.026	0.19	-0.0057		0.047	*PIN1*	-0.24
rs9797822	0.044	0.039	5.2	-0.17				
rs10788522	0.013	0.027	3.9	-0.14	*LDB3*	6.1E-04	*LDB3*	0.42
rs2354358	0.033	0.042	-3.5	0.14		6.1E-04	*LDB3*	0.42
rs867773	0.31	0.033	-2.0	0.11	*PLIN2*	0.048	*PLIN2*	-0.39
rs12379376	0.19	0.034	2.1	-0.11		0.048	*PLIN2*	-0.39

[†]The *cis*-eQTLs were identified in the adipose tissue from the METSIM cohort[8,9]. [±]This SNP is the only genome-wide significant *cis*-eQTL from the set of GxE SNPs with LD $r^2 > 0.2$ in the lipid-responsive peak. [¶]These GxE SNPs are *cis*-eQTLs for more than one gene. For 5 of the significant promoter GxE SNPs listed in Table 2, SNPs with LD $r^2 > 0.2$ in the lipid-responsive region that also exhibited a significant GxE effect of saturated fat intake on BMI are listed. Redundant SNPs are listed together in order of more to less significant. The reported *p*-values are from the βs in the multi-variable linear model (see Equation 2 in the Methods), where g is the number of minor alleles of the genotype and e is saturated fat intake. Here *p*-g indicates the *p*-value for the genotype effect; *p*-g*e, the *p*-value for the GxE effect; beta values follow the same notation. For the multi-variable linear model, there were a total of 290 SNPs and 38,394 individuals with no missing data available for study. Related to Table 2.

Supplementary Table 21. EMSA oligo probes used for analysis of GxE SNP rs10788522.

Probe Name	Probe Sequence
rs10788522 FWD labeled	biotin - 5'- TCTGGGGAGAGGAAG**G/A**GGGACAGGCTGAGAC - 3'
rs10788522 FWD unlabeled	5'- TCTGGGGAGAGGAAG**G/A**GGGACAGGCTGAGAC - 3'
rs10788522 REV unlabeled	5' - GTCTCAGCCTGTCCC**C/T**CTTCCTCTCCCCAGA - 3'

Oligonucleotides were designed to target the GxE SNP rs10788522 in the *LDB3* promoter *Hin*dIII fragment (+/- 15 bp).

Supplementary References

1. Ernst, J. & Kellis, M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat. Biotechnol.* **33**, 364–76 (2015).

2. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).

3. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).

4. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).

5. Wang, J., Duncan, D., Shi, Z. & Zhang, B. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Res.* **41**, W77–W83 (2013).

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

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

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

9. Laakso, M. *et al.* METabolic Syndrome In Men (METSIM) Study: a resource for studies of metabolic and cardiovascular diseases. *J. Lipid Res.* **58**, 481–493 (2017).

Chapter 4

BMI-discordant monozygotic twin preadipocytes exhibit altered compartmentalization in regions with BMI-interacting variants affecting inflammation

BMI-discordant monozygotic twin preadipocytes exhibit altered compartmentalization in regions with BMI-interacting variants affecting inflammation

Kristina M. Garske[1], Caroline Comenho[1], Brunilda Baillu[2], David Z. Pan[1,3], Yash V. Bhagat[1], Gregory Rosenberg[1], Amogha Koka[1], Zong Miao[1,3], Janet S. Sinsheimer[1,2,3], Kirsi H. Pietiläinen[4,5], Päivi Pajukanta[1,2,6,*]

[1]Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, CA 90095, USA

[2]Department of Computational Medicine, UCLA, Los Angeles, CA 90095, USA

[3]Bioinformatics Interdepartmental Program, UCLA, Los Angeles, CA 90095, USA

[4]Obesity Research Unit, Research Program for Clinical and Molecular Metabolism, Faculty of Medicine, University of Helsinki, Helsinki 00014, Finland

[5]Obesity Center, Abdominal Center, Helsinki University Hospital and University of Helsinki, Helsinki 00014, Finland

[6]Institute for Precision Heath, David Geffen School of Medicine at UCLA, Los Angeles, CA 90095, USA

*Correspondence: ppajukanta@mednet.ucla.edu

Summary

Obesity is associated with adipose tissue dysfunction, which can cause low-grade inflammation. In obesity, preadipocytes (PAd) need to efficiently differentiate to adipocytes to store the extra fat, making PAd one of the key cell types affected by obesity and obesity-induced inflammation. To test the hypothesis that increased body mass index (BMI) alters genomic programming of human preadipocytes and thus drives gene-environment interactions underlying inflammation, we leveraged a BMI-discordant monozygotic (MZ) twin cohort. We first defined the higher-order genomic programming across the twin pairs by integrating open chromatin (ATAC-seq) and expression (RNA-seq) data from primary preadipocytes isolated from the twins' adipose tissue with chromosomal interaction (promoter Capture-HiC) data. After identifying subnuclear active (A) and inactive (B) compartments, we observed that connectivity of open chromatin in the active A compartments is significantly impaired in the higher BMI MZ twins compared to their paired lower BMI MZ twin siblings. Extending these epigenetic results to the UK Biobank (UKB) showed that the A compartment regions that differed most (~88.5 Mb) in open chromatin connectivity between the twins also significantly contribute to systemic inflammation in UKB by variants interacting with BMI on C-reactive protein (CRP). Taken together, the obesogenic cellular microenvironment alters genomic programming in human preadipocytes, which drives inflammation through gene-environment interactions.

Introduction

The obesity epidemic is driving concomitant alarming increases in obesity comorbidities, such as type 2 diabetes (T2D), coronary artery disease (CAD), and non-alcoholic fatty liver disease (NAFLD) (Hotamisligil, 2006; Lavie *et al.*, 2018; Cercato and Fonseca, 2019). Obesity is also one of the key risk factors for severe COVID-19 outcomes (Docherty *et al.*, 2020; Petrilli *et al.*, 2020), most likely not only due to the altered mechanics of lung ventilation but also due to the low-grade inflammation induced by obesity (Ritter *et al.*, 2020). Recent assessment of polygenic risk scores (PRSs) for the obesity surrogate trait, body mass index (BMI), highlights how accumulation of risk variants is associated with the level of BMI and higher odds of having obesity comorbidities, T2D or CAD, in the UK Biobank (UKB) (Khera *et al.*, 2019). However, since not all individuals with obesity exhibit metabolic profiles associated with poor cardiometabolic health outcomes, the genetic risk for BMI and related complex traits is not deterministic, and environmental and lifestyle factors also confer the risk (Khera *et al.*, 2019; Blüher, 2020). Therefore, understanding the pathophysiological mechanisms in obesity that contribute to metabolically unhealthy phenotypes can improve risk assessment of genetic and environmental contributions to these clinically important traits. This will ultimately inform treatment strategies to decrease morbidity and mortality due to the obesity epidemic.

Chronic low-grade inflammation is a hallmark of obesity that contributes to the development of obesity comorbidities such as insulin resistance and atherosclerosis (Hotamisligil, 2006). In obesity, the adipose tissue environment develops into a pro-inflammatory state that corresponds with an increase in the overall number, as well as activation, of resident macrophages (Hotamisligil, 2006; Ouchi *et al.*, 2011). However, the factors that initiate these events in obesity

are incompletely understood. There is increasing evidence for the role of preadipocytes in the development of dysfunctional adipose tissue and the pro-inflammatory state seen in obesity. It has been shown that primary preadipocytes isolated from individuals with obesity are less efficient at producing adipocytes *in vitro* when compared with preadipocytes isolated from lean individuals (Isakson *et al.*, 2009). This reduced adipogenic capacity can contribute to adipocyte hypertrophy, wherein already existing adipocytes need to take up and store excess energy in the absence of newly formed adipocytes. Hypertrophic adipocytes produce abnormal levels of adipokines, which contributes to insulin resistance and inflammation (Sorisky, Molgat and Gagnon, 2013; Ghaben and Scherer, 2019; Pyrina *et al.*, 2020). This suggests that one mechanism linking adipose tissue inflammation and downstream obesity comorbidities is reduced preadipocyte differentiation.

Chronic signaling conditions, such as in the state of overnutrition that causes obesity, can elicit changes at the epigenomic level that contribute to dysregulated preadipocytes. For example, preadipocytes that have been cultured with high levels of glucose show altered histone modifications at genes involved in inflammatory responses (Rønningen *et al.*, 2015). This suggests that prolonged exposure to this obesogenic environment affects the epigenomic priming of a pro-inflammatory expression profile in preadipocytes. However, it is unclear if similar mechanisms contribute to the inflammatory response to obesity *in vivo*. Evidence from cell biological (Isakson *et al.*, 2009) and epigenomic (Andersen *et al.*, 2019) studies have shown that preadipocytes from obese individuals exhibit altered differentiation potential relative to preadipocytes from lean individuals, suggesting that the cellular genomic programming can be

affected by obesity *in vivo*. The mechanisms involved in the altered programming and the downstream effects on inflammation remain to be determined.

Here, we leveraged a cohort of monozygotic (MZ) twins who are discordant for BMI (ΔBMI>3 kg/m^2) to study the effects of increased BMI on preadipocyte cellular programming. Because their genetics are identical, epigenomic differences between the MZ twin siblings can be attributed to differences in lifestyle between the twins. We characterized preadipocyte genomic programming through Assay for Transposase-Accessible Chromatin (ATAC) -sequencing and RNA-seq in the MZ twins, and promoter Capture Hi-C (pCHi-C) in an independent source of human primary preadipocytes. By integrating these data and leveraging co-accessibility information across all preadipocyte samples from the twins, we identified subnuclear compartments of chromatin activity, previously defined as active (A) and inactive (B) compartments (Perino and Veenstra, 2016). The active A compartments exhibited significant differences in compartmentalization of chromatin activity, defined by the level of connectivity, between the lower and higher BMI MZ sibling groups. We show that these regions of altered genomic compartmentalization of chromatin activity also contribute to inflammation through gene-BMI interactions in the UKB.

Results

Identification of A/B compartments in human primary preadipocytes

We have an ongoing collection of a unique, deeply phenotyped cohort of 65 Finnish monozygotic (MZ) twin pairs who are discordant for BMI (ΔBMI$\geq$3 kg/m^2) (Granér *et al.*, 2012; Naukkarinen *et al.*, 2014) (see Methods). The clinical and metabolic characteristics of these twins are summarized in Table S1. A paired *t*-test of clinical measurements indicates that many traits are significantly different between the lower and higher BMI groups of siblings, including C-reactive protein (CRP), which is a measure of systemic inflammation, an important comorbidity of obesity (Table S1). We hypothesized that preadipocyte (PAd) genomic programming is altered under conditions of increased BMI, which could lead to adipose tissue dysregulation and contribute to the difference in inflammatory profiles. We therefore isolated the primary subcutaneous PAd from the subcutaneous adipose biopsies of 10 pairs (n = 20) of the BMI-discordant MZ twins (see Methods). The clinical characteristics of this subset of twins for whom we collected the PAd are summarized separately in Table S1.

Cellular genomic programming is defined by various levels of genome organization. To first characterize the PAd genomic programming at the level of subnuclear compartmentalization of chromatin activity, we performed ATAC-seq on the PAd from all 10 pairs of BMI-discordant MZ twins (see Methods). We inferred A and B genomic compartments, which are broadly associated with active or inactive regions of genome (Perino and Veenstra, 2016), respectively, using the co-accessibility information from the ATAC-seq coverage across 100-kb bins, as described previously (Fortin and Hansen, 2015) (see Methods) (Figure 1A; Table S2). In the MZ twin PAd, the A compartments were shorter than B compartments, with a median of 300 kb

compared with 600 kb in the B compartments ($p_{Wilcoxon}$=3.1x10^{-76}) (Figure S1A). B compartments made up an average of 74.8%% +/- 9.7% (s.d.) of each chromosome (Figure S1B). We validated the compartment detection by characterizing the stratification of known functional features of chromatin compartmentalization across the A and B compartments (Lieberman-Aiden *et al.*, 2009; van Steensel and Belmont, 2017). The proportion of chromosomes making up B compartments is significantly correlated (Spearman's ρ=0.68, p=6.8x10^{-4}) with the percent of the chromosome that is comprised of gene deserts (Figure S1C). As previously reported (Lieberman-Aiden *et al.*, 2009; van Steensel and Belmont, 2017), gene deserts were largely restricted to B compartments (Figure S1D) and the gene density in the A compartments was significantly higher than in the B compartments ($p_{Wilcoxon}$=3.34x10^{-69}) (Figure S1E).

Preadipocyte A/B compartments reflect cell-type-specific gene regulatory activity

To address the cell-type-specificity of gene regulatory activity in the PAd A/B compartments, we compared their coverage of chromatin states across 127 ENCODE cell types using the imputed 25 state model from ChromHMM (Ernst and Kellis, 2015). We found that the A compartment coverage of enhancer chromatin states was highest for the mesenchymal stem cell -derived adipocyte cultured cells (MSC-Ad) ($FDR_{Wilcoxon}$<0.05), and next highest in adipose-derived mesenchymal stem cell cultured cells (Ad-MSC), in line with the primary PAd being at a developmental stage similar to these two cell types (Figure S2). There was a similar trend for A compartment coverage of promoter chromatin states for MSC-Ad (Figure S2). Conversely, the A compartment coverage of quiescent chromatin states was significantly lower in the MSC-Ad than all other cell types ($FDR_{Wilcoxon}$<0.05) except Ad-MSCs and primary breast myoepithelial cells

(Figure S2). The coverage of the active transcription chromatin state in the A compartments was not highest in the MSC-Ad (data not shown), suggesting that the promoter, enhancer, and quiescent chromatin states exhibit more cell-type-specific compartment coverage. Since the A compartment coverage was most specific for MSC-Ad, we consider the primary PAd to be most similar to this ENCODE cell type. We therefore used these chromatin states for all subsequent analyses using ChromHMM.

To confirm that measures of active gene regulation are more restricted to A compartments than B compartments, we assessed the coverage of active or inactive chromatin states within the A/B compartments. The A compartment coverage was higher than B compartment coverage for the enhancer ($p_{\text{Wilcoxon}}=2.83\text{x}10^{-79}$) and promoter ($p_{\text{Wilcoxon}}=8.84\text{x}10^{-149}$) chromatin states, whereas the B compartment coverage was higher than the A compartment coverage for quiescent states ($p_{\text{Wilcoxon}}=1.23\text{x}10^{-100}$) (Figure 1B), in line with active gene regulation being more prevalent in the A compartments. Taken together, the A/B compartments inferred from the human primary PAd ATAC-seq data exhibit cell-type-specific gene regulatory chromatin states and functional genomics measures from human MSC-Ad, consistent with the A compartment being comprised of more active gene regulatory states that may be important for regulating PAd cellular programming.

To determine whether the subnuclear compartmentalization of chromatin activity associates with gene expression in the way expected based on the epigenetic data, we performed RNA-seq on the 20 MZ twin PAd samples (see Methods). Genes in the A compartments have higher mean expression than genes in the B compartments ($p_{\text{Wilcoxon}}=4.63\text{x}10^{-49}$), in line with the fact that A

compartments are enriched for molecular signatures of active gene regulation (Figure 1C). A clear example of how the A/B compartments define cell-type-specific genomic programming is presented in Figure 1D and 1E. Consistent with these primary cells being at a later developmental time point (Gulyaeva *et al.*, 2018), the early PAd marker, *DLK1*, is located within a B compartment and has negligible expression in the preadipocytes (Figure 1D). Conversely, the later PAd marker, *PDGFRA*, is located within an A compartment and is clearly expressed (Figure 1E). Furthermore, the well-established adipocyte-specific adipocytokine, *ADIPOQ*, is located within a PAd B compartment, in line with this gene not being expressed until the later stages of adipocyte differentiation (Table S2). Taken together, our data on A and B compartments are in accordance with previously published hallmark features of subnuclear compartments.

Promoter-enhancer interactions are enriched in the A compartments

Cellular genomic programming involves the proper targeting of active or repressive regulatory elements and the necessary DNA-binding proteins to their target promoters through chromosomal interactions (Perino and Veenstra, 2016). We therefore identified promoter interactions in human primary PAd using promoter Capture Hi-C (pCHi-C) (see Methods). To assess whether the PAd pCHi-C interactions correspond to the subnuclear A/B compartmentalization, we examined whether the two ends of the pCHi-C interactions land in the same or different compartments. For 51,974 of the 76,473 PAd interactions (68.0%), both ends landed in the same compartment, with 25,686 of these (49.4%) being contained within the same A compartment, and 26,288 (50.6%) of these interactions being in the same B compartment. Given that the B compartments make up ~75% of the genome (Figure S1B), this suggests that

the pCHi-C interactions are enriched within the A compartments. To determine whether the proportion of pCHi-C interactions in the A compartments is higher than expected by chance alone, we permuted the compartment locations and re-calculated how often both ends of the pCHi-C interactions land in the same A compartment (see Methods). The proportion of interactions that have both ends landing in the same permuted A compartments is on average 18.5% +/- 1.1% (s.d.), meaning that there is a 2.67-fold enrichment of pCHi-C interactions in the A compartments ($p<1x10^{-04}$). This is in line with the pCHi-C interactions being regulatory and thus being more prevalent in A compartment regions of active gene regulation. It has previously been reported that genes involved in pCHi-C interactions are more highly expressed (Schoenfelder *et al.*, 2015). We found that this is only true for genes located within the A compartments (n=8,807 genes, $p_{wilcoxon}=4.2x10^{-08}$), whereas this was not the case for genes in the B compartments (n=9,903) ($p_{wilcoxon}$=n.s.) (Figure S3).

Taken together, we have shown that PAd pCHi-C interactions are enriched in the PAd A compartments, and that the previously reported higher expression of genes involved in pCHi-C interactions is dependent upon the gene landing in A, rather than B compartments (Figure S3). We chose to focus on the A compartments for the remainder of our study, given the evidence that active gene regulation is occurring in these regions, and thus they are likely important for PAd programming and function.

Genome-wide A compartment connectivity is decreased in MZ twin siblings with a higher BMI

Given their identical genomes, in theory the differences in BMI between the BMI-discordant MZ twins are expected to be entirely due to environmental and epigenetic factors. We hypothesized that alterations in cellular signaling resulting from long-term differences in nutrient intake or expenditure between the lower and higher BMI MZ siblings would affect the compartmentalization of the chromatin activity in their preadipocytes. To determine whether we could quantify an overall measure of chromatin activity in the A compartments, we calculated the A compartment connectivity at the genome-wide level (see Methods), following the hypothesis that the compartments that are more highly connected would exhibit coordinated gene regulation and higher accumulation of active regulatory states. We found that the A compartment connectivity is significantly associated with gene expression (p_{KW}=1.27x10^{-16}) (Figure 2A), and all measures of active gene regulation based on chromatin state, PAd pCHi-C interactions, and PAd accessible chromatin (Figure 2B). Taken together, the A compartment connectivity can be used as a metric that captures various degrees of genomic regulation associated with genomic programming.

To assess whether this measure of genomic programming is different between the BMI-discordant MZ twin siblings, we separated the lower and higher BMI twins (n=9 MZ sibling pairs split to into lower and higher BMI subgroups) and re-computed the connectivity of the A compartments in each set separately (see Methods). We found that the genome-wide compartment connectivity is significantly higher in the lower BMI set of twins than in the higher BMI twins ($p_{\text{paired Wilcoxon}}$=4.96x10^{-31}) (Figure 2C). This overall decline of A compartment connectivity in the higher BMI siblings shows that PAd programming is altered when comparing individuals with higher BMI to those with a lower BMI across the same genome.

To identify specific A compartments that exhibit the strongest connectivity differences between the lower and higher BMI MZ twin siblings, we permuted the lower and higher BMI labels between siblings ($n=9^2$ or 512 permutations) and re-calculated the difference in connectivity between the lower and higher BMI subgroups of siblings for all compartments (see Methods). We selected compartments with a permutation *p*-value of less than 0.01, totaling 88.5 Mb, to define the reprogrammed A compartments (Table S2). These A compartments that have been reprogrammed in response to BMI in the BMI-discordant MZ twins thus represent regions that are mechanistic candidates for gene-environment interactions (GxEs) originating in human PAd.

Reprogrammed PAd A compartments are enriched for the heritability of CRP and contribute to gene-BMI interaction effects on CRP in the UK Biobank

To first test whether there is evidence for PAd genomic contribution to the marginal genetic effects on systemic inflammation, we used partitioned LDSC regression (Bulik-Sullivan *et al.*, 2015; Finucane *et al.*, 2015) to partition the heritability of CRP into the B compartments and the A compartments, while also stratifying the latter ones to the A compartments that were reprogrammed in the BMI-discordant MZ twins and the A compartments not reprogrammed in the twins (see Methods). The B compartments were significantly depleted for the heritability of CRP (enrichment=0.840, $p=1.99\text{x}10^{-09}$), whereas the A compartments (enrichment=1.33, $p=7.05\text{x}10^{-04}$) and the reprogrammed A compartments (enrichment=3.10, $p=6.19\text{x}10^{-03}$) were both significantly enriched for the heritability of CRP (Table 1).

To next determine whether the BMI-responsive reprogrammed A compartments identified in the BMI-discordant MZ twin PAd are more likely to harbor GxE effects on inflammation, we performed a GxE scan in the UKB, testing all SNPs in the A compartments for the effect of SNPs interacting with BMI on CRP levels (see Methods). We compared the distribution of the GxE *p*-values in the reprogrammed A compartments (88.5 Mb) to all other A compartments (561 Mb) (Table S2). Indeed, we found that the reprogrammed compartments have a higher accumulation of low *p*-value GxE signals than all other A compartments ($p_{\text{Wilcoxon}}=7.54\text{x}10^{-05}$) (Figure 2D). This shows that the regions exhibiting BMI-dependent PAd programming differences identified in the MZ twins harbor many small-effect GxEs affecting inflammation in humans, above what is seen in the A compartments alone.

Clustering of the A compartments identifies subcommunities of compartments important for distinct preadipocyte functions

To gain insight into the potential genomic regulatory mechanisms that contribute to the regionally enriched heritability of CRP and the observed GxEs affecting inflammation in the UKB, we clustered the A compartments after UMAP dimensionality reduction (McInnes *et al.*, 2018) to 2 variance components (see Methods). This clustering approach identified 10 clusters that exhibit varying levels of connectivity, containing between 107 and 230 A compartments (Figure 3A; Figure S4A; Table S2; Table S3). The total sum of the lengths (in Mbs) of the compartments in each of the A compartment clusters is given in Table S3. As expected, given the observed correlations between connectivity and gene regulatory signatures (Figure 2B), the clusters containing the A compartments with the highest levels of connectivity (clusters 1, 2, 3, and 5; $p_{adj}<0.05$) (Figure 3A; Figure S4A) also exhibit the highest coverage of enhancer (Figure

3B) (clusters 2, 3 and 5; $p_{adj}<0.05$) or promoter (Figure 3C) (clusters 1 and 2; $p_{adj}<0.05$) chromatin states; and have the lowest coverage of quiescent chromatin states ($p_{adj}<0.05$) (Figure 3D). Thus, as clusters 1, 2, 3, and 5 likely represent the most important A compartments for PAd programming and function, we chose to focus on these clusters for the remainder of the study.

The differences in enhancer and promoter chromatin state coverage between the four main A compartment clusters 1, 2, 3 and 5 suggest that there may be differential gene regulatory mechanisms functioning within each cluster. In support of this, we found that cluster 5, which exhibits the highest enhancer chromatin state coverage (Figure 3B), is enriched 3.2-fold for PAd super-enhancers ($p_{adj\ hypergeom}=7.8x10^{-04}$) (see Methods). Cluster 5 also exhibits the highest number of PAd pCHi-C interactions per promoter ($p_{KW}=8.95x10^{-46}$), particularly when compared with cluster 1 (8 interactions per promoter in cluster 5 versus 3 interactions per promoter in cluster 1) (Figure 3E). Accordingly, the gene expression in cluster 1 is significantly lower than in clusters 2, 3, and 5 ($p_{KW}=2.50x10^{-08}$) (Figure S4B). These data suggest that, in contrast to the highly interacting, enhancer-enriched cluster 5, cluster 1 may be more developmentally primed. One feature that has been previously reported to be more common in cells that are primed for differentiation is a higher number of promoter-promoter (P-P) interactions in pCHi-C data (Joshi *et al.*, 2015). We found that cluster 1 does in fact exhibit a higher proportion of P-P interactions relative to clusters 2, 3, and 5 (Figure S4C). We performed a gene ontology (GO) enrichment analysis on the genes in each of the clusters separately (Figure S5; Table S4) (see Methods), and found that cluster 1 is enriched for developmental processes, cell polarity, and cell adhesion, in line with this cluster being important for cellular priming (Noronha Nc *et al.*, 2019) (Figure S5; Table S4). Notably, immune-related processes such as leukocyte chemotaxis and proliferation,

response to cytokine, and apoptotic cell clearance are also enriched in the A compartment cluster 1 (Figure S5; Table S4).

To further examine whether the regions in the A compartment cluster 1 are likely to be developmentally primed relative to the A compartment cluster 5, we assessed the effects of initiating the MZ twin PAd differentiation into adipocytes for 24 hours by performing ATAC-seq on the cells at this developmental time point (see Methods). Intriguingly, the A compartment cluster 5, which has the highest accumulation of enhancer chromatin state coverage, exhibited a higher proportion of ATAC-seq peaks with decreased accessibility after the first 24 hours of differentiation (Figure 3F). This is consistent with cluster 5 being made up of genomic regions that are specifically important for PAd function, in line with the strong enrichment of super-enhancers (see above). On the other hand, cluster 1 showed the opposite trend, with a higher proportion of ATAC-seq peaks being more accessible after the first 24 hours of differentiation, relative to PAd (Figure 3F). Taken together, the chromatin state and enhancer coverage, as well as the differential responses to early differentiation signals, suggest that the A compartment clusters represent regions of the genome that are functionally related and exhibit distinct gene regulatory mechanisms in PAd.

The primed PAd A compartment cluster 1 contributes significantly to the heritability of CRP and is enriched for gene-by-BMI interaction effects on CRP in the UK Biobank

To determine whether any of the four main A compartment clusters are particularly important in participating in the PAd responses to BMI in the MZ twins, we tested whether the reprogrammed A compartments identified in the MZ twins (Figure 2C,D; Table S2) are overrepresented in any

of the clusters. We found that both clusters 1 (2.65-fold enrichment, p_{adj}=2.62x10^{-07}) and 2 (3.47-fold enrichment, p_{adj}=2.81x10^{-13}) are significantly enriched for the reprogrammed A compartments (Table S3). We also found, using partitioned LDSC, that the A compartment clusters 1 and 2 are significantly enriched for the heritability of CRP in the UKB (Figure 4A). Strikingly, when we compare the GxE SNP *p*-values for SNPs interacting with BMI to affect CRP levels in the UKB, we also found that cluster 1, with the highest promoter coverage, has a higher accumulation of low *p*-value GxE SNPs when compared to cluster 5, with the highest enhancer coverage (p_{KW}=0.0164) (Figure 4B). This supports the conclusion that primed, rather than highly regulated and cell-type-specific regions, are important for the immunomodulatory effects of PAd responses to BMI. In summary, the cluster-dependent responses to BMI and contribution to CRP heritability, including to GxEs affecting CRP in the UKB, all support a role for the A compartment cluster 1 regions being important for PAd responses to BMI and affecting systemic inflammation.

To identify genes that may modulate inflammation through PAd mechanisms in the A compartment cluster 1 regions, we referred back to the CRP GWAS loci in the cluster 1 compartments that were found to be reprogrammed in the MZ twins (Figure 2C,D; Table S2). Of the 24 compartments in the A compartment cluster 1 that contain at least one CRP GWAS SNP, seven (29.2%) were reprogrammed in the BMI-discordant MZ twin PAd. These seven compartments contain a total of 17 independent CRP GWAS signals (Table S5). One of these harbors two independent CRP GWAS signals, and is located within the *HLA* locus on chromosome 6, in line with BMI driving PAd cellular responses at this locus, which impacts CRP levels in humans (Table S5). Notably, another reprogrammed A compartment on

chromosome 3 contains five independent CRP GWAS signals (Table S5) and 12 known immune-related genes, which suggests that this reprogrammed A compartment is an immune hub responding to the cellular obesogenic microenvironment. One of the CRP GWAS signals, with the tag SNP rs2271961 (Table S5), is located within a 156-kb region of this reprogrammed A compartment. The SNP LD proxies land in PAd pCHi-C interactions that contain PAd ATAC-seq peaks, and interact with or are in the interacting promoter of three immune-related genes: Inositol Hexakisphosphate Kinase 1 (*IP6K1*), TRAF Interacting Protein (*TRAIP*), and Macrophage Stimulating 1 Receptor (*MST1R*) (Table S5). These genes thus represent strong functional candidates for PAd-origin effects of BMI-induced inflammation. Taken together, the A compartment cluster 1 is enriched for the heritability of CRP and for PAd responses to BMI at the genomic compartmentalization level, likely leading to the enrichment of small-effect SNPs interacting with BMI to affect CRP in humans.

Discussion

Obesity predisposes to COVID-19 complications and a cascade of cardiometabolic disorders (CMDs), likely at least partially by inducing chronic low-grade inflammation in the affected tissues (Hotamisligil, 2006; Kivimäki *et al.*, 2017; Docherty *et al.*, 2020; Petrilli *et al.*, 2020; Ritter *et al.*, 2020). Preadipocytes (PAd) are one of the key cell types in adipose tissue, responding to environmental cues and deciding whether to proceed toward fat storage (differentiation into adipocytes) or alternative pathways. In this study, we showed that increased BMI affects the higher-order compartmentalization of the genome in PAd, in regions that contribute to the heritability of inflammation, measured by CRP in the UK Biobank (UKB). Furthermore, these reprogrammed regions in PAd that span ~88.5 Mb exhibit a higher accumulation of small-effect GxBMI SNPs affecting CRP. Taken together, BMI affects PAd programming in large genomic regions that contribute to systemic inflammation in humans, suggesting an important role for this progenitor cell type in the low-grade inflammatory state that is associated with obesity (Reilly and Saltiel, 2017).

When PAd differentiation is impaired, as occurs in obesity (Isakson *et al.*, 2009; Andersen *et al.*, 2019), already existing adipocytes take up and store excess fat, causing adipocyte hypertrophy. Due to the association between adipocyte hypertrophy and adipose tissue dysregulation, PAd are an important factor to consider when understanding systemic inflammation and downstream obesity comorbidities. However, PAd are an understudied cell type, underrepresented in large collections of epigenetic data such as ENCODE (Davis *et al.*, 2018) and Roadmap (Roadmap Epigenomics Consortium *et al.*, 2015), and mainly studied in mouse or human PAd-like cells, rather than primary cells (MacDougald and Mandrup, 2002; Rauch *et al.*, 2019). This is likely in

part due to the relative difficulty in collecting and propagating primary PAd cells in sufficient numbers for genomics studies. Because of this, there is an overall lack of knowledge surrounding the genomic programming of primary human PAd. Our study identified and characterized primary PAd A/B compartments, which represent the higher-order genomic compartmentalization of chromatin activity across 10 pairs (n=20) of BMI-discordant MZ twins. Thus, this work advances the field by elucidating the genomic context in which local epigenetic signatures function, toward understanding how these cells integrate information from the environment to make important cellular decisions. We showed that the active PAd A compartments are enriched for enhancer and promoter chromatin states, relative to the more inactive B compartments, which are enriched for quiescent chromatin states. Furthermore, subclustering of the A compartments groups genomic regions that contain genes that are important for distinct progenitor cell functions, such as development (cluster 1), signal transduction (cluster 2), and hormone secretion (cluster 5). These clusters exhibit differences in their gene regulatory landscapes, assessed through chromatin state coverage, super-enhancer identification, regulatory interactions, and responses to differentiation signals. Thus, we have shown that higher-order genomic coordination is important for defining functionally related regions of the genome in PAd. By quantifying this coordination through the A compartment connectivity, we then were able to show that the PAd genomic programming at this level is impaired in the higher BMI subgroup of the BMI-discordant MZ twins.

We showed that the reprogrammed A compartments contribute significantly to the heritability of CRP, meaning that environment-responsive regions of the genome also contain an enrichment of marginal SNP effects on inflammation. This suggests an important contribution of PAd BMI-

responsive regions to systemic inflammation. GxEs are difficult to detect in humans for various reasons, including environmental heterogeneity, imprecision in the environmental measurements, and low power to detect interaction effects in the current cohort sizes, even in the UKB. We have previously shown that by restricting the GxE search space to regions of the genome that contain SNPs that respond to environmental cues in experiments, we can reduce the multiple testing burden to only include those regions with prior evidence of being relevant for that environmental stimulus (Garske *et al.*, 2019). This enables the detection of significant GxEs, even when genome-wide significant signals remain difficult to detect. Here, rather than a controlled experiment, we are using BMI as the environmental variable, which in itself is very heterogeneous in its etiology. However, by leveraging the BMI-discordant MZ twins, we identified regions that differed between MZ siblings, thus controlling for the heterogeneity of genetic backgrounds, which complicates and reduces power in environmental studies from cohorts of unrelated individuals.

We found that the reprogrammed genomic regions, marginal effects on CRP, and GxE signals affecting CRP were enriched in regions of the genome that exhibit features of a more primed cellular state (cluster 1), rather than in the regions enriched for super-enhancers and highly interacting gene promoters (cluster 5). Cluster 1 exhibits a higher accumulation of promoter chromatin states, which possibly represent poised promoters. This concept of poised promoters is supported by the lower number of interactions per promoter, the lower gene expression, and the previous knowledge that promoter-promoter interactions are associated with a primed cellular state (Joshi *et al.*, 2015). However, it is important to consider that this conclusion may be driven by the fact that we only assessed the effects of BMI on the higher-order level of genome

organization. The complex etiology of increased BMI reduces the power to detect BMI-driven differences in epigenetic signatures, particularly with the small sample size of the current study. This precluded the assessment of the effects of BMI on individual open chromatin peaks, due to the heavy multiple-testing burden at the genome-wide level (testing for BMI-driven differences in tens to hundreds of thousands of individual peaks). Therefore, it is possible that enhancer-enriched regions in PAd do respond to BMI, but the higher-order coordination of those regions, as we assessed through PAd connectivity, is not as strongly affected. Furthermore, different traits may be affected by different BMI-responsive mechanisms. Thus, larger cohort sizes and investigation of alternative genomic regulatory mechanisms aside from the higher-order coordination of active regions is warranted to further understand the effects of increased BMI on human primary PAd.

In conclusion, we have characterized the higher-order genomic programming of human primary PAd and refined active genomic regions to functionally related clusters that span 30-130 Mbs, thus providing new important information for adipose biology and obesity research. Increased BMI affects this level of PAd genomic programming through dysregulation of the coordination of functionally related regions of the genome. These reprogrammed regions are significantly enriched for the heritability of CRP, and harbor a higher accumulation of small-effect SNPs interacting with BMI to affect CRP levels. Taken together, we have identified PAd-origin genomic regulatory mechanisms that respond to BMI to induce the key obesity consequence, inflammation.

Acknowledgements

This research has been conducted using the UK Biobank Resource under application number 33934. We thank the individuals who participated in the UK Biobank study. We thank the UNGC sequencing core at UCLA for performing the DNA and RNA sequencing. This study was funded by the National Institutes of Health (NIH) grant U01DK105561; the Academy of Finland, grant numbers 266286, 272376, 314383, 335443; Finnish Medical Foundation; Gyllenberg Foundation; Novo Nordisk Foundation, grant numbers NNF20OC0060547, NNF17OC0027232, NNF10OC1013354; Finnish Diabetes Research Foundation; University of Helsinki and Helsinki University Hospital; Government Research Funds. Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number R25GM055052. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. K.M.G. was supported by the NIH-NHLBI grant F31HL142180. D.Z.P. was supported by the NIH-NIDDK grant F31DK118865.

Author contributions

K.M.G. and P.P. designed the study. K.M.G., K.P., and P.P. generated the ATAC-seq and RNA-seq data. K.M.G., B.B., D.Z.P., Z.M., J.S.S. and P.P. developed the analytical and statistical approaches. K.M.G., A.K. and P.P. performed the comparisons of the results against genomic databases and literature. K.M.G., B.B. and D.Z.P. performed computational analysis of the data. K.M.G., C.C., Y.V.B. and G.R. performed the experiments. K.M.G. and P.P. wrote the manuscript and all authors read, reviewed, and/or edited the manuscript.

Declaration of interests

The authors declare no competing interests.

Materials availability

This study did not generate new unique reagents.

Data and code availability

The data that support the GxE findings in this manuscript are available from the UK Biobank. However, restrictions apply to the availability of these data, which were used in this study under UK Biobank Application Number 33934. UK Biobank data are available for bona fide researchers through the application process. The twin ATAC-seq and RNA-seq data are deposited to the THL Biobank, Helsinki, Finland (https://thl.fi/en/web/thl-biobank/for-researchers/sample-collections/twin-study). These twin data are available for bona fide researchers through the application process (https://thl.fi/en/web/thl-biobank/for-researchers/application-process). The PAd pCHi-C data are available at GEO under the accession number GSEXXXXX.

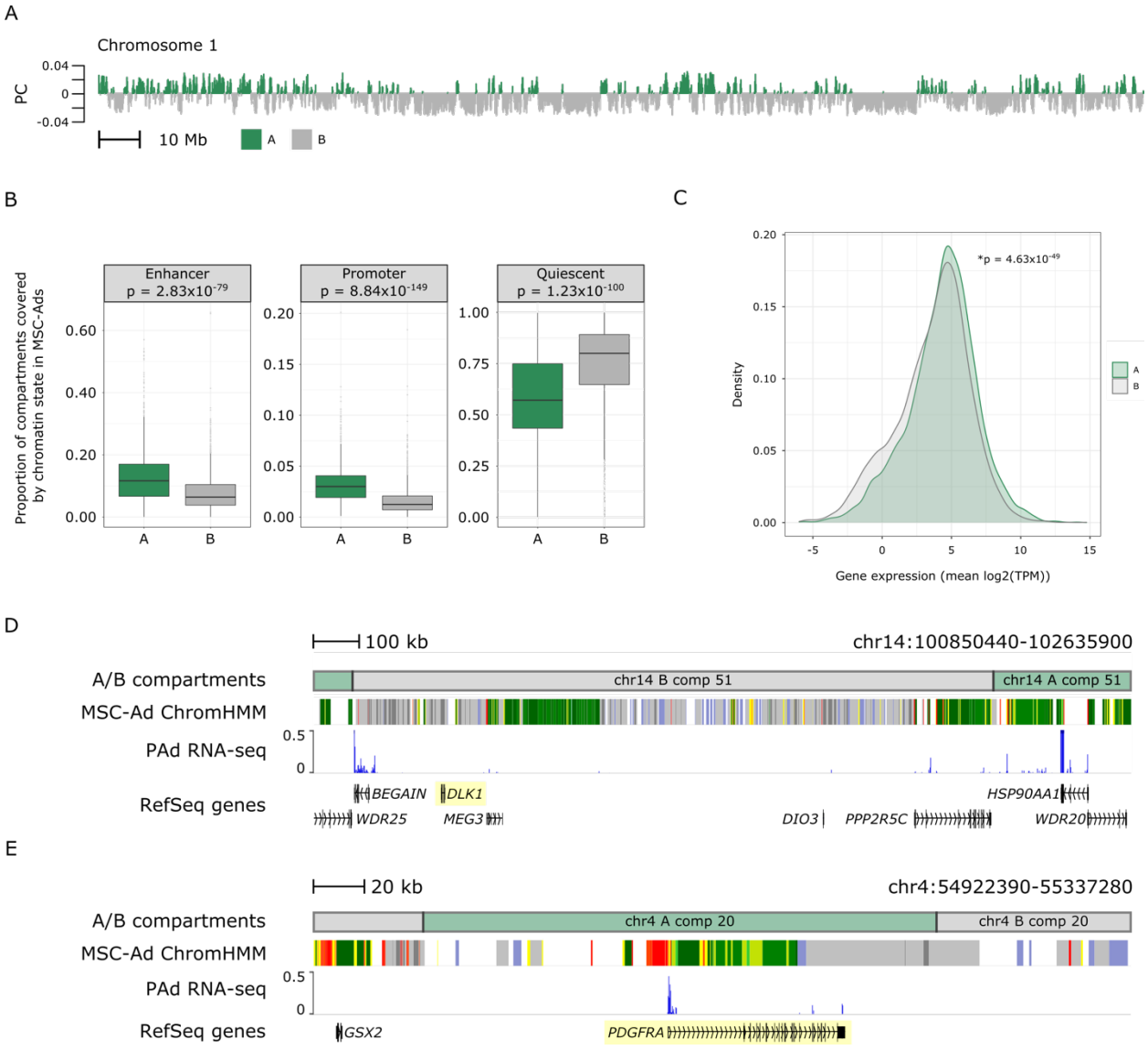


Figure 1. A/B compartment identification using ATAC-seq co-accessibility in human primary preadipocytes. (A) PAd A/B compartments on chromosome 1. Each bar represents a 100-kb bin and the height of the bar is the projection onto the first eigenvector of the 100-kb bin co-accessibility matrix across chromosome 1. The sign switches at A/B compartment boundaries; positive values correspond to A compartments (green) and negative values correspond to B compartments (grey). (B) Coverage of enhancer, promoter, and quiescent chromHMM chromatin

states in the A and B compartments. *P*-values correspond to the Wilcoxon Rank Sum test comparing the A compartment to B compartment coverage for each chromatin state. (C) Density distribution of the gene expression (mean $\log_2$(TPM)) in the A (green) and B (grey) compartments shows higher expression in A compartments. *P*-value corresponds to the Wilcoxon Rank Sum test comparing the gene expression in A compartments to the gene expression in B compartments. Genome browser snapshots of two preadipocyte marker genes: (D) *DLK1*, an early preadipocyte marker, is located within a B compartment on chromosome 4 and is not expressed; and (E) *PDGFRA*, a late preadipocyte marker, is located within an A compartment on chromosome 14 and is expressed. The ChromHMM state track is directly from Roadmap Epigenomics on the WashU Epigenome Browser. PC indicates principal component; MSC-Ad mesenchymal stem cell derived adipocyte cultured cells; and PAd, preadipocyte. See also Figures S1-3 and Table S2.

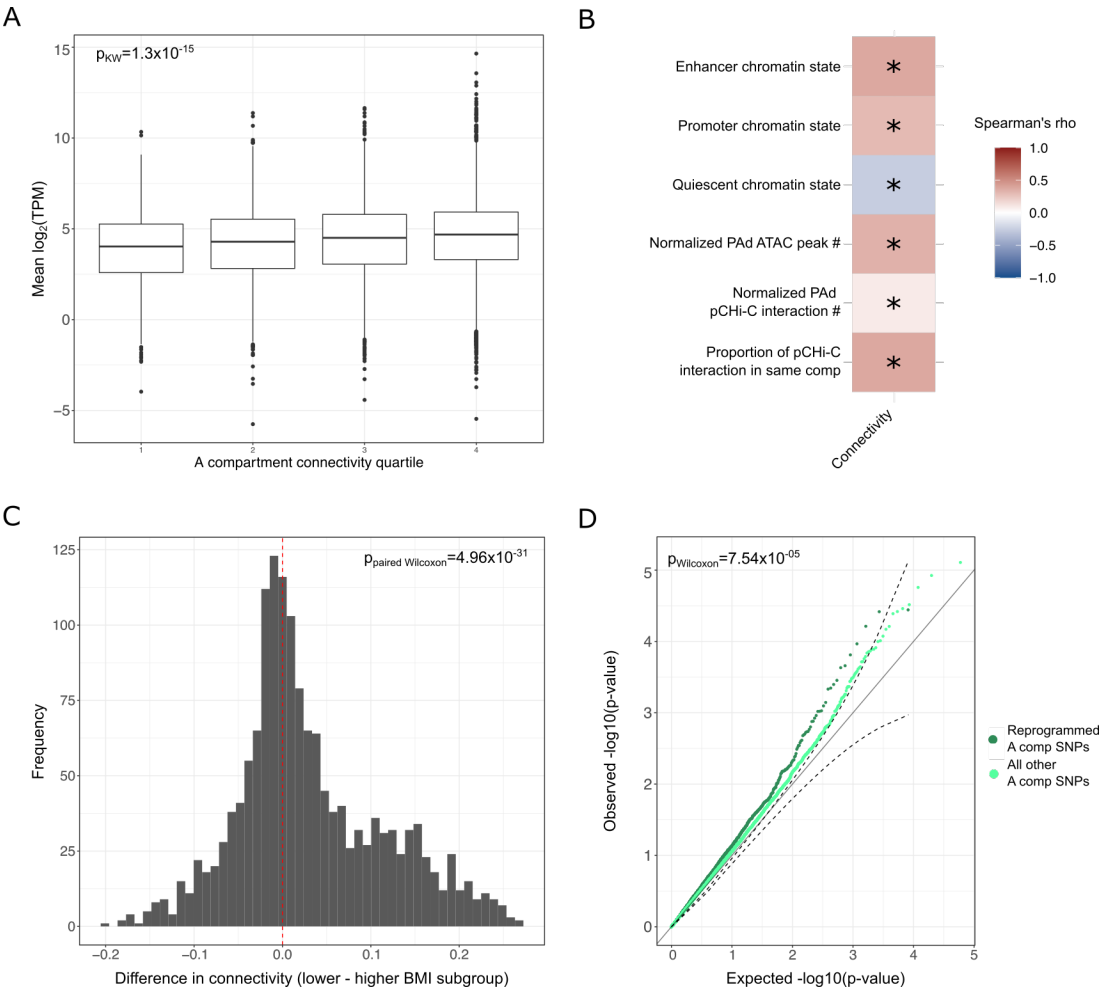


Figure 2. The A compartment connectivity differs between the lower and higher BMI twins and contributes to gene-BMI interactions affecting CRP in the UK Biobank. (A) Boxplots show the mean expression (log_2(TPM)) of genes in the A compartments, stratified into quartiles of the A compartment connectivity. Higher A compartment connectivity is correlated with higher gene expression. The Kruskal-Wallis *p*-value refers to the overall non-parametric test for expression differences across the A compartment connectivity quartiles. All pairwise comparisons were significant in the *post hoc* Dunn test. (B) Heatmap shows the level of correlation of the PAd A compartment connectivity with ChromHMM chromatin states and this

study's PAd ATAC-seq and pCHi-C data. Asterisks denote a significant Spearman's correlation (*p_{adj}<0.05). (C) Histogram of the differences in the A compartment connectivity between the lower and higher BMI subgroups of MZ twin siblings. The red dashed line at x=0 denotes the null hypothesis that there are not genome-wide connectivity differences between the twins. The *p*-value corresponds to the two-sided, one-sample Wilcoxon test for the connectivity differences, showing that the lower BMI twins exhibit a shift toward higher A compartment connectivity compared to the higher BMI twin siblings (shifted to the right of the 0). (D) Q-Q plots for the uniform distribution of the *p*-values for the gene-by-BMI effects on CRP in the UKB, stratified by whether the SNPs land in the reprogrammed A compartments (dark green, connectivity difference permutation $p<0.01$) or the non-reprogrammed A compartments (light green, connectivity difference permutation $p\geq0.01$). Confidence intervals (dashed lines) were calculated for the reprogrammed A compartment *p*-values. The *p*-value corresponds to the Wilcoxon Rank Sum test for differences in the *p*-value distribution between the reprogrammed and non-reprogrammed A compartments, showing that the reprogrammed A compartments have a higher accumulation of low *p*-value GxE SNPs affecting CRP levels in the UKB. See also Table S2.

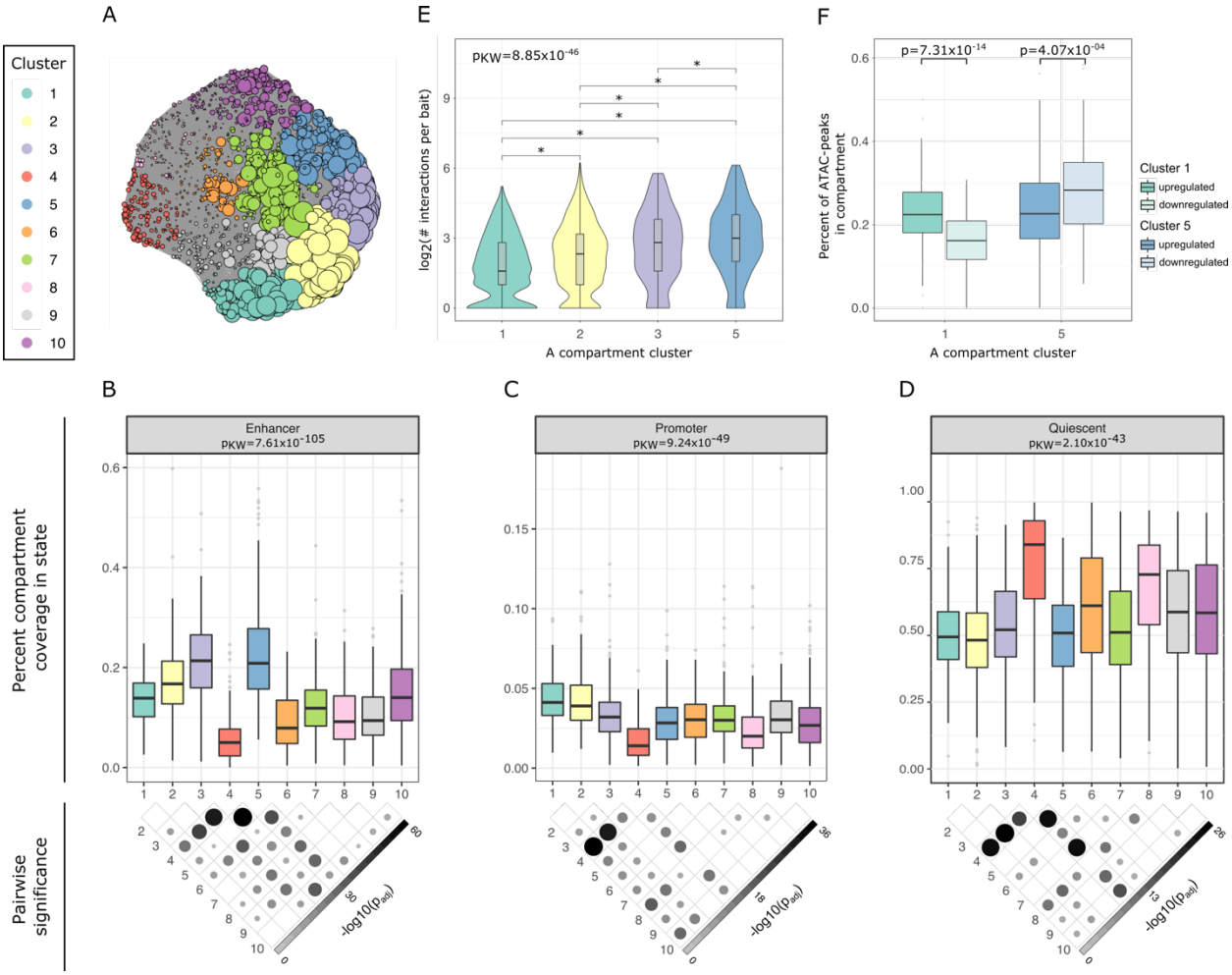


Figure 3. The A compartment clustering reveals differential accumulation of chromatin states and gene regulatory landscapes. (A) iGraph layout of the A compartment clusters after UMAP dimensionality reduction and Louvain clustering. Each circle represents an A compartment. Colors represent the 10 clusters that were identified. The size of the circles is proportional to the level of connectivity of that A compartment, showing that clusters 1, 2, 3, and 5 have the highest levels of connectivity. The A compartment cluster coverage of enhancer (B), promoter (C), and quiescent (D) ChromHMM chromatin states are indicated with boxplots. Overall *p*-values (top) correspond to the Kruskal-Wallis test comparing the coverage across the

A compartment clusters. The *p*-value map below the plot denotes which pairwise differences are significant in the *post hoc* Dunn test, after correcting for multiple testing using the Benjamini-Hochberg procedure. (E) Violin plots show the number of pCHi-C interactions per promoter in the four A compartment clusters. The overall *p*-value corresponds to the Kruskal-Wallis test and the asterisks in the pairwise comparisons denote significant differences in the *post hoc* Dunn test, after correcting for multiple testing using the Benjamini-Hochberg procedure. (F) Boxplots show the proportion of ATAC-peaks in the cluster A compartments that are upregulated (higher accessibility in D1 relative to PAd) or downregulated (lower expression in D1 relative to PAd) after 24 hours of PAd differentiation into adipocytes. The *p*-values correspond to the Wilcoxon Rank Sum test for differences between the proportion of up- or down-regulated peaks in each of the compartment clusters separately. See also Figures S4-5 and Tables S2-4.

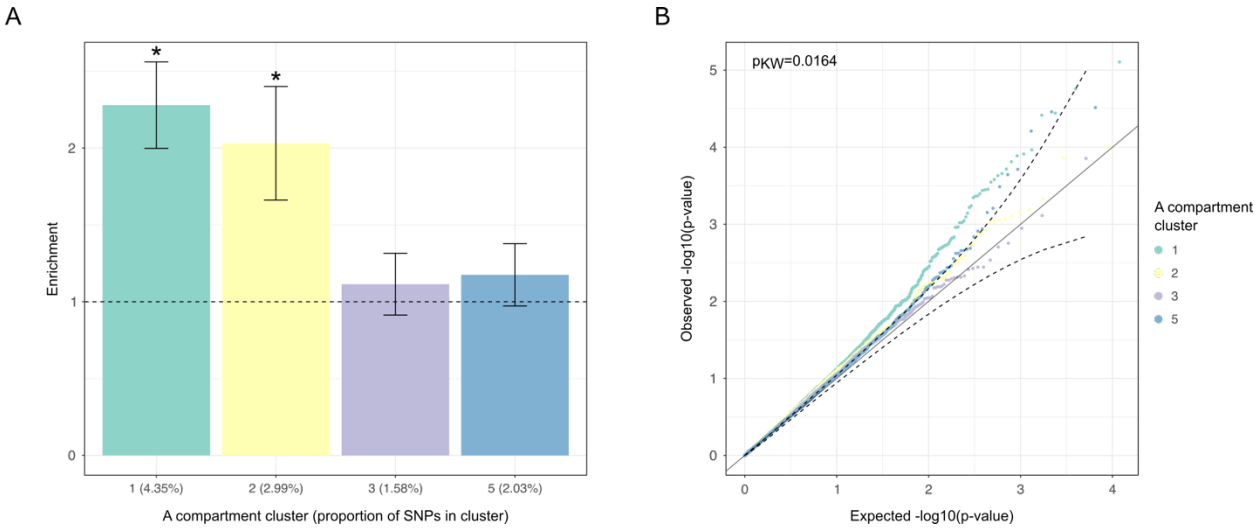


Figure 4. The A compartment cluster 1 contributes significantly to the heritability of CRP and is enriched for gene-BMI effects on CRP in the UK Biobank. (A) Barplot shows the enrichment of heritability for CRP in the different A compartment clusters relative to the null hypothesis of the uniform contribution from all SNPs. Error bars represent the standard error. Asterisks refer to an enrichment adjusted *p*-value of less than 0.05. The x-axis tick marks list the cluster with the proportion of SNPs in that cluster in parentheses. (B) Q-Q plots for the uniform distribution of the *p*-values for the gene-by-BMI effects on CRP in the UKB, stratified by which of the A compartment clusters the SNP lands in. Confidence intervals (dashed lines) were calculated for the A compartment cluster 3. The overall *p*-value corresponds to the Kruskal-Wallis test for differences among all cluster *p*-value distributions. Cluster 1 had a higher accumulation of low *p*-value SNPs than cluster 5 in the *post hoc* Dunn test.

Table 1. Partitioned LDSC analysis shows that the A compartments are significantly enriched for the proportion of CRP heritability while the B compartments are depleted.

PAd A/B compartment category	Proportion of SNPs	Proportion of h^2	Proportion of h^2 SE	Enrichment	Enrichment SE	Enrichment *p*-value
B	0.773	0.650	0.017	0.840	0.022	1.99×10^{-09}
A (not reprogrammed in MZ twins)	0.199	0.265	0.020	1.33	0.099	7.05×10^{-04}
A (reprogrammed in MZ twins)	0.0276	0.0856	0.020	3.10	0.72	6.19×10^{-03}

SE indicates standard error, and h^2**,** heritability. Partitioned LD score regression (LDSC) (Bulik-Sullivan *et al.*, 2015; Finucane *et al.*, 2015) was performed using the C-reactive protein (CRP) summary statistics from the UK Biobank round 2 GWAS results hosted at the Neale Lab website (http://www.nealelab.is/uk-biobank/). Heritability was partitioned using the SNPs in the identified PAd B and A compartments, the latter of which was further stratified into the PAd A compartments that were either not reprogrammed, or were reprogrammed in the higher BMI twin siblings from the BMI-discordant MZ twins. The enrichment *p*-value and the SE for the proportion of h^2 and enrichment were calculated from the block jackknife resampling using the LDSC method. The B compartments exhibit a significant depletion in the proportion of h^2 explained, similar to repressed regions of the genome (Finucane *et al.*, 2015); whereas the A compartments are significantly enriched for the proportion of CRP h^2 explained in these regions.

Methods

Study cohorts

For the preadipocyte (PAd) sample collection, this study included a subset of 10 pairs (n=20) out of the 65 pairs of BMI-discordant (ΔBMI$\geq$3kg/m^2) MZ twins in the Finnish monozygotic twin cohort (FTC), collected and recruited in the Helsinki University Central Hospital, Helsinki, Finland, as described previously (5,6). The cohort mean age is 46 y (+/- 2 y (s.e.)), and it comprises 27 pairs of males (42%) and 38 pairs of females (58%). The subset of the cohort used in this study has a mean age of 42 y (+/- 4 y (s.e.)), and it comprises 6 pairs of males (60%) and 4 pairs of females (40%). The clinical characteristics are shown in Table S1. All participants gave written informed consent, and the study protocol was approved by the local ethics committee. The PAd sample collection is described below. For the C-reactive protein (CRP) gene-environment interaction (GxE) analysis, we used the UK Biobank (UKB) cohort (n=up to 372,652 non-related Europeans) (Bycroft *et al.*, 2018), under the Application Number 33934. The details of the GxE analysis are described below.

PAd collection and cell culture from the BMI-discordant MZ twins' adipose biopsies

We isolated the PAd from the subcutaneous fat biopsy specimens of the twins undergoing adipose biopsies. Briefly, the biopsy specimens were first treated with collagenase and then centrifuged to separate the adipocytes from the stromal vascular fraction (SVF) pellet. Next, the SVF pellet was suspended to PAd basal media with 5% fetal calf serum supplemented with 1% penicillin-streptomycin. Then, the SVF was filtered before plating to allow the PAd to adhere and propagate. Finally, the viable PAd were cryopreserved for the downstream ATAC-seq and RNA-seq experiments (see below). To optimally preserve the in vivo epigenetic characteristics

of these primary human cells, we will use the earliest passages (no more than 5) for all experiments.

For the experiments, cryopreserved cells (passage 3-4) were seeded into PromoCell PAd growth medium (PromoCell C-27410) with 1% Gibco Penicillin-Streptomycin (ThermoFisher 15140122) and cultured according to PromoCell PAd culturing protocols. Cells were maintained in a monolayer culture at 37°C and 5% CO_2. The primary PAd were passaged once for plating after propagation, resulting in fewer than 5 passages before collection for the experiments. We grew cells to <90% confluency for the PAd sample collection. We grew the cells to 100% confluency to begin the differentiation of these cells, using PAd differentiation medium (PromoCell C-27436) for the 24-hr differentiated (D1) time point. We collected two replicates of the PAd (for the ATAC-seq and RNA-seq) and one replicate of the D1 cells (for ATAC-seq), per individual.

Assay for transposase accessible chromatin (ATAC) –sequencing and data processing

We performed the ATAC-seq protocol in the PAd and D1 cells from each individual, including technical replicates from 2 twin pairs. We followed the omni-ATAC protocol(Corces *et al.*, 2017), beginning with the DNase I treatment, and then trypsinized cells using the PromoCell detach kit (PromoCell C-41210), according to the manufacturer's protocols. We then performed the tagmentation and DNA purification reaction as described in (Garske *et al.*, 2019). Libraries were sequenced on the Illumina HiSeq 4000 to produce an average of 40,315,572 (+/- 14,577,770) reads.

We processed the sequencing reads and performed quality control (QC) using the ENCODE ATAC-seq Data Standards and Prototype Processing Pipeline. Briefly, we aligned

reads to the human reference genome (GRCh37/hg19) using Bowtie2 v2.2.9 (Langmead and Salzberg, 2012) (with parameters -k 4 -X 2000 --local), filtering out unpaired mapped reads and reads with MAPQ < 30 (Samtools (Li *et al.*, 2009)) and duplicates (marked with Picard Tools). Only reads from the autosomes were retained for downstream analyses.

Forty out of the 48 samples were retained for downstream analyses. One pair (n=2) of twins failed the differentiation step; one sample had poor tagmentation and did not exhibit the proper fragment size distribution; 4 samples did not pass library complexity thresholds as defined by the ENCODA ATAC-seq Data Standards; and one sample had too few sequencing reads.

For the samples that passed the QC, we called consensus peaks on all samples combined, after removing the technical replicates (n = 33). Peaks were called using MACS2 (Zhang *et al.*, 2008) v2.2.7.1 and peaks with an FDR<0.05 were retained. We filtered out peaks in blacklisted regions, and retained peaks with counts per million (cpm) mapped reads >=1 in at least 10% of the samples.

A/B compartment detection from ATAC-seq data

We performed the A/B compartment detection as described previously (Fortin and Hansen, 2015). Briefly, first we binned the PAd ATAC-seq sequencing reads into 100-kb bins across the genome, except for reads landing in blacklisted regions (Amemiya, Kundaje and Boyle, 2019). We calculated the bins per million mapped reads (BPMs) and corrected the $\log_2$-transformed BPMs for family ID (as a random effect), age, sex, and FRiP, using the lme4 (Bates *et al.*, 2015) v1.1 R package. Next, we obtained the Spearman's rank correlation matrix of the bins to get the pairwise bin co-accessibility measures. To call the A/B compartments, we calculated the first

eigenvector of the correlation matrix, by chromosome, using the nipals function in the mixOmics (Rohart *et al.*, 2017) v6.10.9 R package. Since the sign of the eigenvector is arbitrary, we used the known fact that the B compartments are generally more correlated than A compartments (Fortin and Hansen, 2015). Thus, we correlated the eigenvector with the level of correlation of the compartment (sum of the correlation coefficients with all other bins on the chromosome), and ensured that the positive values in the eigenvector are negatively correlated with the level of bin correlation to denote A compartments as positive an B compartments as negative, changing the sign of the eigenvector if necessary. Next, we smoothed the eigenvector using a simple moving average with a bin size of 3 and obtained the final set of A/B compartments.

To permute the compartment locations for assessing the enrichment of pCHi-C in A compartments, we used the bedtools (Quinlan and Hall, 2010) shuffle command with the -noOverlapping option, the -chrom option to shuffle the compartments within the same chromosome and the -excl option to exclude blacklisted regions that were removed when identifying A/B compartments.

To calculate the chromatin state coverage in the A/B compartments, we downloaded the ChromHMM (Ernst and Kellis, 2015) 25-state segmentation across 127 reference epigenomes from the Roadmap Epigenomics Project. We determined the compartment coverage for each subset of ChromHMM states (enhancers, promoters, quiescent, and active) using bedtools (Quinlan and Hall, 2010) intersect function and dividing by the length of the compartment.

RNA-sequencing and data processing

We isolated and purified RNA from the PAd cells from the 10 pairs of twins, resulting in a total 20 samples. Cells were washed with PBS once before lysing with TriZol (Invitrogen 15596026)

and purified using Direct-zol RNA Mini-Prep (Zymo Research R2061). Libraries were prepared using the Illumina TruSeq Stranded mRNA kit and sequenced on an Illumina HiSeq 4000 instrument for an average sequencing depth of 78M reads (+/- 28M reads) per sample.

Reads were aligned to hg19 with STAR v2.7.0e (Dobin *et al.*, 2013), using the 2-pass method and the following parameters: --outFilterMultimapNmax 1, --outFilterMismatchNmax 6, --alignIntronMin 20, --alignIntronMax 500000, --chimSegmentMin 15. The various technical factors were obtained from STAR v2.7.0e (Dobin *et al.*, 2013) after sequence alignment (uniquely mapped reads) or from the Picard Tools v2.9.0 (option CollectRnaSeqMetrics). We only retained genes with =>1 cpm mapped reads in at least 10% of the samples.

Promoter Capture Hi-C (pCHi-C) cell culture and library preparation

Human subcutaneous primary white PAd were obtained from Zen-Bio (lot L120116E, female, age 52, BMI 26.5). Cells were maintained in a monolayer culture at 37°C and 5% CO_2 using PAd growth medium (PromoCell C-27410) with 1% Gibco Penicillin-Streptomycin (ThermoFisher 15140122) and following PromoCell PAd culturing protocols. We grew cells to <90% confluency (~7M cells) for two isogenic biological replicates of PAd. We fixed the cells directly in the culture plate and proceeded with the pCHi-C library preparation, as described previously (Pan *et al.*, 2018; Garske *et al.*, 2019).

We then followed the Agilent SureSelect^{XT} manufacturer instructions for the pre-capture PCR, using 8 cycles. We amplified the libraries in a post-capture PCR to add indexes, using 12 PCR cycles. In total, six libraries (two libraries were not associated with this publication) were sequenced together across 2 lanes on the Illumina HiSeq 4000 to produce 107M (rep1) and 92M (rep2) sequencing reads.

Promoter Capture Hi-C data processing and interaction calling

We processed the sequencing reads as described previously (Pan *et al.*, 2018; Garske *et al.*, 2019), using the Hi-C User Pipeline (HiCUP) v0.5.9 (Wingett *et al.*, 2015) default settings, except the insert size restrictions for the filtering step were set to 200-600 bp. A total of 46M (rep1) and 39M (rep2) reads per sample were aligned and passed the filtering and deduplication steps, with 38M (rep1) and 32M (rep2) on-target, for an average capture efficiency of 81.8%. We called significant interactions with the Capture Hi-C Analysis of Genome Organization (CHiCAGO) software v1.1.1 (Cairns *et al.*, 2016). We used the default threshold of 5 for calling significant interactions. To create the set of interactions for our downstream analyses, we used CHiCAGO to call interactions on the combined 2 biological replicates.

The A compartment connectivity analysis

To quantify the compartment connectivity, we first calculated the bins per million mapped reads (BPMs) for the A compartments and corrected the $\log_2$-transformed BPMs for family ID (as a random effect), age, sex, and FRiP, using the lme4 (Bates *et al.*, 2015) v1.1 R package. Next, we obtained the Spearman's rank correlation matrix of the bins to get the pairwise bin co-accessibility measures. The connectivity per compartment is calculated as the sum of the compartment adjacency with all other compartments genome-wide, divided by the total number of compartments. Adjacency is equal to 0 if the Spearman's $\rho<0.6$ and equal to 1 if the Spearman's $\rho>=0.6$.

To compute the differences in compartment connectivity between the lower and higher BMI siblings, we separated the MZ twin pairs into subgroups containing the lower BMI siblings

(n=9) and higher BMI siblings (n=9). We then calculated the compartment connectivity in the two subgroups separately, and compared the differences in connectivity at the compartment level by subtracting the connectivity value in the higher BMI MZ sibling group from the connectivity value in the lower BMI MZ sibling group. We used the paired Wilcoxon Rank Sum test to compare the level of connectivity between the lower and higher BMI siblings at the genome-wide level.

To identify individual compartments that are significantly different between the lower and higher BMI subgroups, we permuted the BMI status (higher or lower) within each MZ twin pair and re-calculated the connectivity differences of the compartments for all permutation (9 pairs^2=512 permutations). For each permutation, we compared the difference in connectivity for a given compartment between the groups and compared with the true difference in connectivity between the lower and higher BMI groups of twins. We calculated the number of permutations that exhibited a higher connectivity in the lower BMI compared to higher BMI groups of siblings than the true difference (one-sided), given our previous result that there is a shift toward higher connectivity in the lower BMI sibling group at the genome-wide level. Compartments with a permutation *p*-value of <0.01 (n=121, totaling 88.5 Mb) were defined as reprogrammed in the BMI-discordant twins and selected for downstream analyses.

Partitioned LD Score (LDSC) regression

We used the partitioned LDSC regression method(Bulik-Sullivan *et al.*, 2015; Finucane *et al.*, 2015) v1.0.1 to estimate the heritability of C-reactive protein (CRP) explained by the A/B compartments (stratifying the A compartments into the reprogrammed or non-reprogrammed compartments); or partitioned across the A compartment clusters. We downloaded the summary

statistics for the second round of the UKB GWAS performed by the Neale Group and colleagues (http://www.nealelab.is/uk-biobank/).

GxE analysis in the UKB

We downloaded the imputed genotype data from the UKB cohort (Bycroft *et al.*, 2018). We selected unrelated individuals of European ancestry who had both body mass index (BMI) and CRP measurements collected. We performed inverse normal transformation of the CRP values and corrected for age, age^2, sex, assessment center ID, array type, and the first 20 genetic principal components. The BMI values were centered and scaled.

For the gene-environment interaction (GxE) analysis, we filtered out SNPs with a minor allele frequency <1% and genotyping missing rate of >5%. To test for the interaction between the SNP and BMI, we used plink (Purcell *et al.*, 2007) v1.90b3.45 to test the effect of the SNP, BMI, and BMIxSNP in a linear model.

A compartment dimensionality reduction and clustering

Using UMAP for dimensionality reduction prior to clustering, as opposed to pairwise correlations to create an adjacency matrix, has previously been shown to improve the detection of true genetic interactions (Dorrity *et al.*, 2020). To do this, we first binned the PAd ATAC-seq sequencing reads into the identified PAd A compartments, except for reads landing in blacklisted regions (Amemiya, Kundaje and Boyle, 2019). We calculated the bins per million mapped reads (BPMs) and corrected the $\log_2$-transformed BPMs for family ID (as a random effect), age, sex, and FRiP, using the lme4 (Bates *et al.*, 2015) v1.1 R package. We performed dimensionality reduction and clustering following the previously published methodology (Dorrity *et al.*, 2020).

We performed principal component analysis (PCA) on the corrected PAd BPMs using the prcomp function in R. We then performed an additional dimensionality reduction to 2 components using Uniform Manifold Approximation and Projection (UMAP) (McInnes *et al.*, 2018) in the umap v0.2.7.0 R package, with n_neighbors set to 10 and min_dist set to 0.05. We obtained the 75 nearest neighbors based on the UMAP projections for each compartment, which corresponds to the mean number of compartments each compartment is correlated with in pairwise correlation analyses. This was done using the get.knn function in the FNN (Beygelzimer *et al.*, 2013) v1.1.3 R package. Louvain clustering was performed on the resulting adjacency matrix, using the iGraph (Csardi and Nepusz, 2006) v1.2.6 R package, to obtain the 10 A compartment clusters used for downstream analyses.

For assessing the statistical significance of the differences between the A compartment clusters, we used the Kruskal-Wallis test and applied the dunnTest function in the FSA (Ogle, 2017) v0.8.32 R package for the *post hoc* test to determine which pairwise comparisons are significant after correcting for multiple testing using the Benjamini-Hochberg procedure.

Preadipocyte super-enhancer identification

We downloaded the raw FASTQ ChIP-seq data for the H2K27ac histone mark and MED1 at the day 1 adipogenic time point (Rauch *et al.*, 2019) from bone marrow derived stromal stem cells (BM-hMSC-TERT4) from the GEO database (accession code GSE113253). We processed the ChIP-seq data according to the ENCODE ChIP-seq pipeline. Briefly, sequencing reads were aligned to the hg19 reference genome using Bowtie2 v2.2.9 (Langmead and Salzberg, 2012) (with parameters -k 4 --local), filtering out unmapped reads and reads with MAPQ<30 (Samtools

(Li *et al.*, 2009)) and duplicates (marked with Picard Tools). Only reads from the autosomes were retained for downstream analyses.

Peaks were called on each biological replicate separately using MACS2 (Zhang *et al.*, 2008) v2.2.7.1 and then consensus peaks were called on both replicates together to run the irreproducible discovery rate (IDR) analysis to identify reproducible peaks across both replicates. Only MED1 peaks that overlapped with H3K27ac peaks were retained as the constituent peaks for downstream analyses to identify super-enhancers. The ROSE algorithm (Lovén *et al.*, 2013; Whyte *et al.*, 2013) was used to call super-enhancers based on the MED1 ChIP-seq alignments.

Gene ontology (GO) term enrichment in the A compartment clusters

We performed a gene ontology (GO) enrichment analysis on the compartment genes in each A compartment cluster separately. As performing enrichment analyses on genes selected from large genomic regions can lead to spurious enrichments due to clusters of gene families or genes with similar functions (Pazos Obregón *et al.*, 2018), we used the Network Enrichment Analysis Tool (NEAT) (Signorelli, Vinciotti and Wit, 2016). NEAT uses information about the relationship between genes (e.g., genes in the same co-expression network) to test for functional enrichment, thereby requiring additional information about the gene function in that cell-type, on top of simply the region of the genome in which it lands.

To provide the network information to NEAT, we performed weighted gene co-expression network analysis (WGCNA) (Langfelder and Horvath, 2008) using the RNA-seq data from all of the A compartment genes together. We followed default WGCNA procedures except that we used a soft power value of 12 and performed a signed analysis. Genes from each A compartment cluster were assigned a co-expression module, and this network information was

provided to NEAT for the A compartment GO enrichment. We downloaded the GO slims from the PANTHER (Mi *et al.*, 2021) database. We used an alpha of 0.005 as the cutoff for GO term significance to correct for testing 10 A compartment clusters separately, and then thresholded the within-compartment *p*-values using an FDR<0.05 as the significance cut-point.

To summarize the cluster GO terms based on semantic similarity, we used the online tool REVIGO (Supek *et al.*, 2011). We used the simRel method for clustering, and then quantified the number of GO terms that are listed under each indispensable GO term from the REVIGO output.

Identification of differentially accessible ATAC-seq peaks

We performed the differential accessibility (DA) analysis between the PAd and D1 time points using the R package limma v3.34.9 (Ritchie *et al.*, 2015; Phipson *et al.*, 2016) and the voom (Law *et al.*, 2014) method. We used the duplicateCorrelation function in limma (Smyth, Michaud and Scott, 2005) to account for the repeated measure from the same individual. To decrease confounding, we included age, sex, and fraction of reads in peaks (FRiP) as covariates in the model and the family ID as a blocking factor. We tested for DA between the PAd and D1 time points in the lower and higher BMI individuals separately. We used an FDR < 0.05 cutoff to define significant DA peaks for these comparisons.

Supplemental information titles and legends

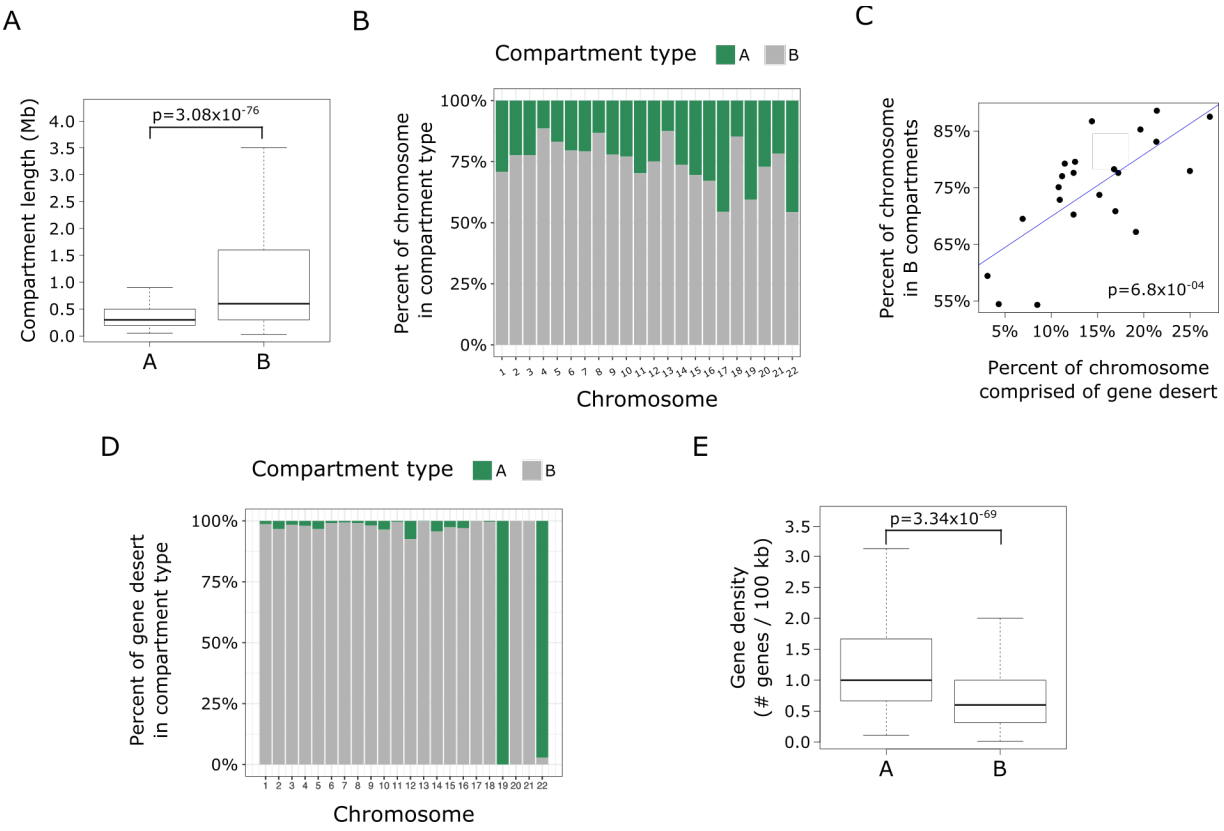


Figure S1. Related to Figure 1. A/B compartment detection in preadipocytes reflects known genomic hallmarks of A and B compartments. (A) Boxplots show the distribution of A/B compartment lengths. Outliers have been removed for clarity. The *p*-value corresponds to the Wilcoxon Rank Sum test comparing the lengths between the A and B compartments. (B) Bar plots show the proportion of each chromosome that makes up A/B compartments. (C) Scatterplot shows the relationship between the proportion of chromosomes made up of gene deserts, and the proportion of chromosomes made up of B compartments. The *p*-value corresponds to the significance of the Spearman's rank correlation coefficient. (D) Bar plots show the percent of gene deserts that overlap A or B compartments. (E) Boxplots show the gene density distribution

in A/B compartments. Outliers have been removed for clarity. The *p*-value corresponds to the Wilcoxon Rank Sum test comparing the gene density between the A and B compartments.

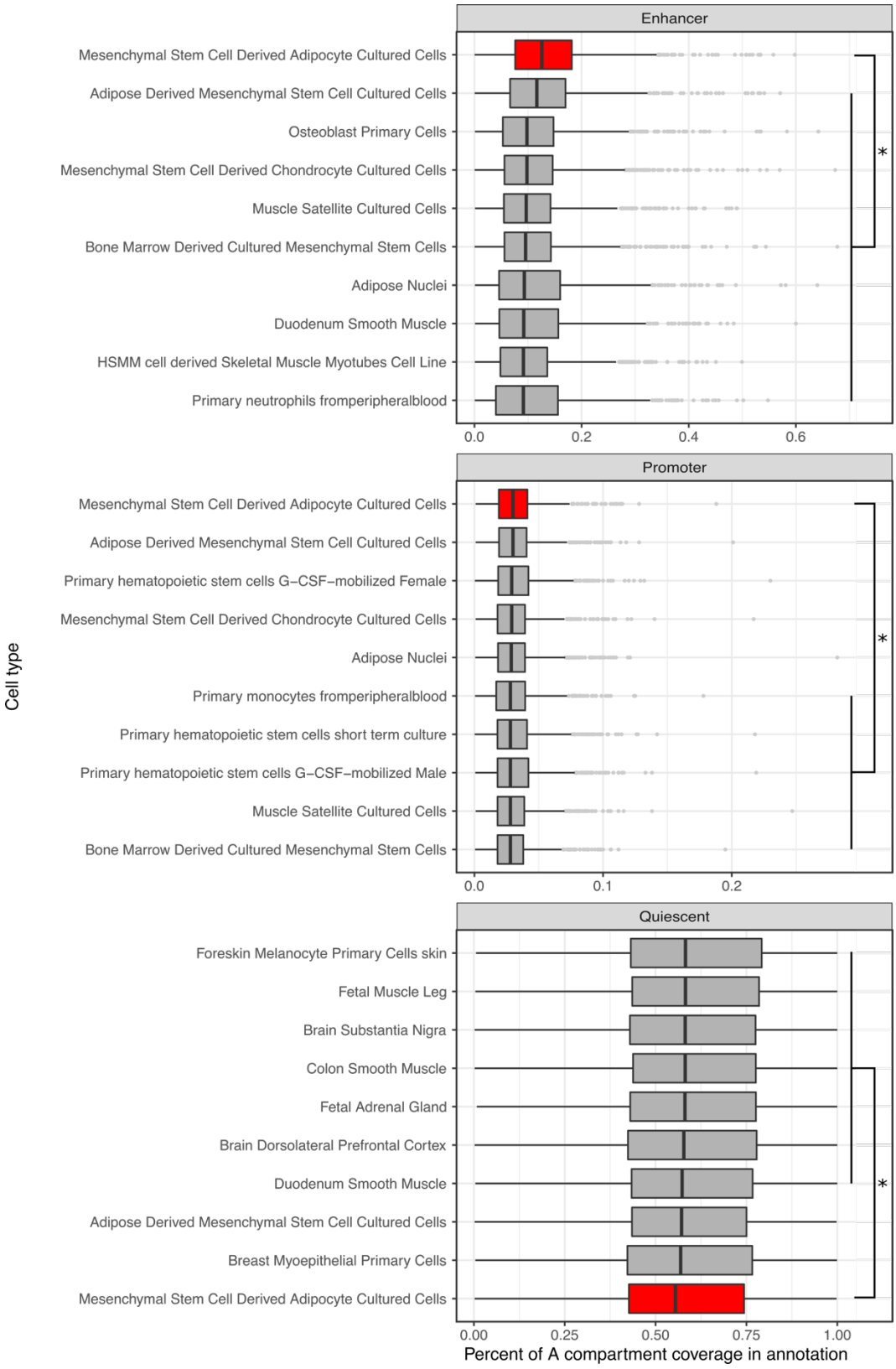

Enhancer
Mesenchymal Stem Cell Derived Adipocyte Cultured Cells
Adipose Derived Mesenchymal Stem Cell Cultured Cells
Osteoblast Primary Cells
Mesenchymal Stem Cell Derived Chondrocyte Cultured Cells
Muscle Satellite Cultured Cells
Bone Marrow Derived Cultured Mesenchymal Stem Cells
Adipose Nuclei
Duodenum Smooth Muscle
HSMM cell derived Skeletal Muscle Myotubes Cell Line
Primary neutrophils fromperipheralblood
0.0
0.2
0.4
0.6
*
Promoter
Mesenchymal Stem Cell Derived Adipocyte Cultured Cells
Adipose Derived Mesenchymal Stem Cell Cultured Cells
Primary hematopoietic stem cells G−CSF−mobilized Female
Mesenchymal Stem Cell Derived Chondrocyte Cultured Cells
Adipose Nuclei
Primary monocytes fromperipheralblood
Primary hematopoietic stem cells short term culture
Primary hematopoietic stem cells G−CSF−mobilized Male
Muscle Satellite Cultured Cells
Bone Marrow Derived Cultured Mesenchymal Stem Cells
0.0
0.1
0.2
*
Quiescent
Foreskin Melanocyte Primary Cells skin
Fetal Muscle Leg
Brain Substantia Nigra
Colon Smooth Muscle
Fetal Adrenal Gland
Brain Dorsolateral Prefrontal Cortex
Duodenum Smooth Muscle
Adipose Derived Mesenchymal Stem Cell Cultured Cells
Breast Myoepithelial Primary Cells
Mesenchymal Stem Cell Derived Adipocyte Cultured Cells
0.00
0.25
0.50
0.75
1.00
*
Cell type
Percent of A compartment coverage in annotation

Figure S2. Related to Figure 1. The A compartment coverage of enhancer chromatin states across 127 ENCODE samples highlights the cell-type-specificity of the A compartment identification. Boxplots show the coverage of enhancer, promoter, and quiescent ChromHMM chromatin states across all of the A compartments for the top 10 ENCODE cell or tissue types (for the enhancer and promoter coverage); or the bottom 10 ENCODE cell or tissue types (for the quiescent coverage). The cell-type with the highest coverage of enhancer marks and promoter marks, as well as with the lowest coverage of quiescent marks, is denoted in red (mesenchymal stem cell -derived adipocyte cultured cells (MSC-Ad)). Asterisks correspond to $p<0.05$ the Wilcoxon Rank Sum test, comparing the MSC-Ad cells to all other cell- and tissue types, after correcting for multiple testing using the Benjamini-Hochberg procedure.

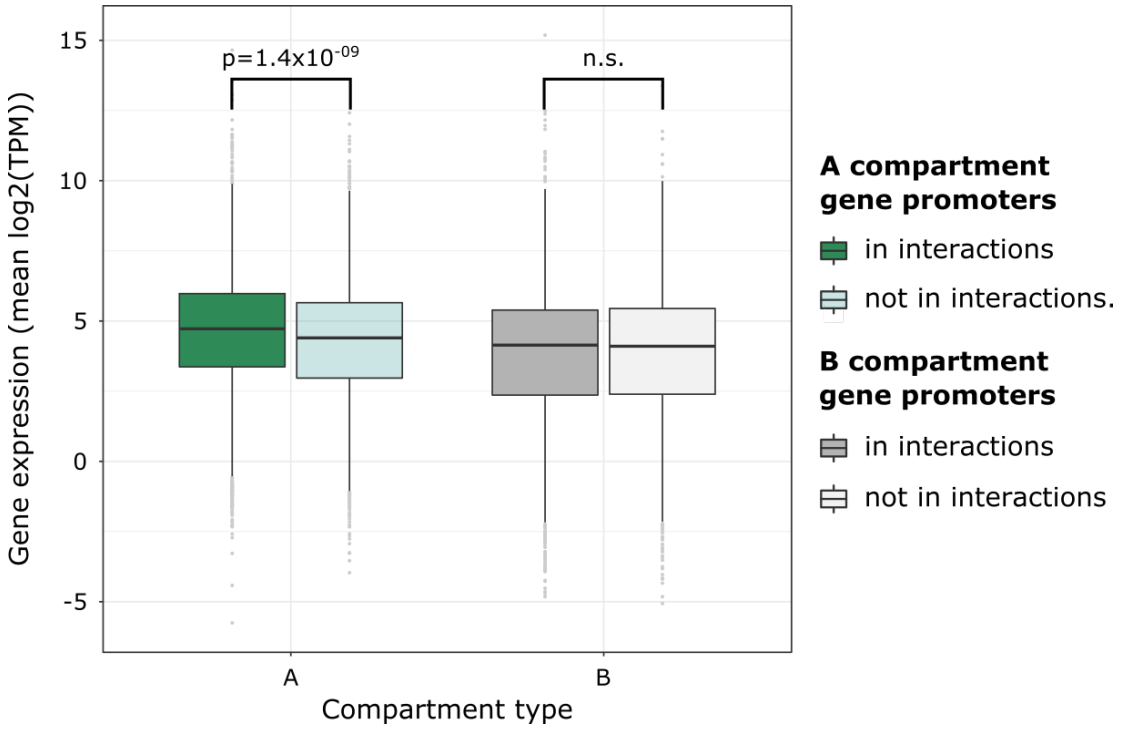


Figure S3. Related to Figure 1. Promoter Capture Hi-C in PAd reveals how promoter interaction effects are dependent upon A/B compartmentalization. (A) Boxplot shows the expression of genes either involved in pCHi-C interactions (darker) or not (lighter) for genes with promoters in the A compartments (left) or the B compartments (right). *P*-value corresponds to the Wilcoxon Rank Sum test comparing gene expression within each compartment type separately.

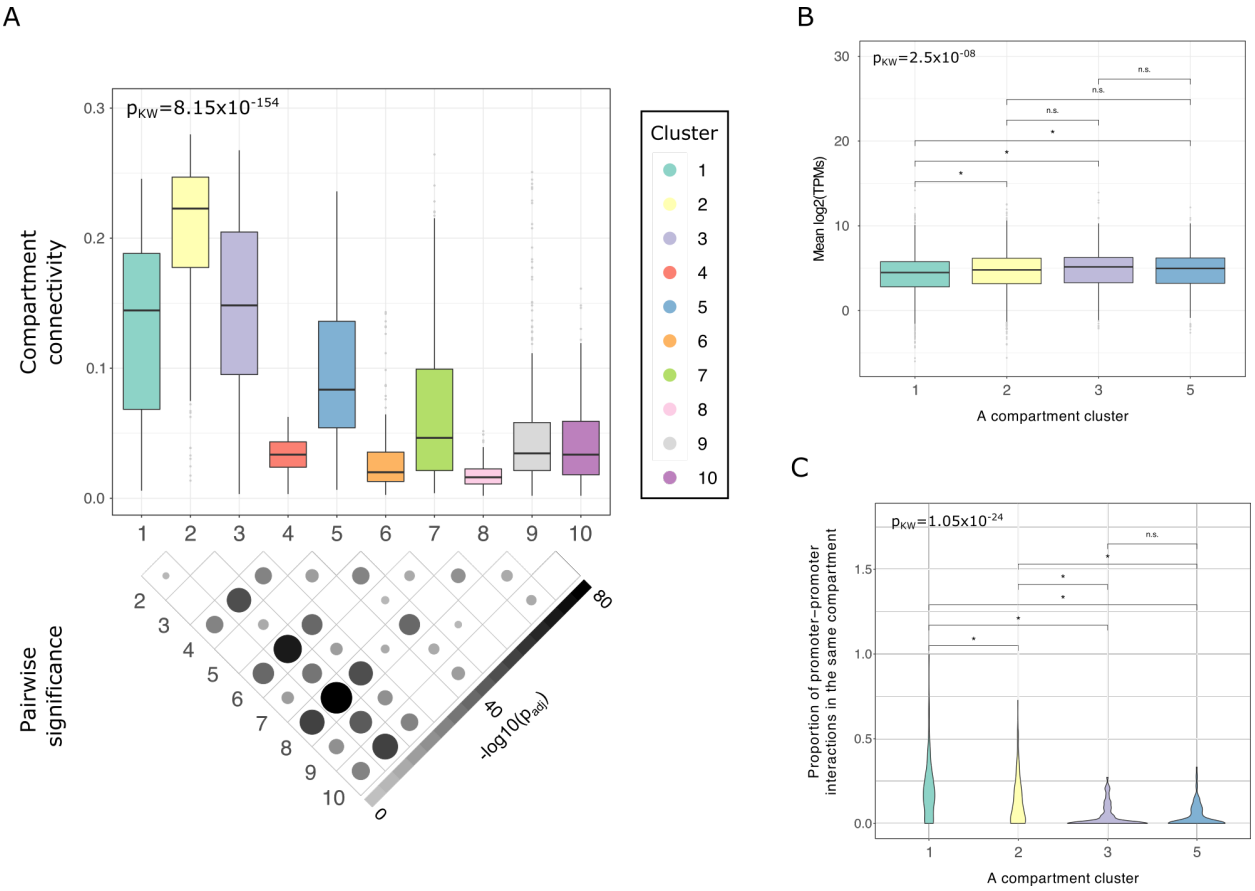


Figure S4. Related to Figure 3. The A compartment clustering stratifies clusters based on the level of connectivity and cellular priming, assessed through promoter-promoter interactions and gene expression. (A) Boxplots show the connectivity distributions across the 10 A compartment clusters. The overall *p*-value corresponds to the Kruskal-Wallis test and the *p*-value map below the plot denotes which pairwise differences are significant in the *post hoc* Dunn test, after correcting for multiple testing using the Benjamini-Hochberg procedure. (B) Violin plots show the proportion of pCHi-C interactions within the same A compartment that are promoter-promoter interactions. The overall *p*-value corresponds to the Kruskal-Wallis test and the pairwise significance is determined by the *post hoc* Dunn test. (C) Boxplots show the mean

gene expression ($\log_2$(TPM) of genes in each of the A compartment clusters. The overall *p*-value corresponds to the Kruskal-Wallis test, and the pairwise significance is determined by the *post hoc* Dunn test.

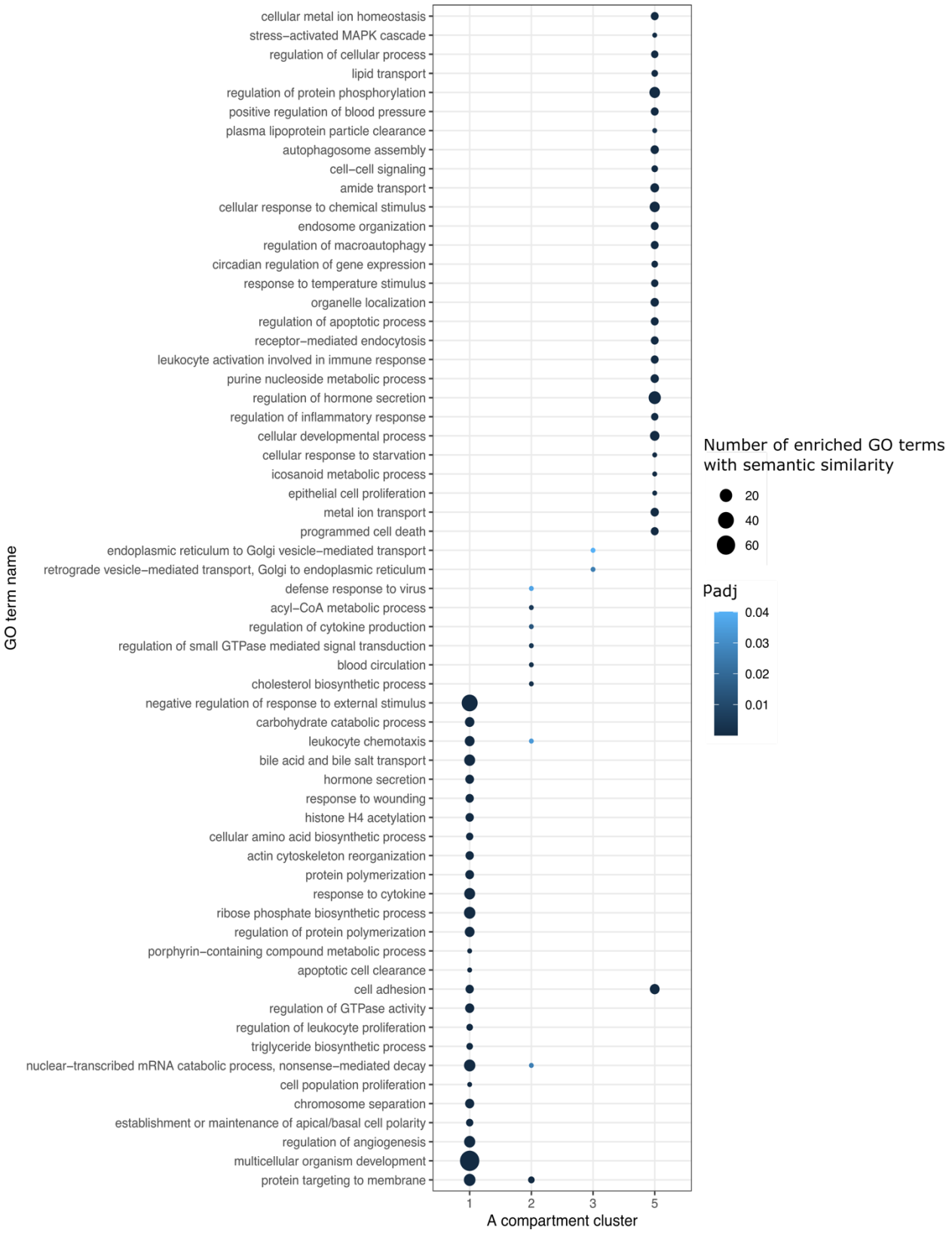

cellular metal ion homeostasis
stress−activated MAPK cascade
regulation of cellular process
lipid transport
regulation of protein phosphorylation
positive regulation of blood pressure
plasma lipoprotein particle clearance
autophagosome assembly
cell−cell signaling
amide transport
cellular response to chemical stimulus
endosome organization
regulation of macroautophagy
circadian regulation of gene expression
response to temperature stimulus
organelle localization
regulation of apoptotic process
receptor−mediated endocytosis
leukocyte activation involved in immune response
purine nucleoside metabolic process
regulation of hormone secretion
regulation of inflammatory response
cellular developmental process
cellular response to starvation
icosanoid metabolic process
epithelial cell proliferation
metal ion transport
programmed cell death
endoplasmic reticulum to Golgi vesicle−mediated transport
retrograde vesicle−mediated transport, Golgi to endoplasmic reticulum
defense response to virus
acyl−CoA metabolic process
regulation of cytokine production
regulation of small GTPase mediated signal transduction
blood circulation
cholesterol biosynthetic process
negative regulation of response to external stimulus
carbohydrate catabolic process
leukocyte chemotaxis
bile acid and bile salt transport
hormone secretion
response to wounding
histone H4 acetylation
cellular amino acid biosynthetic process
actin cytoskeleton reorganization
protein polymerization
response to cytokine
ribose phosphate biosynthetic process
regulation of protein polymerization
porphyrin−containing compound metabolic process
apoptotic cell clearance
cell adhesion
regulation of GTPase activity
regulation of leukocyte proliferation
triglyceride biosynthetic process
nuclear−transcribed mRNA catabolic process, nonsense−mediated decay
cell population proliferation
chromosome separation
establishment or maintenance of apical/basal cell polarity
regulation of angiogenesis
multicellular organism development
protein targeting to membrane
GO term name
1
2
3
5
A compartment cluster
Number of enriched GO terms with semantic similarity
20
40
60
Padj
0.04
0.03
0.02
0.01

Figure S5. Related to Figure 3. Gene ontology (GO) term enrichment across the A compartment clusters. Dot plot shows significantly enriched GO terms related to biological processes in the A compartment clusters 1, 2, 3, and 5. The size of the circle is proportional to how many of the significantly enriched GO terms for that A compartment cluster were listed together after semantic similarity analysis was done to cluster similar GO terms. The semantic similarity analysis was performed using REVIGO.

Table S1. Differences in metabolic traits between the lower and higher BMI sibling groups in the BMI-discordant MZ twin cohort.

Trait	Full cohort lower BMI sibling mean (SE)	Full cohort higher BMI sibling mean (SE)	*p*-value (paired t-test)	This study lower BMI sibling mean (SE)	This study higher BMI sibling mean (SE)
BMI (n_{cohort}=130)	26.5 (0.6)	32.3 (0.7)	7.25E-27	27.7 (1.8)	33.9 (2.3)
body fat percent (n_{cohort} =130)	33.9 (1.1)	41 (0.9)	2.79E-19	33.4 (2.6)	40 (2.7)
body fat mass (kg) (n_{cohort} =126)	27.1 (1.4)	39.3 (1.5)	1.68E-19	30.7 (4.6)	43.7 (5.2)
fat free mass (kg) (n_{cohort} =126)	48.1 (1.3)	52.6 (1.5)	1.85E-08	54.9 (3.2)	59.6 (3.4)
waist (cm) (n_{cohort} =123)	90.2 (1.7)	105.5 (1.9)	1.11E-21	99 (4.5)	112.6 (4.8)
hip (cm) (n_{cohort} =124)	99.3 (1.4)	110.1 (1.7)	6.72E-23	103.9 (4.9)	114.4 (5.7)
WHR (n_{cohort} =129)	0.91 (0.01)	0.95 (0.01)	2.00E-08	0.96 (0.02)	0.99 (0.03)
liver fat percent (n_{cohort} =80)	1.4 (0.3)	4.7 (0.8)	8.46E-08	1.7 (0.5)	6.4 (2.4)
subcutaneous fat volume (cm^3) (n_{cohort} =52)	3814 (417)	6359 (540)	4.12E-10	n.a.	n.a.
intraabdominal fat volume (cm^3) (n_{cohort} =52)	790 (179)	1644 (247)	4.49E-09	n.a.	n.a.
adipocyte diameter (um) (n_{cohort} =122)	83 (2)	93 (2)	1.03E-07	80 (3)	82 (3)
adipocyte volume (um^3) (n_{cohort} =122)	450 (24)	631 (29)	1.28E-10	461 (38)	538 (45)

HbA1c (mmol/mol) (n=126)	5.4 (0.1)	5.7 (0.1)	4.37E-07	5.2 (0.1)	5.8 (0.4)
fasting glucose (mmol/L) (n_{cohort} =128)	5.5 (0.1)	5.9 (0.2)	1.74E-04	5.4 (0.1)	5.8 (0.3)
fasting insulin (pmol/L) (n_{cohort} =124)	6.5 (0.6)	9.7 (0.8)	9.54E-06	6.9 (1.9)	7.8 (1.2)
Matsuda index (n_{cohort} =113)	7.4 (0.5)	4.9 (0.4)	4.42E-07	7.0 (1.0)	4.7 (0.6)
HOMA-IR (n_{cohort} =119)	1.6 (0.2)	2.6 (0.3)	1.35E-06	1.7 (0.4)	1.9 (0.3)
total cholesterol (mmol/L) (n_{cohort} =124)	4.7 (0.1)	4.8 (0.1)	3.78E-01	4.7 (0.2)	4.8 (0.3)
HDL (mmol/L) (n_{cohort} =130)	1.6 (0.1)	1.4 (0.1)	4.95E-05	1.6 (0.2)	1.4 (0.1)
TG (mmol/L) (n_{cohort} =130)	0.98 (0.07)	1.19 (0.07)	2.19E-04	0.8 (0.1)	1.1 (0.1)
LDL (mmol/L) (n_{cohort} =130)	2.9 (0.1)	3.1 (0.1)	7.07E-02	3.0 (0.2)	3.2 (0.3)
CRP (mg/L) (n_{cohort} =121)	2.5 (0.4)	3.9 (0.6)	1.48E-04	1.59 (0.42)	3.69 (0.89)
AST (U/L) (n_{cohort} =130)	27.7 (1.2)	30.4 (2.2)	1.53E-01	26.0 (3.9)	39.3 (10.9)
ALT (U/L) (n_{cohort} =130)	25.9 (1.9)	31.7 (2.9)	8.04E-03	25.3 (5.4)	41.9 (11.2)
GGT (U/L) (n_{cohort} =84)	24.1 (3.3)	33 (5.2)	5.28E-03	25.2 (5.3)	29.1 (5.0)

SE indicates standard error; BMI, body mass index; WHR, waist-to-hip ratio; HbA1c, hemoglobin A1c; HOMA-IR, homeostatic model assessment for insulin resistance; HDL, high-density lipoprotein; TG, triglycerides; LDL, low-density lipoprotein; CRP, C-reactive protein; AST, aspartate transaminase; ALT, alanine transaminase; and GGT, gamma-glutamyltransferase.

Table S3. Related to Figure 3. The A compartment cluster numbers and Mb spanned.

A compartment cluster	# of A compartments in cluster	Total Mb	# of reprogrammed A compartments in cluster	Total Mb (reprogrammed)
1	145	129	30	37.3
2	144	88.8	39	32.5
3	107	42.6	13	5.0
4	156	52.8	8	1.8
5	138	60.6	11	4.0
6	137	39.3	1	0.1
7	193	78.1	7	4.0
8	121	27.1	6	1.1
9	180	69.2	4	2.4
10	230	62.3	2	0.3

Table S5. Related to Figure 4. CRP GWAS SNPs that land in pCHi-C interactions in the reprogrammed primed A compartment cluster 1.

Cluster 1 A compartment	# GWAS SNPs (LD R^2<0.2)	# GWAS signals in PAd pCHi-C and ATAC-seq	Tag SNP	Genes with GWAS SNP in interacting promoter	Genes interacting with GWAS SNP
chr1_A_comp22	2	-	-	-	-
chr3_A_comp31	5	3	rs374407247	*AMT, DAG1, BSN*	*RHOA, TCTA, APEH, QRICH1*
			rs2271961	*GMPPB, AMIGO3, IP6K1, FAM212A, CDHR4*	*IP6K1, TRAIP, CAMKV, MST1R, MON1A, RBM6*
			rs76367790	*MST1R, RBM6*	*UBA7, FAM212A, CDHR4, MST1R, 2, RBM6, IP6K1, MON1A*
chr6_A_comp26	2	1	rs154977	-	*HLA-DMB*
chr7_A_comp51	1	1	rs12333760	*PTCD1, ZNF789, ATP5J2, ZNF394, ZKSCAN5*	*ZNF394, ZKSCAN5, ZSCAN25*
chr11_A_comp25	1	-	-	-	-
chr12_A_comp66	5	1	rs28861152	-	*SETD8*
chr17_A_comp1	1	1	rs7502910	*SERPINF2, RP11-961A15.1*	-

Chr indicates chromosome; LD, linkage disequilibrium; GWAS, genome-wide association study; SNP, single nucleotide polymorphism; ATAC, Assay for Transposase Accessible Chromatin. The “Cluster 1 A compartment” column is the unique A compartment identifier. Only Ensembl protein-coding and lincRNAs are shown for clarity.

References

Amemiya, H. M., Kundaje, A. and Boyle, A. P. (2019) 'The ENCODE Blacklist: Identification of Problematic Regions of the Genome', *Scientific Reports*, 9(1), p. 9354. doi: 10.1038/s41598-019-45839-z.

Andersen, E. *et al.* (2019) 'Preadipocytes from obese humans with type 2 diabetes are epigenetically reprogrammed at genes controlling adipose tissue function', *International Journal of Obesity*, 43(2). doi: 10.1038/s41366-018-0031-3.

Bates, D. *et al.* (2015) 'Fitting linear mixed-effects models using lme4', *Journal of Statistical Software*, 67(1), pp. 1–48. doi: 10.18637/jss.v067.i01.

Beygelzimer, A. *et al.* (2013) 'Package "FNN": Fast Nearest Neighbor Search Algorithms and Applications', *Cran*.

Blüher, M. (2020) 'Metabolically healthy obesity', *Endocrine Reviews*, 41(3), pp. 405–420. doi: 10.1210/endrev/bnaa004.

Bulik-Sullivan, B. *et al.* (2015) 'LD score regression distinguishes confounding from polygenicity in genome-wide association studies', *Nature Genetics*, 47(3). doi: 10.1038/ng.3211.

Bycroft, C. *et al.* (2018) 'The UK Biobank resource with deep phenotyping and genomic data', *Nature*, 562, pp. 203–209. doi: 10.1038/s41586-018-0579-z.

Cairns, J. *et al.* (2016) 'CHiCAGO : robust detection of DNA looping interactions in Capture Hi-C data', *Genome Biology*. Genome Biology, 17(127), p. 127. doi: 10.1186/s13059-016-0992-2.

Cercato, C. and Fonseca, F. A. (2019) 'Cardiovascular risk and obesity', *Diabetology & Metabolic Syndrome*, 11(1), p. 74. doi: 10.1186/s13098-019-0468-0.

Corces, M. R. *et al.* (2017) 'An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues', *Nature Methods*, 14, pp. 959–962. doi: 10.1038/nmeth.4396.

Csardi, G. and Nepusz, T. (2006) ‘Package “igraph” Title Network Analysis and Visualization’, *Interjournal*.

Davis, C. A. *et al.* (2018) ‘The Encyclopedia of DNA elements (ENCODE): Data portal update’, *Nucleic Acids Research*, 46(D1), pp. D794–D801. doi: 10.1093/nar/gkx1081.

Dobin, A. *et al.* (2013) ‘STAR: Ultrafast universal RNA-seq aligner’, *Bioinformatics*, 29(1), pp. 15–21.

Docherty, A. B. *et al.* (2020) ‘Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: Prospective observational cohort study’, *The BMJ*, 369, p. m1985. doi: 10.1136/bmj.m1985.

Dorrity, M. W. *et al.* (2020) ‘Dimensionality reduction by UMAP to visualize physical and genetic interactions’, *Nature Communications*, 11(1), p. 1537. doi: 10.1038/s41467-020-15351-4.

Ernst, J. and Kellis, M. (2015) ‘Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues.’, *Nature biotechnology*, 33(4), pp. 364–76. doi: 10.1038/nbt.3157.

Finucane, H. K. *et al.* (2015) ‘Partitioning heritability by functional annotation using genome-wide association summary statistics’, *Nature Genetics*. Nature Publishing Group, 47(11), pp. 1228–1235. doi: 10.1038/ng.3404.

Fortin, J. P. and Hansen, K. D. (2015) ‘Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data’, *Genome Biology*, 16, p. 180. doi: 10.1186/s13059-015-0741-y.

Garske, K. M. *et al.* (2019) ‘Reverse gene–environment interaction approach to identify variants influencing body-mass index in humans’, *Nature Metabolism*, 1(6), pp. 630–642. doi: 10.1038/s42255-019-0071-6.

Ghaben, A. L. and Scherer, P. E. (2019) ‘Adipogenesis and metabolic health’, *Nature Reviews Molecular Cell Biology*, 20, pp. 242–258. doi: 10.1038/s41580-018-0093-z.

Granér, M. *et al.* (2012) 'Epicardial fat, cardiac dimensions, and low-grade inflammation in young adult monozygotic twins discordant for obesity', *American Journal of Cardiology*, 109(9), pp. 1295–1302. doi: 10.1016/j.amjcard.2011.12.023.

Gulyaeva, O. *et al.* (2018) 'Sox9-Meis1 Inactivation Is Required for Adipogenesis, Advancing Pref-1+ to PDGFRα+ Cells', *Cell Reports*, 25(4). doi: 10.1016/j.celrep.2018.09.086.

Hotamisligil, G. S. (2006) 'Inflammation and metabolic disorders', *Nature*, pp. 860–867. doi: 10.1038/nature05485.

Isakson, P. *et al.* (2009) 'Impaired preadipocyte differentiation in human abdominal obesity: Role of Wnt, tumor necrosis factor-α, and inflammation', *Diabetes*, 58(7), pp. 1550–1557. doi: 10.2337/db08-1770.

Joshi, O. *et al.* (2015) 'Dynamic Reorganization of Extremely Long-Range Promoter-Promoter Interactions between Two States of Pluripotency', *Cell Stem Cell*, 17(6), pp. 748–757. doi: 10.1016/j.stem.2015.11.010.

Khera, A. V. *et al.* (2019) 'Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood', *Cell*, 177(3), pp. 587–596. doi: 10.1016/j.cell.2019.03.028.

Kivimäki, M. *et al.* (2017) 'Overweight, obesity, and risk of cardiometabolic multimorbidity: pooled analysis of individual-level data for 120 813 adults from 16 cohort studies from the USA and Europe', *The Lancet Public Health*, 2(6), pp. E277–E285. doi: 10.1016/S2468-2667(17)30074-9.

Langfelder, P. and Horvath, S. (2008) 'WGCNA: An R package for weighted correlation network analysis', *BMC Bioinformatics*, 9, p. 559. doi: 10.1186/1471-2105-9-559.

Langmead, B. and Salzberg, S. L. (2012) 'Fast gapped-read alignment with Bowtie 2', *Nat Methods*, 9(4), pp. 357–359. doi: 10.1038/nmeth.1923.

Lavie, C. J. *et al.* (2018) 'Management of cardiovascular diseases in patients with obesity', *Nature Reviews Cardiology*, pp. 45–56. doi: 10.1038/nrcardio.2017.108.

Law, C. W. *et al.* (2014) 'Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts', *Genome Biology*, 15(2), p. R29. doi: 10.1186/gb-2014-15-2-r29.

Li, H. *et al.* (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25(16), pp. 2078–2079. doi: 10.1093/bioinformatics/btp352.

Lieberman-Aiden, E. *et al.* (2009) 'Comprehensive mapping of long-range interactions reveals folding principles of the human genome.', *Science (New York, N.Y.)*, 326(5950), pp. 289–293. doi: 10.1126/science.1181369.

Lovén, J. *et al.* (2013) 'Selective inhibition of tumor oncogenes by disruption of super-enhancers', *Cell*, 153(2), pp. 320–334. doi: 10.1016/j.cell.2013.03.036.

MacDougald, O. A. and Mandrup, S. (2002) 'Adipogenesis: Forces that tip the scales', *Trends in Endocrinology and Metabolism*, pp. 5–11. doi: 10.1016/S1043-2760(01)00517-3.

McInnes, L. *et al.* (2018) 'UMAP: Uniform Manifold Approximation and Projection', *Journal of Open Source Software*, 3(29), p. 861. doi: 10.21105/joss.00861.

Mi, H. *et al.* (2021) 'PANTHER version 16: A revised family classification, tree-based classification tool, enhancer regions and extensive API', *Nucleic Acids Research*, 49(D1), pp. D394–D403. doi: 10.1093/nar/gkaa1106.

Naukkarinen, J. *et al.* (2014) 'Characterising metabolically healthy obesity in weight-discordant monozygotic twins', *Diabetologia*, 57(1), pp. 167–176. doi: 10.1007/s00125-013-3066-y.

Noronha Nc, N. D. C. *et al.* (2019) 'Priming approaches to improve the efficacy of mesenchymal stromal cell-based therapies', *Stem Cell Research and Therapy*, p. 131. doi: 10.1186/s13287-019-1224-y.

Ogle, D. H. (2017) *Package 'FSA' - Simple Fisheries Stock Assessment Methods*, *cran.fhcrc.org*.

Ouchi, N. *et al.* (2011) 'Adipokines in inflammation and metabolic disease', *Nature Reviews Immunology*, pp. 85–97. doi: 10.1038/nri2921.

Pan, D. Z. *et al.* (2018) 'Integration of human adipocyte chromosomal interactions with adipose gene expression prioritizes obesity-related genes from GWAS', *Nature Communications*, 9(1), p. 1512. doi: 10.1038/s41467-018-03554-9.

Pazos Obregón, F. *et al.* (2018) 'Cluster Locator, online analysis and visualization of gene clustering', *Bioinformatics*, 34(19), pp. 3377–3379. doi: 10.1093/bioinformatics/bty336.

Perino, M. and Veenstra, G. J. C. (2016) 'Chromatin Control of Developmental Dynamics and Plasticity', *Developmental Cell*, pp. 610–620. doi: 10.1016/j.devcel.2016.08.004.

Petrilli, C. M. *et al.* (2020) 'Factors associated with hospital admission and critical illness among 5279 people with coronavirus disease 2019 in New York City: Prospective cohort study', *The BMJ*, 369, p. m1966. doi: 10.1136/bmj.m1966.

Phipson, B. *et al.* (2016) 'Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression', *Annals of Applied Statistics*, 10(2), pp. 946–963. doi: 10.1214/16-AOAS920.

Purcell, S. *et al.* (2007) 'PLINK: a tool set for whole-genome association and population-based linkage analyses', *The American Journal of Human Genetics*, 81(3), pp. 559–575. doi: 10.1086/519795.

Pyrina, I. *et al.* (2020) 'Fate of Adipose Progenitor Cells in Obesity-Related Chronic Inflammation', *Frontiers in Cell and Developmental Biology*, p. 644. doi: 10.3389/fcell.2020.00644.

Quinlan, A. R. and Hall, I. M. (2010) 'BEDTools: A flexible suite of utilities for comparing genomic features', *Bioinformatics*, 26(6), pp. 841–842.

Rauch, A. *et al.* (2019) 'Osteogenesis depends on commissioning of a network of stem cell transcription factors that act as repressors of adipogenesis', *Nature Genetics*, 51, pp. 716–727. doi: 10.1038/s41588-019-0359-1.

Reilly, S. M. and Saltiel, A. R. (2017) 'Adapting to obesity with adipose tissue inflammation', *Nature*

Reviews Endocrinology, pp. 633–643. doi: 10.1038/nrendo.2017.90.

Ritchie, M. E. *et al.* (2015) 'Limma powers differential expression analyses for RNA-sequencing and microarray studies', *Nucleic Acids Research*, 43(7), p. e47. doi: 10.1093/nar/gkv007.

Ritter, A. *et al.* (2020) 'Obesity and covid-19: Molecular mechanisms linking both pandemics', *International Journal of Molecular Sciences*, p. 5793. doi: 10.3390/ijms21165793.

Roadmap Epigenomics Consortium *et al.* (2015) 'Integrative analysis of 111 reference human epigenomes', *Nature*, 518, pp. 317–330. doi: 10.1038/nature14248.

Rohart, F. *et al.* (2017) 'mixOmics: An R package for 'omics feature selection and multiple data integration', *PLoS Computational Biology*, 13(11), p. e1005752. doi: 10.1371/journal.pcbi.1005752.

Rønningen, T. *et al.* (2015) 'Epigenetic priming of inflammatory response genes by high glucose in adipose progenitor cells', *Biochemical and Biophysical Research Communications*, 467(4), pp. 979–986. doi: 10.1016/j.bbrc.2015.10.030.

Schoenfelder, S. *et al.* (2015) 'The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements', *Genome Research*, 25(4), pp. 582–597. doi: 10.1101/gr.185272.114.

Signorelli, M., Vinciotti, V. and Wit, E. C. (2016) 'NEAT: An efficient network enrichment analysis test', *BMC Bioinformatics*, 17(1), p. 352. doi: 10.1186/s12859-016-1203-6.

Smyth, G. K., Michaud, J. and Scott, H. S. (2005) 'Use of within-array replicate spots for assessing differential expression in microarray experiments', *Bioinformatics*, 21(9), pp. 2067–75. doi: 10.1093/bioinformatics/bti270.

Sorisky, A., Molgat, A. S. D. and Gagnon, A. M. (2013) 'Macrophage-induced adipose tissue dysfunction and the preadipocyte: Should I stay (and differentiate) or should I go?', *Advances in Nutrition*, pp. 67–75. doi: 10.3945/an.112.003020.

van Steensel, B. and Belmont, A. S. (2017) 'Lamina-Associated Domains: Links with Chromosome

Architecture, Heterochromatin, and Gene Repression', *Cell*, pp. 780–791. doi: 10.1016/j.cell.2017.04.022.

Supek, F. *et al.* (2011) 'Revigo summarizes and visualizes long lists of gene ontology terms', *PLoS ONE*, 6(7), p. e21800. doi: 10.1371/journal.pone.0021800.

Whyte, W. A. *et al.* (2013) 'Master transcription factors and mediator establish super-enhancers at key cell identity genes', *Cell*, 153(2), pp. 307–319. doi: 10.1016/j.cell.2013.03.035.

Wingett, S. W. *et al.* (2015) 'HiCUP: pipeline for mapping and processing Hi-C data', *F1000Research*, 4, p. 1310. doi: 10.12688/f1000research.7334.1.

Zhang, Y. *et al.* (2008) 'Model-based Analysis of ChIP-Seq (MACS)', *Genome Biology*, 9(9), p. R137. doi: 10.1186/gb-2008-9-9-r137.

Chapter 5

Conclusions, limitations and future directions

Understanding the genetic etiology and pathophysiology of increased BMI in humans is a challenging but important task toward understanding and reducing risk for CVD. In this dissertation, I have outlined three integrative genomics approaches that incorporated a context-specific assessment of genomic regulatory mechanisms in adipose tissue cell-types that contribute to or respond to BMI in humans. In chapter 2, we showed that adipocyte promoter interactions are enriched for the heritability of adipose gene expression, supporting the utility of using these interactions as a fine-mapping tool for genetic variants that are associated with gene expression. By also identifying obesity GWAS variants that land within these regulatory loops, we highlighted four putative mechanisms that originate in adipocytes and contribute to these obesity traits. These results enabled the initial fine-mapping of obesity GWAS loci and contributed to a better understanding of how adipose tissue cell-types can affect energy homeostasis and fat outcomes in humans.

In chapter 3, we tested the hypothesis that significant GxE associations can be identified after prioritizing candidate variants based on their genomic responses to the environmental stimulus in question. We performed a dietary lipid challenge in human primary adipocytes using saturated or monounsaturated fatty acids, and genetic variants that land in dietary lipid -responsive interacting regions in adipocytes were tested for an interaction with dietary saturated fat intake to affect BMI in the UKB. This resulted in the identification of 14 promoter GxE variants and 24 enhancer GxE variants for BMI.

Finally, in chapter 4 we extended the context-specific chromatin accessibility comparisons to a cohort of MZ twins who are discordant for BMI (ΔBMI$\geq$3 kg/m^2), enabling the assessment of environment (BMI) -driven changes in the epigenome while controlling for the genetic background. We found that increased BMI alters the higher-order chromatin activity

state in the preadipocytes isolated from these twins, and that these reprogrammed regions are significantly enriched for the heritability of CRP in the UKB. Furthermore, these regions contain a higher accumulation of small-effect variant interactions with BMI to affect CRP, suggesting a causal role for BMI-dependent preadipocyte reprogramming in driving the important obesity comorbidity, chronic low-grade inflammation.

In this dissertation work, BMI was used as the obesity measurement phenotype. BMI has been criticized as a measure for obesity due to the inability to estimate the contribution of lean versus fat mass to the resulting BMI value. Instead, waist-to-hip ratio (WHR) has been shown to be a better representation of adiposity in humans, given that it is better correlated with measures of body fat distribution. However, BMI correlates significantly with WHR, suggesting that at the population level, BMI does capture adiposity to a substantial degree (Speakman *et al.*, 2018). Furthermore, BMI is quick and easy to measure, being simply weight (in kilograms) divided by height (in meters) squared. This has supported the collection of very large cohorts with BMI measurements, increasing the power to detect many causal variants, typically with very small effect sizes, for obesity. It has been shown that some of the first hits in cohorts with better defined adiposity phenotypes overlap with BMI associations (Speakman *et al.*, 2018). However, collecting these phenotypes is less straightforward than collecting the measurements for the BMI calculation. The collection procedures often necessitate more training for the collection staff, and they are often more time-consuming, precluding the collection in large enough numbers to reach the power attained by BMI GWAS (Speakman *et al.*, 2018). Furthermore, looking beyond the scientific research, BMI is the measure that can be calculated from data collected during routine clinical procedures: height and weight. Thus, understanding this measure directly, including the caveats and differences in cutoffs between sexes and ethnicities, should be of more clinical

relevance (Dalton *et al.*, 2003). Taken together, despite the phenotypic limitations, understanding BMI is likely beneficial for discovering obesity biology and treating obesity.

These studies started with the characterization of certain epigenetic marks in adipose tissue cell-types *in vitro*. *In vitro* methods that involve the culturing of cells, even if they are primary cells, can be criticized. Some reasons relevant for this dissertation include 1) cells are being cultured in a monolayer on a plate (2-dimensional culture rather than the 3-dimensional organization in tissues); 2) the cell media does not recapitulate the *in vivo* cellular environment; or 3) there is a lack of the *in vivo* interactions of cells with other cell-types in tissues, or lack of signaling within the entire organism. We show that we can integrate *in vitro* genomic findings with human genetic and gene expression cohort studies to extrapolate from the experimental evidence and identify likely genomic regulatory mechanisms underlying human CMDs. By leveraging the environmentally controlled nature of *in vitro* cell culture, we were able to increase the signal-to-noise ratio (through the collection of data from single cell-type rather than tissue) and reduce biological variation (through culturing primary cells briefly to standardize the external environment between individuals) in the functional genomics assessments utilized in this work. This enabled us to identify robust epigenetic signals from the cultured cells that could be validated through large human genetics cohorts with gene expression (METSIM and GTEx) and deep phenotyping (METSIM and UKB), thus providing putative causal genomic regulatory mechanisms underlying obesity and related CMDs.

One major limitation of the work presented here is that it was solely performed in individuals of European ancestry. GWAS have been overwhelmingly performed in individuals of European ancestry. This is detrimental to equity in healthcare as well as understanding the biological underpinnings of disease (Popejoy and Fullerton, 2016; Sirugo, Williams and

Tishkoff, 2019). In the clinic, the generalizability of the results we obtain from European ancestry studies could be limited by differences in effect sizes between ancestries or population-specific variants. This can make PRSs and other genetic risk assessments imprecise or incorrect. Some ancestry-specific mechanisms are likely involved in obesity and related CMDs. Mexicans and Mexican-Americans exhibit the highest risk for T2D (Barquera *et al.*, 2018), whereas Black Americans have a higher risk for hypertension (Ford, 2011). Furthermore, different genetic backgrounds likely shape the genomic regulatory landscape, from promoter interactions and the higher-order organization of the chromatin, to the local genetic regulation by enhancers and other regulatory elements. While the genetic effects on the 3-dimensional conformation of the chromatin is less understood, the existence of chromatin accessibility and histone mark quantitative trait loci exist has been established (Kasowski *et al.*, 2013; Kilpinen *et al.*, 2013; McVicker *et al.*, 2013; Gate *et al.*, 2018). Thus, it is important to place an emphasis on collecting large enough GWAS and functional genomics cohorts for obesity and related comorbidities in more diverse populations in the coming years.

For all of the studies presented in this dissertation, the goal was to identify strong candidates for the functional assessment of genes that drive obesity and related CMDs in adipose tissue cell-types. One unanticipated result from the GWAS era was the highly polygenic nature of some complex traits, including BMI. We learned that it is not feasible to test all potential causal genes and variants in a single GWAS locus, let alone tens to hundreds. Thus, the genome-wide assessment of functional genomics for the initial GWAS fine-mapping have proven to be an important approach in the post-GWAS era. While many disease associations may exhibit a functional enrichment in certain cell- or tissue-types, in reality, a wide range of cell-types can be important for disease. Functional genomics provides a way for researchers to prioritize cell-types

and fine-map mechanisms that are relevant for their expertise. This can refine a starting list of many GWAS loci to a handful that have strong genomic evidence of being involved in their cell-type or context of interest. Integrative functional genomics approaches are highly flexible and depend on the priority of the research. For instance, rather than focusing on cell- or context-specific mechanisms, projects could instead focus on a cross cell-type fine-mapping analysis looking for evidence that known drug targets underlie biological mechanisms at GWAS loci. The vast amount of information available to the public, with much more to come, is enabling creative and meaningful research aims that have provided valuable insight into complex trait biology and genetics.

Next steps following the work presented in this dissertation could involve functional probing at the loci that exhibit evidence of being involved in obesity mechanisms originating in adipose tissue cell-types. In chapter 2, our integrative approach identified putative genomic regulatory mechanisms that underlie four obesity GWAS loci. For genes with limited information regarding their function, gene knockdown combined with RNA-seq as well as a number of other functional genomics assays, such as ATAC-seq or TF and histone mark ChIP-seq, could be performed to get a sense of the pathways the gene is involved in. Experimental designs that incorporate more high-throughput, functional genomics assessments such as these are becoming more widely used as these assays become more affordable due to kits and reduced sequencing costs. By obtaining a network-level view of gene perturbation effects across many individuals, for example, we can more reliably identify of the pathway(s) in which the gene is involved in that cell-type. In addition, approaches such as this can be performed on many genes at once by profiling the effects of gene knockdown across many cells, rather than individuals. This is the general idea behind Perturb-seq (Dixit *et al.*, 2016), which combines a high-

throughput activating or inactivating CRISPR assay with single-cell RNA-seq readouts to assess the effects of gene perturbations across many cells for many genes in one experiment. This would be highly valuable for following up projects with many hits that cannot be screened on a locus-by-locus basis, such as the GxEs identified in chapters 3 and 4.

Understanding both the causes of obesity and its downstream mechanisms that contribute to the increased prevalence of other CMDs will be important for reducing cardiometabolic risk. In this dissertation, I have outlined approaches using cell- and context-specific functional genomics assays to highlight putative genomic regulatory mechanisms underlying genetic associations and GxEs for CMDs in humans. GWASs have highlighted many loci to be involved in complex traits, which necessitates more systematic fine-mapping approaches than previously anticipated. Furthermore, understanding the mechanisms through which the obesogenic environment interacts with our genetics to drive variation in BMI and related comorbidities can be valuable for developing novel therapeutics. This will aid the development of personalized prevention and treatment plans to counteract the continued rise in the prevalence of obesity in the United States and worldwide.

References

Barquera, Simón *et al.* (2018) 'Collaborative research and actions on both sides of the US-Mexico border to counteract type 2 diabetes in people of Mexican origin', *Globalization and Health*, 14(1), p. 84. doi: 10.1186/s12992-018-0390-5.

Dalton, M. *et al.* (2003) 'Waist circumference, waist-hip ratio and body mass index and their correlation with cardiovascular disease risk factors in Australian adults', *Journal of Internal Medicine*, pp. 555–563. doi: 10.1111/j.1365-2796.2003.01229.x.

Dixit, A. *et al.* (2016) 'Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens', *Cell*, 167(7), pp. 1853–1866. doi: 10.1016/j.cell.2016.11.038.

Ford, E. S. (2011) 'Trends in mortality from all causes and cardiovascular disease among hypertensive and nonhypertensive adults in the united states', *Circulation*, 123(16), pp. 1737–1744. doi: 10.1161/CIRCULATIONAHA.110.005645.

Gate, R. E. *et al.* (2018) 'Genetic determinants of co-accessible chromatin regions in activated T cells across humans', *Nature Genetics*, 50, pp. 1140–1150. doi: 10.1038/s41588-018-0156-2.

Kasowski, M. *et al.* (2013) 'Extensive variation in chromatin states across humans', *Science*, 342(6159), pp. 750–752. doi: 10.1126/science.1242510.

Kilpinen, H. *et al.* (2013) 'Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription', *Science*, 342(6159), pp. 744–747. doi: 10.1126/science.1242463.

McVicker, G. *et al.* (2013) 'Identification of genetic variants that affect histone modifications in human cells', *Science*, 342(6159), pp. 747–749. doi: 10.1126/science.1242429.

Popejoy, A. B. and Fullerton, S. M. (2016) 'Genomics is failing on diversity', *Nature*, pp. 161–164. doi: 10.1038/538161a.

Sirugo, G., Williams, S. M. and Tishkoff, S. A. (2019) 'The Missing Diversity in Human Genetic Studies', *Cell*, 177(1), pp. 26–31. doi: 10.1016/j.cell.2019.02.048.

Speakman, J. R. *et al.* (2018) 'GWAS for BMI: a treasure trove of fundamental insights into the genetic basis of obesity', *International Journal of Obesity*, pp. 1524–1531. doi: 10.1038/s41366-018-0147-5.