Title
Integrative genomic approaches to understand human disease mechanisms: applications to cardiometabolic traits

Author
Ko, Arthur

Publication Date
2018

UNIVERSITY OF CALIFORNIA

Los Angeles

Integrative genomic approaches to understand human disease mechanisms: applications to cardiometabolic traits

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Molecular Biology

by

Arthur Ko

2018

ABSTRACT OF THE DISSERTATION

Integrative genomic approaches to understand human disease mechanisms: applications to cardiometabolic traits

by

Arthur Ko

Doctor of Philosophy in Molecular Biology

University of California, Los Angeles, 2018

Professor Päivi Elisabeth Pajukanta, Chair

With more efficient genotyping technologies and lower sequencing cost, genome-wide association studies (GWAS) have been broadly applied to many complex human traits. However, people of European descent remain the most prominent subjects in genetic research and other ethnic groups might not fully benefit from the effort of GWAS. In addition to expanding GWAS to include more diverse populations, new approaches that enable trans-ethnic or multi-ethnic analyses in GWAS will also be a crucial stepping stone for future genetic studies. To address this disparity and knowledge gap, we developed and applied a new approach, cross-population allele screen (CPAS) prior to GWAS, to identify population-specific variants that are associated with complex traits or diseases (Chapter 2). In our study, we identified novel genetic variants that are associated with serum triglycerides (TGs), high-density lipoprotein cholesterol (HDL-C), and

body mass index (BMI), exhibiting differential allele frequencies between Finns and Mexicans. Notably, one of the novel TGs-associated genes, SIK family kinase 3 (*SIK3*), harbors an Amerindian-specific common risk variant (allele frequency=18% in Mexicans), which is not observed in other continental populations, and the risk allele carriers also exhibit higher serum TG levels after a high-fat meal. In addition, this locus displays a signal of positive selection in Mexicans, suggesting that a delayed serum lipid clearance might have been evolutionally advantageous for ancient Amerindian people.

While GWAS have uncovered many trait-associated loci, translating GWAS results to actionable medical information remains nontrivial due to the difficulty of pinpointing the true causal variants and genes. To understand the molecular mechanism of GWAS variants, many functional genomic approaches have been developed. In this dissertation, I will present two computational methods to integrate genetic and transcriptomic data to infer functional variants and possible underlying genes. First, we developed Functional Summary-based Imputation (FUSION) that can leverage GWAS summary statistics and a relatively small reference panel of transcriptomes to infer the association between gene expression and traits (Chapter 3). Using FUSION as well as subcutaneous adipose and whole blood RNA-sequence (RNA-seq) data, we performed transcript-wide associated studies (TWAS) and identified 69 novel genes associated with BMI, serum lipids, and height. With the constantly growing GWAS summary statistics and transcriptomic data, we can further utilize FUSION to apply TWAS to many different traits and tissues.

To account for the increasing presence of large-scale RNA-seq cohorts, we created a new computational tool, ASElux, which can efficiently perform allele-specific expression (ASE) estimation that was previously prohibited due to excessive computing time (Chapter 4). We

implemented a hybrid index system in ASElux to first build an individualized reference genome with available genotype data, and ASElux will then only align variant-carrying reads that are informative for ASE calculation. Thus, ASElux can correct for the reference allele bias during alignment with much shorter computing time. In our comparison test, ASElux is 4-33 times faster than other commonly used software or pipelines for ASE and obtain a similar or better accuracy. We applied ASElux to 273 lung RNA-seq samples, and uncovered a splice variant, rs11078928, which could explain the molecular mechanism of an asthma GWAS hit, rs11078927. We envision that the speed and efficiency of ASElux can facilitate ASE analysis in many RNA-seq datasets to uncover functional variants in the future.

In Chapters 5 and 6, I will present our studies utilizing epigenomic and transcriptomic data to gain insight into the causal mechanisms of obesity and non-alcoholic fatty liver disease (NAFLD). To elucidate molecular mechanisms underlying obesity-related GWAS variants, we integrated promoter-enhancer interactions in human primary adipocytes with adipose *cis* expression quantitative trait locus (eQTL) variants (Chapter 5). Using promoter capture Hi-C, we first assayed chromosomal interactions in human primary adipocytes. In combination with human subcutaneous adipose transcriptomes, we then identified four genes associated with BMI or obesity-related traits that are also under *cis* regulation via chromosomal looping. We further performed electrophoretic mobility shift assays (EMSAs) to validate the allelic effect of a *cis* eQTL, rs4776984, regulating mitogen-activated protein kinase 5 (*MAP2K5)*. The reference allele displayed a lower protein binding affinity than the alternative allele, in line with the computationally predicted disruptive effect. Finally, we also reported 38 additional BMI candidate genes under the regulation of chromosomal interactions for future studies of obesity.

In our NAFLD study (Chapter 6), we tested the hypothesis that obesity may impair the function of adipose tissue, which can lead to ectopic fat accumulation in the liver, resulting in NAFLD. To understand the molecular pathogenesis of NAFLD driven by obesity, we examined the liver histology and subcutaneous adipose transcriptomes from 259 morbidly obese Finnish individuals that underwent a bariatric surgery. One year after the surgery, we re-profiled their adipose transcriptomes to assess the effect of the weight loss on adipose gene expression. At baseline, we identified adipose expression of 43 genes downregulated in non-alcoholic steatohepatitis (NASH) patients. Of these, the adipose expression of 17 genes was negatively correlated with liver steatosis and serum TGs. In a large panel of mouse strains, expression of five of the 17 genes was also correlated with a diet-induced liver steatosis. Specifically, the adipose expression of one of the five genes, death associated protein kinase 2 (*DAPK2*), recovered after the weight-loss at the one-year follow-up. Combining phenotype and longitudinal transcriptome data, we performed mediation analyses to demonstrate the causal effect of *DAPK2* adipose expression on NAFLD. When *DAPK2* expression was knocked down in human primary preadipocytes, five key genes involved in autophagy, of which two also function in adipocyte differentiation, were also downregulated. Our findings suggest an obesity-induced reduction of *DAPK2* expression as a new pathogenic mechanism of NAFLD through impairment of autophagy pathway and adipocyte differentiation. In summary, our work presented in Chapters 5 and 6, employing functional genomic approaches and computational methods to decipher disease mechanisms of obesity and NAFLD, highlights strategies to understand the molecular pathogenesis of human disease beyond GWAS.

The dissertation of Arthur Ko is approved.

Eleazar Eskin

Leonid Kruglyak

Stephen G Young

Yibin Wang

Päivi Elisabeth Pajukanta, Committee Chair

University of California, Los Angeles

2018

DEDICATION

I dedicate this work to my parents, my brother, and my wife for their unconditional support and always pushing me forward.

Table of Contents

List of figures

Chapter 1

Chapter 6

List of supplementary materials

Chapter 6

The following items are available as external Excel files.

Chapter 6

Supplementary Table 2. The associations between the liver co-expression module eigen-genes and traits from the weighted gene co-expression network analysis (WGCNA).

Supplementary Table 3. The associations between the adipose co-expression module eigen-genes and traits from the weighted gene co-expression network analysis (WGCNA).

Acknowledgements

I would like to thank my advisor, Dr. Paivi Pajukanta, for her support throughout my graduate career as well as all members of the Pajukanta lab for their help and effort to make all work presented here possible. My most sincere gratitude goes to Drs. Teresa Tusie-Luna, Carlos Aguilar-Salinas, Markku Laakso, and Jussi Pihlajamäki for their insight and collaboration. I wish to thank Drs. Rita Cantor and Janet Sinsheimer for their statistical guidance and career support that help me develop as a scientist. I would also like to thank Dr. Jake Lusis for his kind advice and encouragement. Finally, I would like to thank the funding support from the Genomic Analysis Training Program (GATP, T32HG002536) and NIH NRSA Predoctoral Fellowship F31 (F31HL127921).

Chapter 2 is a reprint of "Amerindian-specific regions under positive selection harbour new lipid variants in Latinos" by Ko A, Cantor RM, Weissglas-Volkov D, Nikkola E, Reddy PMVL, Sinsheimer JS, Pasaniuc B, Brown R, Alvarez M, Rodriguez A, Rodríguez-Guillén R, Bautista IC, Arellano-Campos O, Muñoz-Hernández LL, Salomaa V, Kaprio J, Jula A, Jauhiainen M, Heliövaara M, Raitakari O, Lehtimäki T, Eriksson JG, Perola M, Lohmueller KE, Matikainen N, Taskinen MR, Rodriguez-Torres M, Riba L, Tusie-Luna T, Aguilar-Salinas CA, Pajukanta P. *Nature Communications*; 2014;5:3983 and appears under the Creative Commons License.

Chapter 3 is a reprint of "Integrative approaches for large-scale transcriptome-wide association studies" by Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BWJH, Jansen R, de Geus EJC, Boomsma DI, Wright FA, Sullivan PF, Nikkola E, Alvarez M, Civelek M, Lusis AJ, Lehtimäki T, Raitoharju E, Kähönen M, Seppälä I, Raitakari OT, Kuusisto J, Laakso M,

Price AL, Pajukanta P, Pasaniuc B. *Nature Genetics*; 2016;48:245-252 appears with my author copyright for reprint and reuse.

Chapter 4 is a reprint of "ASElux: an ultra-fast and accurate allelic read counter" by Miao Z, Alvarez M, Pajukanta P, and Ko A. Bioinformatics; 2018;34(8):1313–20 and appears with the permission of Oxford University Press.

Chapter 5 is reprint "Integration of human adipocyte chromosomal interactions with adipose gene expression prioritizes obesity-related genes from GWAS" by Pan DZ, Garske KM, Alvarez M, Bhagat YV, Boocock J, Nikkola E, Miao Z, Raulerson CK, Cantor RM, Civelek M, Glastonbury CA, Small KS, Boehnke M, Lusis AJ, Sinsheimer JS, Mohlke KL, Laakso M, Pajukanta P, Ko A. *Nature Communications*; 2018;9:1512. and appears under the Creative Commons License.

Chapter 6 is a version of the article in submission "Obesity-induced reduction of Death Associated Protein Kinase 2 predisposes to non-alcoholic fatty liver disease" by Ko A, Benhammou JN, Garske K, Kaminska D, Nikkola E, Pisegna JR, Hui ST, Lusis AJ, Männistö V, Kärjä V, Cantor RM, Sinsheimer JS, Mohlke KL, Laakso M, Pihlajamäki J, Pajukanta P. This study is supported by National Institutes of Health (NIH) grants HL-095056, HL-28481, U01DK105561, CURE:DDRC NIH P30DK41301 and by grants from the Academy of Finland (contract numbers 120979 and 138006), the Finnish Diabetes Research Foundation, the Finnish Cultural Foundation and Northern Savo Regional Fund, and the Kuopio University Hospital State Research Funding (EVO and VTR). Ko A and Garske K. were supported by NIH grants F31HL127921 and F31HL142180, respectively. Benhammou JN was supported by an NIH Training Grant (DK007180). Kaminska D was supported by the Finnish Cultural Foundation and the Foundations' Post Doc Pool. Contribution: Study design: Ko A, Pihlajamäki J, and

Pajukanta P. Methods development and statistical analysis: Ko A, Cantor RM, and Sinsheimer JS. Computational analysis: Ko A. Functional analysis: Benhammou JN, Garske K, Pisegna JR, Ko A, and Pajukanta P. Phenotyping: Kärjä V, Männistö V, Kaminska D, Lusis AJ, Hui ST, Laakso M, Benhammou JN, Pisegna JR, and Pihlajamäki J. Data collection and GWAS genotyping: Pihlajamäki J, Kärjä V, Männistö V, Kaminska D, Laakso M, Nikkola E, Mohlke KL, Lusis AJ, Hui ST, Ko A, and Pajukanta P. Manuscript: Ko A and Pajukanta P wrote the manuscript and all authors read, reviewed, and/or edited the manuscript.

VITA

University of Waterloo	B.C.S.	2006-2010
Waterloo, Ontario, Canada	Computer Science-Bioinformatics	

Computational Instructor

UCLA	Human Genetics 236A: Advanced Human Genetics	2015/09-2015/11

Teaching Assistantships

UCLA	Life Science 4: Genetics	2015/01-2015/04
UCLA	MCDB 187AL: Research Immersion Laboratory in Genomic Biology	2014/01-2014/03
University of Waterloo	Math 235: Linear Algebra 2	2009/01-2009/04
University of Waterloo	Math 239: Combinatorics	2008/09-2008/12

PUBLICATIONS

*** Indicates corresponding authorship and **co-corresponding authorship**

1. **Ko A****, Kaminska D, Nikkola E, Benhammou JN, Pisegna JR, Hui ST, Lusis AJ, Männistö V, Kärjä V, Cantor RM, Sinsheimer JS, Mohlke KL, Laakso M, Pihlajamäki J, Pajukanta P. Obesity-induced reduction of Death Associated Protein Kinase 2 predisposes to non-alcoholic fatty disease. Submitted 2018.
2. Lea A, Subramaniam M, **Ko A**, Lehtimäki T, Raitoharju E, Kähönen M, Seppälä I, Mononen N, Raitakari TO, Ala-Korpela M, Pajukanta P, Zaitlen NA, Ayroles JF. Genetic and environmental effects disrupt molecular co-regulation. Submitted 2018.
3. Pan DZ, Garske KM, Alvarez M, Bhagat YV, Boocock J, Nikkola E, Raulerson CK, Cantor RM, Civeleck M, Glastonbury CA, Small KS, Boenke M, Lusis AJ, Sinsheimer JS, Mohlke KL, Laakso M, Pajukanta P, **Ko A***. Integration of human adipocyte chromosomal interactions with local adipose gene expression identifies obesity genes beyond GWAS. ***Nature Communications***; 2018;9:1512.
4. Miao Z, Alvarez M, Pajukanta P, **Ko A***. ASElux: an ultra-fast and accurate allelic reads counter. ***Bioinformatics***; 2018;34(8):1313-20.
5. Nikkola E, **Ko A**, Alvarez M, Cantor MR, Garske K, Kim E, Gee S, Rodriguez A, Muxel R, Matikainen N, Soderlund S, Motazacker MM, Boren J, Lamina C, Kronenberg F, Schneider WJ, Palotie A, Laakso M, Taskinen MR, Pajukanta P. Family-specific aggregation of lipid GWAS variants confers the susceptibility to familial hypercholesterolemia in a large Austrian family. ***Atherosclerosis***; 2017; 264:58-66.
6. Civelek M, Wu Y, Pan C, Raulerson CK, **Ko A**, He A, Tilford C, Saleem NK, Stančáková A, Scott LJ, Fuchsberger C, Stringham HM, Jackson AU, Narisu N, Chines PS, Small KS, Kuusisto J, Parks BW, Pajukanta P, Kirchgessner T, Collins FS, Gargalovic PS, Boehnke M, Laakso M, Mohlke KL, Lusis AJ. Genetic regulation of adipose gene expression and cardio-metabolic traits. ***The American Journal of Human Genetics***; 2017;100:428-443.
7. Rodriguez A, Gonzalez L, **Ko A**, Alvarez M, Miao Z, Bhagat YV, Nikkola E, Cruz-Bautista I, Arellano-Campos O, Muñoz-Hernández LL, Ordoñez-Sánchez ML, Rodríguez-Guillén R, Mohlke KL, Laakso M, Tusie-Luna T, Aguilar-Salinas CA, Pajukanta P. Molecular characterization of the lipid genome-wide association study signal on chromosome 18q11.2 implicates HNF4A-mediated regulation of the TMEM241 gene. ***Arteriosclerosis Thrombosis and Vascular Biology***; 2016;48:245-252.

8. Gusev A, **Ko A**, Shi H, Bhatia G, Chung W, Penninx BWJH, Jansen R, de Geus EJC, Boomsma DI, Wright FA, Sullivan PF, Nikkola E, Alvarez M, Civelek M, Lusis AJ, Lehtimäki T, Raitoharju E, Kähönen M, Seppälä I, Raitakari OT, Kuusisto J, Laakso M, Price AL, Pajukanta P, Pasaniuc B. Integrative approaches for large-scale transcriptome-wide association studies. ***Nature Genetics***; 2016;48:245-252.

9. Nikkola E, Laiwalla A, **Ko A**, Alvarez M, Connolly M, Ooi YC, Hsu W, Bui A, Pajukanta P, Gonzalez NR. Remote Ischemic Conditioning Alters Methylation and Expression of Cell Cycle Genes in Aneurysmal Subarachnoid Hemorrhage. ***Stroke***; 2015;46:2445-2451.

10. **Ko A**, Cantor RM, Weissglas-Volkov D, Nikkola E, Reddy PMVL, Sinsheimer JS, Pasaniuc B, Brown R, Alvarez M, Rodriguez A, Rodríguez-Guillén R, Bautista IC, Arellano-Campos O, Muñoz-Hernández LL, Salomaa V, Kaprio J, Jula A, Jauhiainen M, Heliövaara M, Raitakari O, Lehtimäki T, Eriksson JG, Perola M, Lohmueller KE, Matikainen N, Taskinen MR, Rodriguez-Torres M, Riba L, Tusie-Luna T, Aguilar-Salinas CA, Pajukanta P. Amerindian-specific regions under positive selection harbour new lipid variants in Latinos. ***Nature Communications***; 2014;5:3983.

11. Campbell CD, Mohajeri K, Malig M, Hormozdiari F, Nelson B, Du G, Patterson KM, Eng C, Torgerson DG, Hu D, Herman C, Chong JX, **Ko A**, O'Roak BJ, Krumm N, Vives L, Lee C, Roth LA, Rodriguez-Cintron W, Rodriguez-Santana J, Brigino-Buenaventura E, Davis A, Meade K, LeNoir MA, Thyne S, Jackson DJ, Gern JE, Lemanske RF, Shendure J, Abney M, Burchard EG, Ober C, Eichler EE. Whole-genome sequencing of individuals from a founder population identifies candidate genes for asthma. ***PLoS ONE***; 2014;9(8):e104396.

12. Hormozdiari F, Konkel MK, Prado-Martinez J, Chiatante G, Herraez IH, Walker JA, Nelson B, Alkan C, Sudmant PH, Huddleston J, Catacchio CR, **Ko A**, Malig M, Baker C, Genome Project GA, Marques-Bonet T, Ventura M, Batzer MA, Eichler EE. Rates and patterns of great ape retrotransposition. ***Proc. Natl. Acad. Sci. U.S.A***.; 2013;110:13457-13462.

13. Kiezun A, Pulit SL, Francioli LC, van Dijk F, Swertz M, Boomsma DI, van Duijn CM, Slagboom PE, van Ommen GJB, Wijmenga C, de Bakker PIW, Sunyaev SR, Genome of the Netherlands Consortium. Deleterious alleles in the human genome Are on average younger than neutral alleles of the same frequency. ***PLoS Genetics***; 2013;9(2):e1003301.

14. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. ***Nature***; 2012;491:56-65.

15. Campbell CD, Chong JX, Malig M, **Ko, A**, Dumont BL, Han, L, Vives L, O'Roak BJ, Sudmant PH, Shendure J, Abney M, Ober C, Eichler EE. Estimating the human mutation rate using autozygosity in a founder population. ***Nature Genetics***; 2012;44:1277-1281.

16. Krumm N, Sudmant PH, **Ko A**, O'Roak BJ, Malig M, Coe BP, NHLBI Exome Sequencing Project, Quinlan AR, Nickerson DA, Eichler EE. Copy number variation detection and genotyping from exome sequence data. ***Genome Research***; 2012;2:1525-1532.

17. O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP, Levy R, **Ko A**, Lee C, Smith JD, Turner EH, Stanaway IB, Vernot B, Malig M, Baker C, Reilly B, Akey JM, Borenstein E, Rieder MJ, Nickerson DA, Bernier R, Shendure J, Eichler EE. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. ***Nature***; 2012;485:246-250.

Chapter 1

Introduction

For the past 15 years, genome-wide association studies (GWAS) have successfully uncovered hundreds of genomic loci associated with complex human traits and diseases[1,2]. However, the focus of GWAS has been heavily skewed toward European populations, and the landscape of genetic studies does not well represent the diversity of global populations. In particular, a meta-analysis of the GWAS Catalog in 2016 reported that only 20% of studies were performed on non-Europeans (Figure 1)[3]. This misrepresentation could lead to a serious medical disparity, as the current drive of precision medicine will only benefit a few privileged populations. For example, a common application of GWAS findings is to calculate polygenic risk score (PRS) as a genetic predictor for an individual[4]. When PRS is derived from summary statistics estimated from a different population, the difference in linkage disequilibrium (LD) pattern or the underlying causal variants might lead to a poor prediction[5]. Indeed, several studies have shown that when applying GWAS summary statistics from European cohorts to other populations for multiple traits and diseases, PRS can be directionally inconsistent or have a low accuracy[6-8]. In addition, some risk variants might be exclusive to specific populations or display varying effect sizes in different populations[5]. Therefore, future GWAS and sequencing endeavors should encompass more diverse ethnic groups. Furthermore, there is also a pressing need to develop better methods that can translate or leverage results across populations. To reduce this knowledge gap, we developed and applied a new approach that specifically maps trait-associated variants that exhibit a differential allele frequency between populations to identify several Mexican-specific variants associated with metabolic traits[9]. This work will be further discussed in Chapter 2.

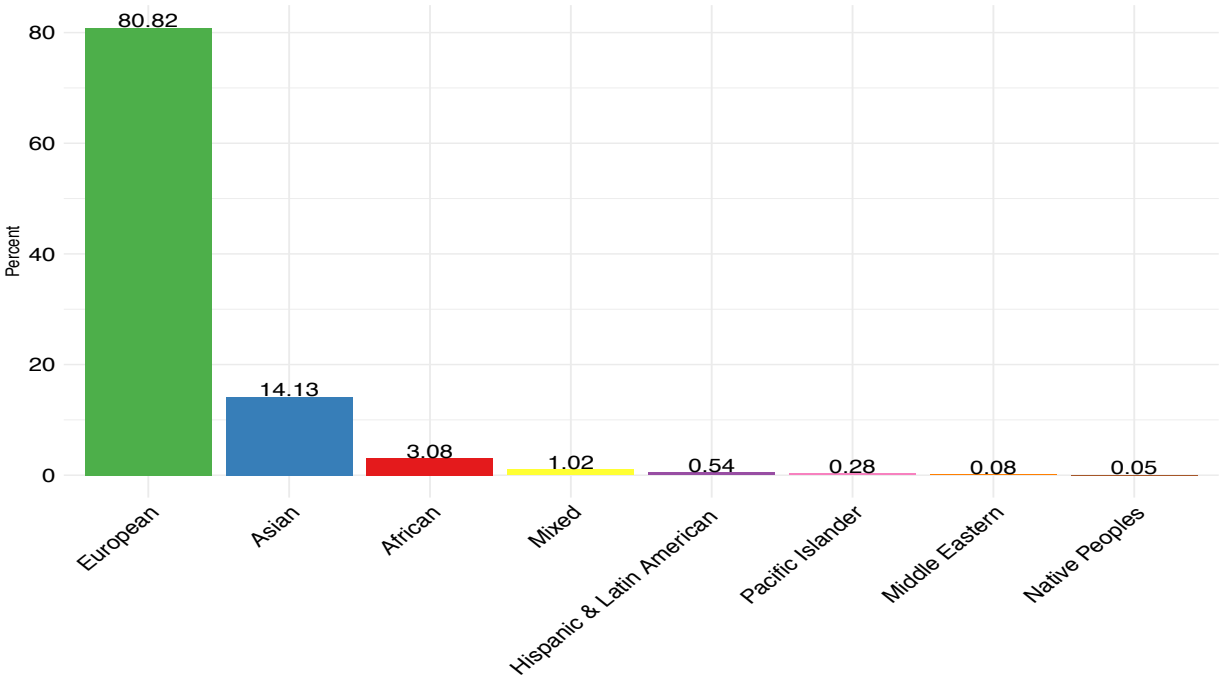


Figure 1. A meta-analysis of the GWAS Catalog in 2016. More than 80% of published GWAS still focus on European populations, demonstrating the potential disparity in precision medicine. The percent of each population represented in the GWAS Catalog is shown on top of each bar. The Native Peoples include indigenous populations, such as Native Americans and Australian Aboriginals. Adapted from Popejoy et al.[3].

In Chapter 2, we identified population-specific variants in Latinos that are associated with metabolic traits[9]. Dyslipidemia and obesity are especially prevalent in populations with Amerindian backgrounds[10], such as Mexican-Americans, predisposing these populations to cardiovascular disease. We designed an approach, known as the cross-population allele screen (CPAS), which restricts GWAS to only variants that differ in allele frequency between two populations. We performed CPAS prior to a GWAS in 19,273 Europeans and Mexicans, to first identify population-stratified loci between Europeans and Mexicans and then performed GWAS to pinpoint Amerindian risk loci associated with lipids and BMI. Using CPAS-GWAS, we uncovered novel Amerindian lipid genes, receptor-related orphan receptor alpha (*RORA*) and salt-inducible kinase 3 (*SIK3*) as well as three loci previously unassociated with dyslipidemia or obesity. We also observed that the lipoprotein lipase (*LPL*) and apolipoprotein A5 (*APOA5*)

genes harbor specific Amerindian signatures of risk variants and haplotypes. Notably, we found that *SIK3* and one novel lipid locus underwent positive selection in Mexicans. Furthermore, after a high-fat meal, the *SIK3* risk variant carriers displayed higher serum triglyceride (TG) levels than the non-carriers. These findings suggest that Amerindian-specific genetic architecture leads to a higher incidence of dyslipidemia and obesity in modern Mexicans. Furthermore, CPAS should facilitate future GWAS studies in diverse populations.

Other challenges of GWAS include identifying the true causal variants and genes at the associated loci as well as transforming the association results into biological mechanisms[1]. Since genotyping arrays are often designed with single-nucleotide polymorphisms (SNPs) that tag as many variants as possible via LD, disentangling the underlying correlations and finding the true functional variants pose a major obstacle[11,12]. Several strategies that apply functional genomic approaches to fine map causal genetic variants and genes at the GWAS loci have emerged[13] (Figure 2). For example, one strategy is to determine whether a variant impacts the expression of a nearby gene, and thus infer the potential causal gene[14]. These regulatory variants can be identified either by mapping them as expression quantitative loci (eQTLs) or by quantifying allele-specific expression (ASE), tagged by the GWAS variants[15]. Another approach is to overlap a GWAS locus with epigenomic information, such as chromatin interactions, transcription factor binding, or methylation that can be assayed with a plethora of experimental methods[16]. The establishments of many large transcriptomic and epigenomic reference panels and datasets from as ENCODE[17], FANTOM[18,19], and GTEx[20], further empower fine-mapping efforts by providing public resources and annotations for many human tissues. These functional genomic strategies have been fruitful in elucidating the underlying mechanisms of some type 2 diabetes (T2D) GWAS variants. For instance, a study discovered 14 known T2D loci that co-localize with islet

cis-eQTL signals and active chromatin regions, and when the expression of one of the candidate genes, *ZMIZ1*, is increased in islet and beta cells, insulin secretion is impaired[21]. We have also utilized ASE and promoter capture chromatin conformation capture (capture Hi-C)[22] to pinpoint causal variants at asthma[23] and obesity[24] GWAS loci, respectively. In this dissertation, I will present our work leveraging or integrating multiple genomic, transcriptomic, and epigenomic data to disentangle the underlying molecular mechanisms of GWAS variants.

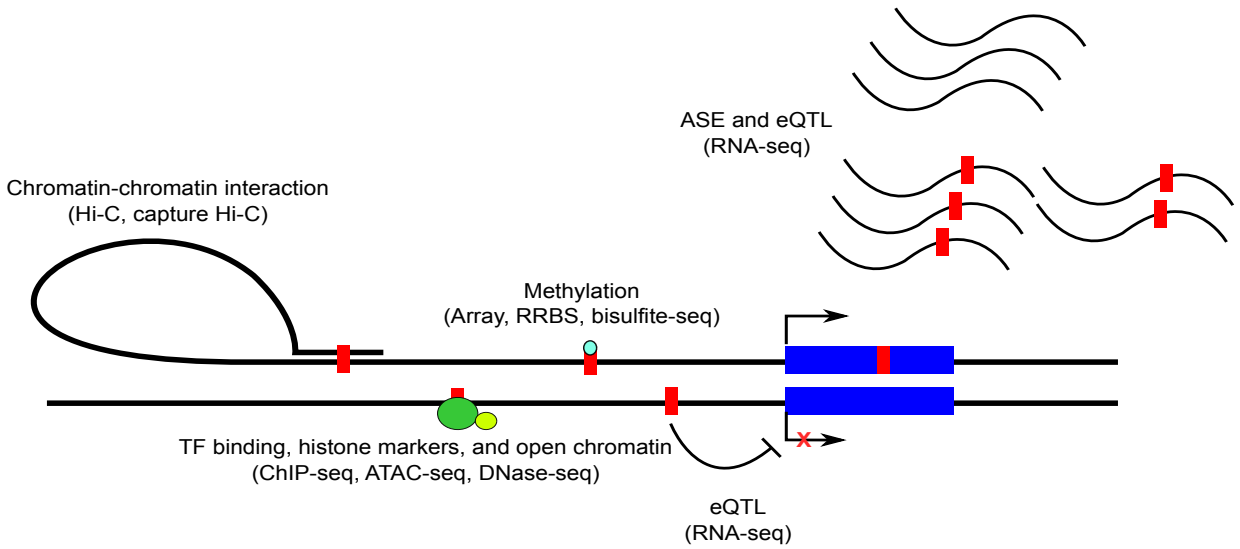


Figure 2. Functional genomic approaches to delineate the molecular mechanisms of GWAS variants. Red bars indicate causal GWAS variants and the gene body is shown as the blue boxes. Expression quantitative trait loci (eQTL) and allele-specific expression (ASE) can help infer the causal gene, regulated by the GWAS variant or its LD proxy. Epigenomic information, such as chromosome conformation, methylation, and transcription factor (TF) binding sites, can further help fine-map the causal variant. Experimental methods to assay these molecular phenotypes are given in parenthesis.

Transcriptome-wide association studies (TWAS) is a powerful approach that tests the association between phenotypes and the expression of all genes or transcripts to infer the underlying causal genes of human complex traits and disease[25,26]. However, TWAS are often limited by the sample sizes due to the high cost of whole-transcriptome profiling, typically by RNA-seq. In Chapter 3, I will discuss our newly developed method, Functional Summary-based

Imputation (FUSION), which predicts the association between gene expression and traits using only summary statistics[25]. FUSION utilizes a relatively small reference panel of gene expression profiles and proximal genetic variants of each gene to impute expression into much larger cohorts. Imputed expression is then correlated with traits based on summary statistics from GWAS. We applied FUSION to ~3,000 transcriptomes and identified 69 new genes associated with BMI, lipid, and height, demonstrating FUSION as an effective TWAS approach without direct measurement of gene expression. The availability of a suitable population-based transcriptomic reference panel may limit the application of FUSION, but we expect that data from large transcriptomic cohorts, such as GTEx[20], will help circumvent this issue. Furthermore, FUSION is highly flexible and can be extended to association studies of other molecular phenotypes.

As the amount of RNA-seq data grows rapidly, efficient computational tools are required to fulfill the analytical needs. In Chapter 4, I will present our novel computational tool, ASElux[23], which can quickly and accurately calculate ASE using short-read data in large RNA-seq data sets. ASE analysis in large-scale studies has been obstructed by the reference allele bias[27] and the slow speed of the variant-aware alignment[28,29]. ASElux utilizes a hybrid index system to build an individualized reference genome to correct for the reference bias. Subsequently, ASElux achieves its high speed by only aligning reads that carry a SNP that is informative for ASE calculation. We benchmarked ASElux against other popular tools, such as the combination of STAR[30] and WASP[29] for read alignment and bias correction, and the SNP-aware aligner, GSNAP[28]. Overall, ASElux is 4-33X faster while obtaining a better or comparable accuracy. We applied ASElux to 273 lung RNA-seq samples from GTEx[20] and identified a splice-QTL, rs11078928, which revealed the underlying mechanism of an asthma GWAS hit[31]. Taken

together, our new method, ASElux, should greatly advance the field of transcriptomics and facilitate future ASE investigations in extensive RNA-seq cohorts without compromising the accuracy.

Increased adiposity is a hallmark of obesity and overweight, which affect 2.2 billion people world-wide. Understanding the genetic and molecular mechanisms in adipose tissue underlying obesity-related phenotypes can improve treatment options and help develop drugs. In Chapter 5, we examined the association between body mass index (BMI) and gene expression, regulated by local enhancers via chromosomal interactions in primary human white adipocytes, a cell type highly relevant for obesity[24]. We utilized the promoter Capture-Hi-C[22] method, which detects physical loops between gene promoters and other genomic regions, to construct a 3D interaction map of the genome. These chromosomal interactions helped us identify the regional functional variant and its target gene for 4 obesogenic GWAS loci, including mitogen-activated protein kinase 5 (*MAP2K5*)[32]. Furthermore, we uncovered 38 replicated non-GWAS genes as new candidates for future studies of obesity. Promoter-interacting elements in human adipocytes are substantially enriched for adipose-related transcription factor motifs, such as PPARG and CEBPB, and contribute a significant amount of heritability to the gene expression, indicating that chromosomal interactions are important for local transcriptional regulation of adipocytes. Overall, our study drives the functional understanding of 4 metabolic GWAS loci forward and identifies 42 replicated genes involved in obesity.

The global obesity epidemic has led to the rising prevalence of non-alcoholic fatty liver disease (NAFLD), which affects almost 25% of adults worldwide[33]. Without intervention or treatment, NAFLD can escalate to cirrhosis or hepatocellular carcinoma, which are irreversible and often require a liver transplant[34]. However, our understanding of the underlying mechanisms

of NAFLD remains elusive, hampering the development of treatment. GWAS and an exome-wide scan have recently identified NAFLD-associated variants within patatin-like phospholipase domain-containing 3 protein (*PNPLA3*)[35] and transmembrane 6 superfamily member 2 (*TM6SF2*)[36]; however their mechanisms are not well-understood yet. In Chapter 6, I will present our work to elucidate the largely unknown molecular pathogenesis of NAFLD using human liver and subcutaneous adipose tissues, statistical causal inference, mouse model, and experiments in human primary preadipocytes to show Death Associated Protein Kinase 2 (*DPAK2*) as a new causal gene for NAFLD. We profiled the transcriptomes of liver and subcutaneous adipose tissues from patients undergoing a bariatric surgery. In parallel, liver histology was performed to accurately stage NAFLD and nonalcoholic steatohepatitis (NASH), which is a more severe form of NAFLD and a precursor of cirrhosis or hepatocellular carcinoma[37,38]. One year after the surgery, we re-sampled their adipose transcriptomes to investigate the effect of weight loss on gene expression. We identified 43 genes that are downregulated in the adipose tissue of NASH patients. Furthermore, the adipose expression of 17 of them is negatively correlated with liver steatosis score and serum TGs. Adipose expression of five of the 17 genes was also correlated with diet-induced liver steatosis in a large panel of mouse strains[39]. We examined adipose transcriptomes at follow-up, and found the expression of one of the five genes, *DAPK2*, recovered after the weight loss. We carried out mediation analyses and leveraged longitudinal transcriptome information to infer a causal pathway among obesity, serum TGs, *DAPK2* adipose expression, and NAFLD. Our results support the causal effect of obesity on NAFLD, mediated by adipose expression of *DAPK2* and serum TGs. We knocked down *DAPK2* in human primary preadipocytes via siRNA to determine the functional consequences of reduced *DAPK2* expression in the adipose tissue. Five key genes involved in autophagy were significantly

downregulated, including *TGM2* and *ULK2*, which also regulate adipocyte differentiation[40,41]. In summary, our genomic and experimental data suggest impaired autophagy and adipocyte differentiation due to *DAPK2* reduction in the adipose tissue as a novel pathogenic mechanism of NAFLD.

Since the inception of GWAS, we have gained important insight into the genetic architectures of many human diseases. Nevertheless, we still fall short on encompassing more diverse populations that remain grossly underrepresented in genomic studies. From 2009 to 2016, the percent of non-European GWAS has steadily increased from 4% to 20%[3,42], but it will continue to take the combined effort of the research community to remedy this disparity. The contribution of GWAS to delineate disease mechanisms remains limited; however there are few promising examples utilizing functional genomic data to elucidate the underlying molecular mechanisms of variants and genes at T2D, schizophrenia, and obesity GWAS loci[21,24,43]. Better methods to seamlessly combine genetic and functional genomic data will greatly facilitate the translation of GWAS findings to applicable medical information. In the following chapters, I will provide examples from our work that aim to reduce these knowledge gaps.

References

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

2. Burton, P.R., Clayton, D.G., Cardon, L.R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D.P., McCarthy, M.I., Ouwehand, W.H., Samani, N.J., et al. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661–678.

3. Popejoy, A.B., and Fullerton, S.M. (2016). Genomics is failing on diversity. Nature *538*, 161–164.

4. Wray, N.R., Goddard, M.E., and Visscher, P.M. (2007). Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res. *17*, 1520–1528.

5. Rosenberg, N.A., Huang, L., Jewett, E.M., Szpiech, Z.A., Jankovic, I., and Boehnke, M. (2010). Genome-wide association studies in diverse populations. Nature Reviews Genetics *11*, 356–366.

6. Martin, A.R., Gignoux, C.R., Walters, R.K., Wojcik, G.L., Neale, B.M., Gravel, S., Daly, M.J., Bustamante, C.D., and Kenny, E.E. (2017). Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. Am. J. Hum. Genet. *100*, 635–649.

7. Vilhjalmsson, B.J., Yang, J., Finucane, H.K., Gusev, A., Lindstrom, S., Ripke, S., Genovese, G., Loh, P.-R., Bhatia, G., Do, R., et al. (2015). Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. Am. J. Hum. Genet. *97*, 576–592.

8. Reisberg, S., Iljasenko, T., Läll, K., Fischer, K., and Vilo, J. (2017). Comparing distributions of polygenic risk scores of type 2 diabetes and coronary heart disease within different populations. PLoS ONE *12*, e0179238.

9. Ko, A., Cantor, R.M., Weissglas-Volkov, D., Nikkola, E., Reddy, P.M.V.L., Sinsheimer, J.S., Pasaniuc, B., Brown, R., Alvarez, M., Rodriguez, A., et al. (2014). Amerindian-specific regions under positive selection harbour new lipid variants in Latinos. Nat Commun *5*.

10. Miller, M., Stone, N.J., Ballantyne, C., Bittner, V., Criqui, M.H., Ginsberg, H.N., Goldberg, A.C., Howard, W.J., Jacobson, M.S., Kris-Etherton, P.M., et al. (2011). Triglycerides and cardiovascular disease: a scientific statement from the American Heart Association. Circulation *123*, 2292–2333.

11. Schaid, D.J., Chen, W., and Larson, N.B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. Nature Reviews Genetics *360*, 1759.

12. Wang, Z., and Chatterjee, N. (2017). Increasing mapping precision of genome-wide association studies: to genotype and impute, sequence, or both? Genome Biol. *18*.

13. Spain, S.L., and Barrett, J.C. (2015). Strategies for fine-mapping complex traits. Hum. Mol. Genet. *24*, R111–R119.

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

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

16. Claussnitzer, M., Dankel, S.N., Kim, K.-H., Quon, G., Meuleman, W., Haugen, C., Glunk, V., Sousa, I.S., Beaudry, J.L., Puviindran, V., et al. (2015). FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. N. Engl. J. Med. *373*, 895–907.

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

18. FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest, A.R.R., Kawaji, H., Rehli, M., Baillie, J.K., de Hoon, M.J.L., Lassmann, T., Itoh, M., Summers, K.M., Suzuki, H., et al. (2014). A promoter-level mammalian expression atlas. Nature *507*, 462–470.

19. (2014). A promoter-level mammalian expression atlas. *507*, 462–.

20. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet. *45*, 580–585.

21. van de Bunt, M., Fox, J.E.M., Dai, X., Barrett, A., Grey, C., Li, L., Bennett, A.J., Johnson, P.R., Rajotte, R.V., Gaulton, K.J., et al. (2015). Transcript Expression Data from Human Islets Links Regulatory Signals from Genome-Wide Association Studies for Type 2 Diabetes and Glycemic Traits to Their Downstream Effectors. PLoS Genet. *11*.

22. Mifsud, B., Tavares-Cadete, F., Young, A.N., Sugar, R., Schoenfelder, S., Ferreira, L., Wingett, S.W., Andrews, S., Grey, W., Ewels, P.A., et al. (2015). Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat. Genet. *47*, 598–606.

23. Miao, Z., Alvarez, M., Pajukanta, P., and Ko, A. (2018). ASElux: an ultra-fast and accurate allelic reads counter. Bioinformatics *34*, 1313–1320.

24. Pan, D.Z., Garske, K.M., Alvarez, M., Bhagat, Y.V., Boocock, J., Nikkola, E., Miao, Z., Raulerson, C.K., Cantor, R.M., Civelek, M., et al. (2018). Integration of human adipocyte chromosomal interactions with adipose gene expression prioritizes obesity-related genes from GWAS. Nat Commun *9*, 1512.

25. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W.J.H., Jansen, R., de Geus, E.J.C., Boomsma, D.I., Wright, F.A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. Nat. Genet. *48*, 245–252.

26. Gusev, A., Mancuso, N., Won, H., Kousi, M., Finucane, H.K., Reshef, Y., Song, L., Safi, A., McCarroll, S., Neale, B.M., et al. (2018). Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. Nat. Genet. *50*, 538–548.

27. Stevenson, K.R., Coolon, J.D., and Wittkopp, P.J. (2013). Sources of bias in measures of allele-specific expression derived from RNA-seq data aligned to a single reference genome. BMC Genomics 2013 14:1 *14*, 536.

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

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

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

31. Bouzigon, E., Corda, E., Aschard, H., Dizier, M.-H., Boland, A., Bousquet, J., Chateigner, N., Gormand, F., Just, J., Le Moual, N., et al. (2008). Effect of 17q21 variants and smoking exposure in early-onset asthma. N. Engl. J. Med. *359*, 1985–1994.

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

33. Younossi, Z.M., Koenig, A.B., Abdelatif, D., Fazel, Y., Henry, L., and Wymer, M. (2016). Global epidemiology of nonalcoholic fatty liver disease—Meta-analytic assessment of prevalence, incidence, and outcomes. Hepatology *64*, 73–84.

34. Schuppan, D., and Afdhal, N.H. (2008). Liver cirrhosis. Lancet *371*, 838–851.

35. Romeo, S., Kozlitina, J., Xing, C., Pertsemlidis, A., Cox, D., Pennacchio, L.A., Boerwinkle, E., Cohen, J.C., and Hobbs, H.H. (2008). Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. Nat. Genet. *40*, 1461–1465.

36. Kozlitina, J., Smagris, E., Stender, S., Nordestgaard, B.G., Zhou, H.H., Tybjærg-Hansen, A., Vogt, T.F., Hobbs, H.H., and Cohen, J.C. (2014). Exome-wide association study identifies a TM6SF2 variant that confers susceptibility to nonalcoholic fatty liver disease. Nat. Genet. *46*, 352–356.

37. Adams, L.A., Lymp, J.F., St Sauver, J., Sanderson, S.O., Lindor, K.D., Feldstein, A., and Angulo, P. (2005). The natural history of nonalcoholic fatty liver disease: a population-based cohort study. Gastroenterology *129*, 113–121.

38. Nagaoki, Y., Hyogo, H., Aikata, H., Tanaka, M., Naeshiro, N., Nakahara, T., Honda, Y., Miyaki, D., Kawaoka, T., Takaki, S., et al. (2012). Recent trend of clinical features in patients with hepatocellular carcinoma. Hepatol. Res. *42*, 368–375.

39. Hui, S.T., Parks, B.W., Org, E., Norheim, F., Che, N., Pan, C., Castellani, L.W., Charugundla, S., Dirks, D.L., Psychogios, N., et al. (2015). The genetic architecture of NAFLD among inbred strains of mice. Elife *4*.

40. Myneni, V.D., Melino, G., and Kaartinen, M.T. (2015). Transglutaminase 2-a novel inhibitor of adipogenesis. Cell Death Dis *6*, –e1868.

41. Ro, S.-H., Jung, C.H., Hahn, W.S., Xu, X., Kim, Y.-M., Yun, Y.S., Park, J.-M., Kim, K.H., Seo, M., Ha, T.-Y., et al. (2013). Distinct functions of Ulk1 and Ulk2 in the regulation of lipid metabolism in adipocytes. Autophagy *9*, 2103–2114.

42. Need, A.C., and Goldstein, D.B. (2009). Next generation disparities in human genomics: concerns and remedies. Trends Genet. *25*, 489–494.

43. Won, H., la Torre-Ubieta, de, L., Stein, J.L., Parikshak, N.N., Huang, J., Opland, C.K., Gandal, M.J., Sutton, G.J., Hormozdiari, F., Lu, D., et al. (2016). Chromosome conformation elucidates regulatory relationships in developing human brain. Nature *538*, 523–527.

Chapter 2

Amerindian-specific regions under positive selection harbour new lipid variants in Latinos

nature COMMUNICATIONS

ARTICLE

Received 24 Feb 2014 | Accepted 29 Apr 2014 | Published 2 Jun 2014

DOI: 10.1038/ncomms4983 OPEN

Amerindian-specific regions under positive selection harbour new lipid variants in Latinos

Arthur Ko[1,2], Rita M. Cantor[1], Daphna Weissglas-Volkov[1], Elina Nikkola[1], Prasad M.V. Linga Reddy[1], Janet S. Sinsheimer[1], Bogdan Pasaniuc[1,3,4], Robert Brown[4], Marcus Alvarez[1], Alejandra Rodriguez[1], Rosario Rodriguez-Guillen[5,6], Ivette C. Bautista[5], Olimpia Arellano-Campos[5], Linda L. Muñoz-Hernández[5], Veikko Salomaa[7], Jaakko Kaprio[7,8,9], Antti Jula[7], Matti Jauhiainen[7], Markku Heliövaara[7], Olli Raitakari[10,11], Terho Lehtimäki[12], Johan G. Eriksson[7,13,14], Markus Perola[7,9,15], Kirk E. Lohmueller[16], Niina Matikainen[17], Marja-Riitta Taskinen[17], Maribel Rodriguez-Torres[5], Laura Riba[5,6], Teresa Tusie-Luna[5,6], Carlos A. Aguilar-Salinas[5] & Päivi Pajukanta[1,2]

Dyslipidemia and obesity are especially prevalent in populations with Amerindian backgrounds, such as Mexican–Americans, which predispose these populations to cardiovascular disease. Here we design an approach, known as the cross-population allele screen (CPAS), which we conduct prior to a genome-wide association study (GWAS) in 19,273 Europeans and Mexicans, in order to identify Amerindian risk genes in Mexicans. Utilizing CPAS to restrict the GWAS input variants to only those differing in frequency between the two populations, we identify novel Amerindian lipid genes, receptor-related orphan receptor alpha (RORA) and salt-inducible kinase 3 (*SIK3*), and three loci previously unassociated with dyslipidemia or obesity. We also detect lipoprotein lipase (LPL) and apolipoprotein A5 (*APOA5*) harbouring specific Amerindian signatures of risk variants and haplotypes. Notably, we observe that *SIK3* and one novel lipid locus underwent positive selection in Mexicans. Furthermore, after a high-fat meal, the *SIK3* risk variant carriers display high triglyceride levels. These findings suggest that Amerindian-specific genetic architecture leads to a higher incidence of dyslipidemia and obesity in modern Mexicans.

[1] Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, California 90095, USA. [2] Molecular Biology Institute at UCLA, Los Angeles, California 90095, USA. [3] Department of Pathology and Laboratory Medicine, David Geffen School of Medicine at UCLA, Los Angeles, California 90095, USA. [4] Bioinformatics Interdepartmental Program, UCLA, Los Angeles, California 90095, USA. [5] Instituto Nacional de Ciencias Médicas y Nutrición, Salvador Zubiran, 14000 Mexico City, Mexico. [6] Instituto de Investigaciones Biomédicas de la UNAM, 04510 04510 Mexico City, Mexico. [7] National Institute for Health and Welfare, 00271 Helsinki, Finland. [8] Department of Public Health, Hjelt Institute, University of Helsinki, 00014 Helsinki, Finland. [9] Institute for Molecular Medicine Finland FIMM, University of Helsinki, 00014 Helsinki, Finland. [10] Research Centre of Applied and Preventive Cardiovascular Medicine, University of Turku, 20520 Turku, Finland. [11] Department of Clinical Physiology and Nuclear Medicine, Turku University Hospital, 20520 Turku, Finland. [12] Department of Clinical Chemistry, Fimlab Laboratories and University of Tampere School of Medicine, 33100 Tampere, Finland. [13] Folkhälsan Research Center, University of Helsinki, 00290 Helsinki, Finland. [14] Department of General Practice and Primary Health Care, University of Helsinki, 00014 Helsinki, Finland. [15] University of Tartu, Estonian Genome Center, 51010 Tartu, Estonia. [16] Department of Ecology and Evolutionary Biology, UCLA, Los Angeles, California 90095, USA. [17] Department of Medicine, University of Helsinki, 00014 Helsinki, Finland. Correspondence and requests for materials should be addressed to P.P. (email: ppajukanta@mednet.ucla.edu).

Dyslipidemia is a highly prevalent (53%)[1] cardiovascular risk factor in the United States that will drastically increase medical and economic burdens in the subsequent decades if prevention and treatment cannot be better tailored for those most susceptible. In addition to socioeconomic status, the prevalence of lipid disorders also varies among ethnic groups, with Hispanics being more prone to dyslipidemia than any of the other US groups[2]. With 40% of Mexican–American men and 35% of women exhibiting high triglycerides (TGs) ($>1.69\,mmol\,l^{-1}$)[2], a large portion of the population has a high risk of cardiovascular disease (CVD), especially as a direct causal relationship between hypertriglyceridemia and CVD was recently demonstrated[3]. Strikingly, the decreasing rate of CVD currently observed in Europeans[4] does not extend to Hispanic-origin populations, as exemplified by the four times higher incidence of CVD among the Amerindians when compared with Europeans[2]. Thus, identifying Hispanic-specific lipid variants is critical to deciphering the genetic pathogenesis of dyslipidemia and CVD in this rapidly growing US minority, and ultimately personalizing prevention and treatment of this major risk factor.

Despite their increased predisposition[5], Mexicans and other groups with Amerindian heritage have been substantially underrepresented in genomic studies[6,7]. Most lipid studies focus on recapturing European-origin signals in the Latino populations[8–13], with only a single Mexican lipid genome-wide association study (GWAS) reported[14]. GWAS in admixed populations are hindered by a complex population substructure that can reduce power[15]. Statistical methods, such as local ancestry inference or admixture mapping, have been employed to overcome or even utilize such ancestral variations to identify disease-associating loci in diverse populations; however, they often rely on ancestry-informative markers or parental population haplotype panels that are not readily available in all populations, as is the case with Latinos[16–18]. Fitting a mixed model or adjusting for ancestry in GWAS can circumvent the confounding effect of ancestry, but may lead to a higher false-negative rate and losing ancestry-specific variants[14,15].

To this end, we design an approach utilizing cross-population allele screen prior to GWAS (CPAS-GWAS) to identify Amerindian-origin lipid variants in Mexicans. Utilizing the CPAS-GWAS approach, we identify 18 Amerindian risk variants for lipids and obesity and one risk haplotype for TGs in Mexicans. Interestingly, the Amerindian-specific TG risk haplotype and 10 of the Amerindian lipid and obesity variants have not been implicated in lipid traits or obesity in other populations. Two of the new TG loci also show signs of potential positive selection, reflecting the possibility that maintaining high serum lipid levels was favourable during the Amerindian population history.

Results

A novel cross-population allele screen approach. To search for Amerindian-specific genetic variants that contribute to the high risk of dyslipidemia and obesity in Mexicans, we developed a CPAS-GWAS approach that first screens across the genome for variants that differ in frequency between the two ancestry populations, Europeans and Amerindians, and subsequently includes only these variants (CPAS variants) in the actual Mexican GWAS. Thus, we restricted the Mexican GWAS to variants only present in Mexicans and not in Europeans, and variants that show statistically significant differences in allele frequency between Mexicans and Europeans, as explained in detail below (see Supplementary Fig. 1 for CPAS design).

CPAS enriches for Amerindian TG variants. We first screened for population-specific variants between the admixed Mexican population and its European ancestry population represented by Finns, using Finnish and Mexican controls matched on the tested phenotype (that is, Finns and Mexicans with normal levels of TGs, total cholesterol (TC), high density lipoprotein cholesterol (HDLC) or body mass index (BMI), respectively). The purpose of the phenotypic matching is to ensure that the differences in allele frequencies are strictly due to population structure in order to focus on the variants that are population-stratified instead of confounded by other phenotypes. Based on our local ancestry estimates, African ancestry is low (2.3%) in the Mexican cohort, and accordingly, no screening between Mexican and African controls was performed.

For screening across the genome, we first imputed the GWAS data in the Finnish and Mexican cohorts to increase both the number of overlapping common variants between the cohorts and the number of low-frequency single-nucleotide polymorphisms (SNPs) (minor allele frequency (MAF) 1-5%), known to differ most between populations[19]. Overlapping SNPs with MAF > 5% in Mexicans were pruned using an R^2 cutoff of 0.5 in the Mexican controls to reduce redundancy and multiple testing. To avoid overestimation of linkage disequilibrium (LD) among the low-frequency variants, all overlapping SNPs with MAF 1–5% in Mexicans were retained.

In the actual TG CPAS screen, 967,056 SNPs (61%) exhibited a difference in allele frequencies between Mexican and Finnish TG controls that passed the Bonferroni correction ($P<3.16\times10^{-16}$) for 1,584,455 SNPs tested. A Mantel–Haenszel test showed that the MAF distribution difference is significantly greater between populations after CPAS ($P<2.20\times10^{-16}$), indicating that population-stratified variants were indeed detected (Fig. 1a,b). In addition, we compared these variants between Europeans and admixed Native Americans from the 1000 Genomes Project, and 74% of them displayed >10% difference in MAF, demonstrating that our screening does filter for variants that differ between the populations. We also included in the GWAS the 694,185 Mexican-specific SNPs that after imputations were only present in Mexicans but not in Finns to further enrich the GWAS for Amerindian-specific variants. Taken together, 1,661,241 CPAS SNPs filtered by CPAS to significantly differ between Finns and Mexicans or not present in Finns were carried forward for association testing between Mexican TG cases and controls. CPAS was also carried out for three additional traits, HDLC, TC and BMI in a similar way.

GWAS results and independent replication. We performed GWAS for high TGs in Mexicans using only the CPAS SNPs as the input. HDLC, TC and BMI were analysed as continuous traits instead to demonstrate that CPAS-GWAS is effective for quantitative traits as well. As the four phenotypes are highly correlated, we only corrected for the number of SNPs using Bonferroni in the GWAS step, followed by the replication step in which we also corrected for multiple testing using Bonferroni. The top 1% of the TG GWAS results are shown in Supplementary Data 1. We selected 15 non-redundant TG SNPs with *P*-values $1.07\times10^{-5}-6.08\times10^{-33}$ for replication in 6,159 additional Mexican individuals based on *P*-value, functional annotation and MAF difference between Mexicans and Finns (Table 1 and Supplementary Table 1). Three of the 15 SNPs were Mexican-specific as their frequencies were less than 1% in the Finnish cohort or Europeans (the 1000 Genomes database). The Mexican replication sample ($n=6,159$) consisted of an unrelated cohort and a family-based cohort (see Supplementary Table 2 for clinical characteristics). We combined the results from the two replication cohorts by performing a meta-analysis using METAL[20].

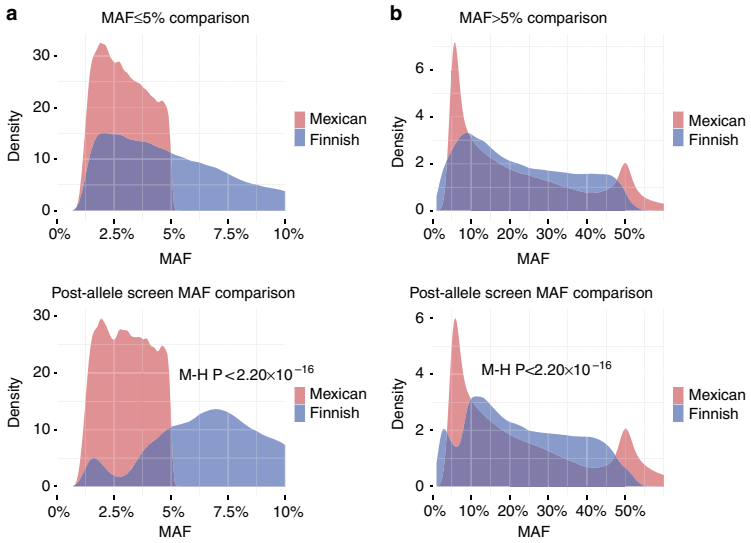


Figure 1 | Minor allele frequency distributions in the Finnish and Mexican low TG controls before and after the cross-population allele screen. (**a**) Displays the SNPs with a MAF ≤ 5% in the Mexicans. These SNPs with a MAF ≤ 5% were not pruned based on LD. (**b**) shows the SNPs with a MAF > 5% in the Mexicans. These SNPs with a MAF > 5% were pruned based on LD in the Mexican controls using an R^2 cutoff of 0.5. Mantel–Haenszel (M-H) *P*-value is displayed, indicating that the difference between the Mexican and Finnish frequencies was significantly different after the screen, and therefore, population-stratified variants are enriched due to the CPAS.

Table 1 | CPAS-GWAS and replication results for case-control comparison of TGs.

SNP	Chr	Position (bp)	MAF (%) (risk allele)	GWAS		Replication: qualitative*		Replication: quantitative*		Type	Gene or adjacent genes
				P	OR(95% CI)	*P*	*Z*	*P*	*Z*		
rs78536982	6	70,182,710	0/9/12(T)	7.62×10^{-6}	1.38 (1.18-1.61)	NS	1.01	0.044	2.01	Intergenic	*BAI3,LMBRD1*
rs62436827	6	167,548,547	12†/6/9(G)	6.88×10^{-6}	1.49 (1.24-1.78)	0.044	2.02	0.019	2.34	Intronic	*CCR6*
rs28680850	8	1,373,720	37†/50/55(A)	2.77×10^{-6}	1.26 (1.15-1.39)	0.00024	3.68	0.00011	3.86	Intergenic	*LOC286083,DLGAP2*
rs79236614	8	19,860,460	9/6/3(G)	3.79×10^{-8}	0.53 (0.42-0.67)	5.87×10^{-7}	−5.00	6.31×10^{-8}	−5.41	Intergenic	*LPL,SLC18A1*
rs4360309	8	126,523,523	49/79/84(T)	1.60×10^{-6}	1.35 (1.19-1.53)	NS	1.56	0.027	2.22	Intergenic	*TRIB1,LINC00861*
rs964184	11	116,648,917	16/29/43(G)	6.08×10^{-33}	1.89 (1.69-2.08)	1.80×10^{-39}	13.15	2.05×10^{-57}	15.97	Intergenic	*BUD13,ZNF259*
rs139961185	11	116,807,343	0/15/20(A)	1.15×10^{-12}	1.44 (1.27-1.63)	3.63×10^{-5}	4.13	3.99×10^{-10}	6.25	Intronic	*SIK3*
rs72925845	17	76,439,361	8†/6/4(A)	1.07×10^{-5}	0.61 (0.48-0.77)	NS	−1.15	0.036	−2.10	Intronic	*DNAH17*

Chr, chromosome; CI, confidence interval; MAF, minor allele frequency; NS, non-significant, $P\geq0.05$; OR, odds ratio; P, *P*-value from linear and logistic regression for quantitative and qualitative TG traits respectively; and Z, the standard score from meta-analysis.
*Meta-analysis of the family and unrelated cohorts in the replication stage. MAFs (on the scale 0–100%) are listed in the following order: Finnish low TG controls/Mexican low TG controls/Mexican high TG cases.
†Finnish MAFs of these SNPs were obtained from the Finnish population in the 1000 Genomes Project as they were missing in our Finnish cohort.

Four variants (rs28680850, rs79236614, rs964184 and rs139961185) on chromosomes 8 and 11 resulted in *P*-values less than the Bonferroni correction significance level ($P<0.0033$) in the replication stage (Table 1). Furthermore, their overall meta-analysis in all Mexican cohorts (GWAS combined with the replication cohorts, total $n=9{,}482$) resulted in *P*-values between $7.1\times10^{-9}-1.8\times10^{-67}$. Interestingly, the intergenic variant rs79236614 that resides ~100 kb downstream of the lipoprotein lipase (*LPL*) gene is in high LD ($R^2=0.91$) in Mexicans with an early stop variant in *LPL*, rs328 (S474X), that cuts off the last exon (Table 1). The novel TG variant rs28680850 on chr8p21 resides in a predicted CpG site. We verified its allele-specific effect on methylation by pyrosequencing bisulphite-treated whole blood-derived DNA samples from Mexicans. The homozygous individuals with the rs28680850 A risk allele ($n=11$) all had a 0% methylation status, whereas individuals with AG and GG genotypes ($n=48$) had a methylated CpG site with an average methylation of 57% (range 36–100%), implicating potential epigenetic regulation of TG levels. The new TG-associated variant on chr11, rs139961185 that resides in an intron of salt-inducible kinase 3 (*SIK3*), is common in Mexicans but not observed in Finns (Table 1). To eliminate the possibility that the association signal came from a correlation with the nearby, known TG-associated gene, apolipoprotein A5 (*APOA5*), we carried out a regional LD analysis (Supplementary Fig. 2). We did not observe any pair-wise $R^2>0.2$ between rs139961185 and any of the *APOA5* or *APOC3* variants, indicating that this novel Mexican-specific TG

variant in *SIK3* is independent from *APOA5* and *APOC3*. In addition, four other SNPs (rs62436827, rs4360309, rs72925845 and rs78536982) showed suggestive TG signals ($P<0.05$) in replication for the same allele and direction as in the GWAS (Table 1).

Six HDLC variants and three TC SNPs passed the genome-wide significance threshold ($P<5\times10^{-8}$) in the Mexican CPAS-GWAS (Table 2). Two HDLC hits (rs78557978 and rs148533712) and the top BMI signal (rs6027281) that reside near novel genes that have never been implicated for these traits were selected for replication. HDLC and TC variants near or in known lipid genes such as *CETP* and *CELSR2* were not selected for replication. Two novel HDLC loci, an intronic variant in receptor-related orphan receptor alpha (*RORA*) (rs148533712) and an intergenic variant near UDP glycosyltransferase 8 (rs78557978) were replicated (Table 2). Since a known HDLC-associated gene, hepatic lipase (*LIPC*), is 2.3 Mb away from rs148533712, we performed a regional LD analysis (Supplementary Fig. 3) to investigate whether this Mexican HDLC signal is independent from *LIPC*. The regional LD analysis demonstrated that the LD (in R^2) decays drastically before reaching *LIPC*, and there was no strong LD between rs148533712 and any variant within *LIPC* ($R^2<0.2$), indicating that the Mexican HDLC signal in *RORA* is independent from the previously known European *LIPC* lipid signal, as is also suggested by the relative long distance of 2.3 Mb. Interestingly, the associated interval around the latter SNP rs78557978 ($R^2>0.5$) includes only one gene, UDP glycosyltransferase 8. The replicated BMI hit, rs6027281 (Table 2) resides between *C20orf197* and *LOC284757*. However, the associated interval ($R^2>0.5$) does not extend to these adjacent predicted genes, suggesting an intergenic regulatory effect for this BMI hit or its proxy.

TG CPAS-GWAS loci are enriched for Amerindian ancestry. To provide additional support for our CPAS approach, we compared the four replicated TG signals with regions displaying enriched Amerindian ancestry in Mexican TG cases versus controls identified by using LAMP-LD[17]. Figure 2a,b shows that the four replicated TG variants reside in regions with the highest Amerindian ancestry difference across the whole genome (a percent difference > 3% and a *z*-score > 3 for an ancestry enrichment between the Mexican TG cases and controls). Supplementary Figs 4–6 show the close-up views of these loci with regional genes. Furthermore, three (rs78536982, rs72925845, and rs4360309) of the four suggestive loci also reside in regions with Amerindian enrichment (a percent difference > 2%) in Mexican high TG subjects (Supplementary Figs 7 and 8). Genome-wide ancestry difference is shown in Supplementary Fig. 9.

Genome-wide SKAT analysis supports replicated TG loci. To utilize the imputed low-frequency variants that are more likely to be population-specific[19], we examined the combined effect of common and rare variants using combined sum test with sequence kernel association test (SKAT-C) analysis[21]. Only the CPAS SNPs were included as input variants in the SKAT-C. Both 11q23 and 8p21 loci where three (rs964184, rs139961185 and rs79236614) of the replicated SNPs from the single-marker analysis reside were significant in SKAT-C ($P<7.64\times10^{-7}$) after correcting for 65,428 regions tested (Supplementary Fig. 10a–c). An additional peak near *LPL* with no GWAS hits is likely due to a cluster of regional rare variants driving the signal (Supplementary Fig. 10a). The 8p23.3 region where the fourth replicated GWAS SNP rs28680850 resides resulted in a suggestive SKAT-C *P*-value of $P=2.70\times10^{-5}$ (Supplementary Fig. 10b). These results indicate that the use of CPAS variants in SKAT helps identify regions with population-based combined effects of common and rare variants.

A mexican-specific TG risk haplotype. We observed a well-known TG- and coronary heart disease (CHD)- associated locus on chromosome 11q23[8–14,22] in three separate analyses, CPAS-GWAS, LAMP-LD (Fig. 2b), and CPAS-SKAT (Supplementary Fig. 10c), with rs964184 showing the strongest genome-wide signal for TGs ($P=6.08\times10^{-33}$) (Table 1). Interestingly, 15 additional non-redundant CPAS variants ($R^2<0.5$), all within 500 kb of the lead SNP rs964184, produced *P*-values of $5.77\times10^{-7}-1.58\times10^{-16}$ in the GWAS, four of these were Mexican specific (European MAF < 1%). When conditioned on rs964184, the 15 SNPs were no longer associated ($P>0.05$) (see Supplementary Data 2 for a detailed LD structure among these SNPs). This raised the possibility of a TG-associated, Mexican-specific haplotype on chr11. To investigate this issue, we performed the LD analysis using D′. All 15 SNPs showed high D′ (>0.5) with rs964184, and a haplotype association analysis of these 16 SNPs resulted in an overall *P*-value of 1.04×10^{-16} between the Mexican TG cases and controls. Consequently, we

Table 2 | CPAS-GWAS and replication results for quantitative lipid traits and BMI.

SNP	Chr	Position (bp)	MAF (%) (risk allele)	GWAS		Replication		Gene or adjacent genes	Status
				P	Beta	*P*	Type		
HDLC									
rs78557978	4	115,638,601	7/17(C)	4.09×10^{-8}	−0.16	0.014*	Intergenic	*UGT8, NDST4*	Novel V/Novel G
rs11216230	11	116,884,789	0/11(A)	3.26×10^{-10}	0.22	—	Intronic	*SIK3*	Novel V/Known G
rs148533712	15	61,244,884	20/50(C)	3.41×10^{-8}	0.12	0.011*	Intronic	*RORA*	Novel V/Novel G
rs9989419	16	56,985,139	35/28(G)	2.71×10^{-9}	0.14	—	Intergenic	*HERPUD1, CETP*	Novel V/Known G
chr16:56997349:I	16	56,997,349	17/26(CA)	6.75×10^{-20}	-0.22	—	Intronic	*CETP*	Known V/Known G
rs5880	16	57,015,091	3†/11(C)	1.76×10^{-16}	-0.29	—	Ns/exonic	*CETP*	Novel V/Known G
TC									
rs3902354	1	109,819,296	34/25(A)	1.16×10^{-8}	0.15	—	Intergenic	*CELSR2*	Known V/Known G
chr16:56997349:I	16	56,997,349	20/26(CA)	1.58×10^{-8}	−0.15	—	Intronic	*CETP*	Known V/Known G
rs118146573	16	57,000,938	10/14(A)	3.79×10^{-10}	−0.20	—	Intronic	*CETP*	Known V/Known G
BMI									
rs6027281	20	58,656,151	28/12(C)	7.10×10^{-8}	−0.19	0.0082	Intergenic	*C20orf197, LOC284757*	Novel V/Novel G

beta, effect size; Chr, chromosome; G, gene; MAF, minor allele frequency; P, *P*-value from linear regression analysis; and NS, nonsynonymous; *UGT8*, UDP glycosyltransferase 8; V, variant.
MAFs (on the scale 0–100%) are listed in the following order: Finnish controls/Mexican overall.
*Proxy variants with $R^2>0.94$ were used in replication.
†The Finnish MAF of this SNP was obtained from the Finnish population in the 1000 Genomes Project as it was missing in our Finnish cohort.

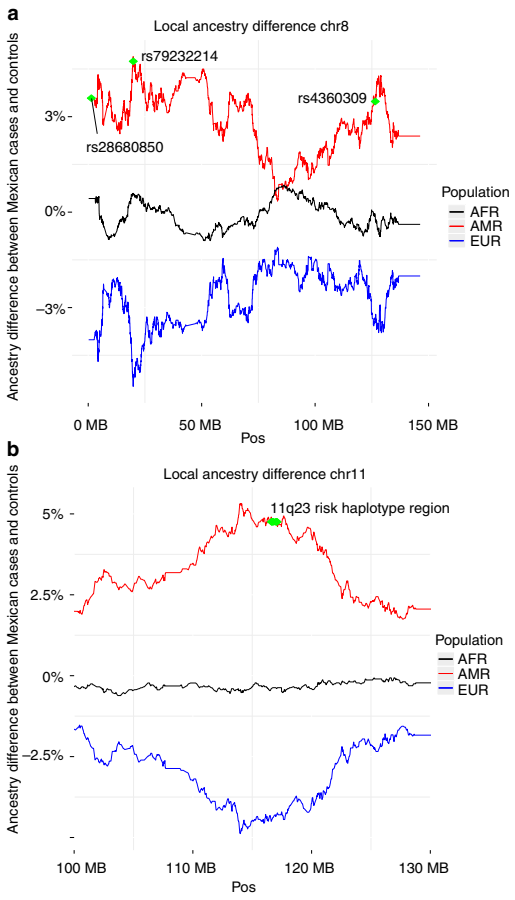


Figure 2 | Local ancestry difference between Mexican low TG controls and high TG cases in the genomic regions implicated by the CPAS-GWAS. (**a**) Local ancestry results are shown for chromosome 8. Rs28680850 and rs79236614 were both significant after Bonferroni correction and rs4360309 displayed a suggestive signal in the GWAS. All three variants reside in regions that show Amerindian enrichment in Mexican high TG cases (>3% Amerindian ancestry difference). (**b**) Local ancestry difference between Mexican low TG controls and high TG cases on chromosome 11q23 where the TG risk haplotype region resides. The seven haplotype-tagging SNPs are shown as green diamonds that are clustered together in the plot. These LAMP-LD[17] results indicate that the 11q23 region is highly enriched for Amerindian ancestry in the Mexican high TG cases.

identified a 460-kb TG-associated risk haplotype (HT1) formed by seven SNPs (Fig. 3a, Supplementary Table 3) with a haplotype frequency of 18% in Mexican TG cases (overall haplotype $P = 1.93 \times 10^{-24}$ and the risk haplotype $P = 1.29 \times 10^{-12}$) (Supplementary Table 3). The two other TG-increasing haplotypes (HT2 and HT3) resulted in *P*-values of 9.96×10^{-9} and 1.40×10^{-5} (Supplementary Table 3). Figure 3b shows the MAFs of the seven haplotype SNPs in the Finns and Mexicans.

Logistic regression of the TG case/control status on the HT1 haplotype carrier status resulted in OR = 1.65 ($P = 7.79 \times 10^{-14}$), suggesting that HT1 is a significant risk factor for high TGs in Mexicans. Interestingly, HT1 is Mexican-specific and not observed in Finns, because it is tagged by rs964184 and rs139961185 (Supplementary Table 3) of which rs139961185 is Mexican-specific (not observed in Finns and MAF = 0.5% in the 1000 Genomes Europeans). This Mexican-specific risk HT1 also showed strong association with high TGs in the replication cohorts ($P = 7.09 \times 10^{-12}$, OR = 1.46) with a frequency of 20% in the Mexican TG cases (overall haplotype $P = 2.83 \times 10^{-41}$ and the risk haplotype $P = 2.51 \times 10^{-13}$).

Two causative TG variants on the haplotype background. To identify causative variants travelling on the haplotype background, we examined all SNPs in the haplotype region, focusing on the Mexican-specific HT1. Eight exonic SNPs on the HT1 background, as well as one known hypertriglyceridemia promoter SNP[14,23] on the HT2 background, were further investigated based on differences in allele frequencies and potential deleterious effect (Fig. 3c; Supplementary Table 4). To identify variants that best explain the Mexican TG case/control status, we carried out a stepwise logistic regression including all nine SNPs. Rs11820589 and rs662799 were retained in the model ($P < 0.00001$; Fig. 3c) with a pseudo-R^2 value of 0.057, indicating that these SNPs tagged by the risk haplotypes explain ~6% of high TG levels in the Mexican cohort. Interestingly, rs11820589 is in LD ($R^2 = 0.82$) with a known non-synonymous variant, rs3135506 in *APOA5* (ref. 24). A PolyPhen 2 score of 0.993 and a SIFT score of 0 for rs3135506 indicate a possible damaging effect on the protein. Thus, a change in TGs attributed to rs11820589 is likely due to the effect of rs3135506 on *APOA5*. Based on ENCODE data, rs662799 (2 kb upstream of *APOA5*) is a strong enhancer in a HepG2 liver cell line, probably regulating *APOA5* in *cis*, as *APOA5* is highly expressed in liver. In summary, these two variants explain ~6% of TG levels in Mexicans likely due to a change of function of *APOA5*.

Positive selection on Amerindian TG loci. To examine if the top TG GWAS loci were favourably retained in the Mexican population due to recent positive natural selection, we examined the integrated haplotype score (iHS) statistics of neutrality (see Methods) for all genotyped and imputed SNPs with MAF > 5% across the chr8 and chr11 regions (Fig. 4) instead of just the CPAS variants, because focusing only on the CPAS variants that differ in allele frequency between the two populations would have introduced a bias into our selection analysis[25]. In our selection analysis, we found multiple peaks of extreme |iHS| values (>4.0) in the chr11 risk haplotype region within the *SIK3* gene (Fig.4c). It is worth noting that both the Mexican-specific, TG-associated haplotype-tagging SNP, rs139961185 and the novel Mexican-specific HDLC-associated variant rs11216230 also reside in *SIK3* (Tables 1 and 2). We estimated that these extreme |iHS| scores in *SIK3* rank among the top 0.1% chromosome-wide scores based on our iHS analysis on all genotyped SNPs on the entire chr11, suggesting that *SIK3* has been under recent positive selection and thus retained unusually high homozygosity. We also identified peaks with |iHS| >4.0 near the novel TG variant on chr8, residing inside a lincRNA gene, *LOC286083*, expressed in most human tissues (Fig. 4b). The *LPL* region resulted in several |iHS| values >3.0, although no extreme |iHS| scores (>4.0) were seen in this TG region (Fig. 4a). Interestingly, the extreme iHS scores were observed with imputed SNPs, suggesting that the genotype panel does not represent well Mexican-specific variants and Latino populations in general.

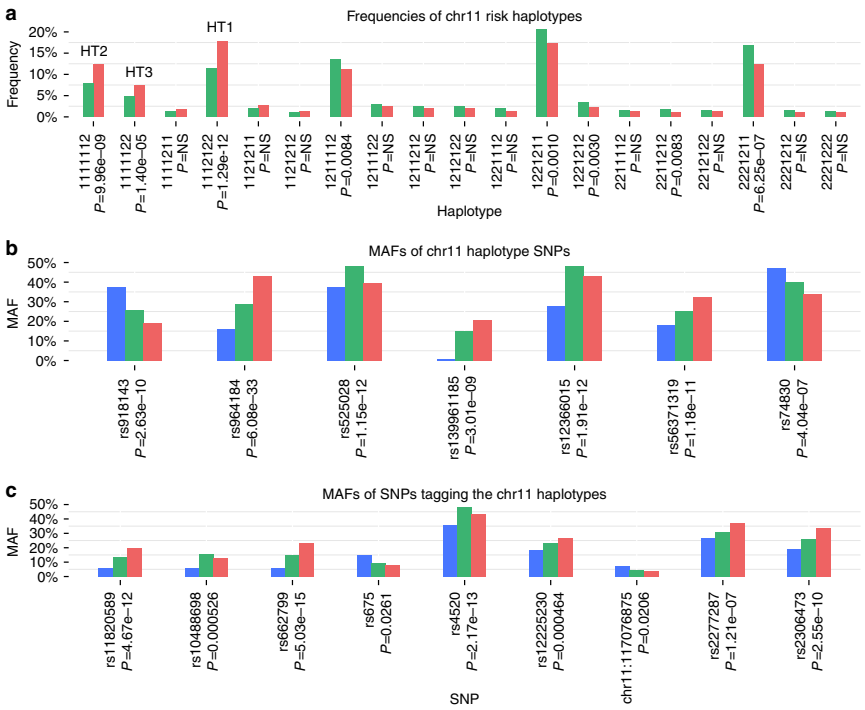


Figure 3 | Frequencies of the chr11 haplotypes and variants in the TG risk region. (**a**) Frequencies of chr11 risk haplotypes between the Mexican TG cases and controls. The *P*-value of the omnibus haplotype was 1.93×10^{-24} using the haplotype case/control test in PLINK. Red bars represent the haplotype frequencies in the Mexican cases and green bars the frequencies in the Mexican controls. NS indicates nonsignificant ($P > 0.05$). The order of the SNPs on the haplotype is rs918143 (1/C), rs964184 (1/G), rs525028 (1/G), rs139961185 (2/A), rs12366015 (1/A), rs56371319 (2/A) and rs74830 (2/T) with the TG-increasing allele given in parenthesis. (**b**) The minor allele frequencies of the seven haplotype SNPs in the order of Finnish TG controls, Mexican TG controls and Mexican TG cases. (**c**) The minor allele frequencies of the nine SNPs travelling with the two chr11 risk haplotypes in the same order of groups as in Fig. 3b above.

To investigate whether admixed ancestry confounds selection signal on chr11, we also performed the iHS analysis in all subjects homozygous for the Amerindian ancestry in the chr11 region ($n = 1{,}217$), as estimated by LAMP-LD. We observed iHS scores of 3.3 (rs609177) and 2.8 (rs111809212) in *SIK3*. Interestingly, these variants are in LD with the Mexican-specific TG risk haplotype SNP rs139961185 in *SIK3*, both resulting in $R2 > 0.54$ and $D' > 0.99$ with rs139961185. Accordingly, they were also associated with high TGs when analysed in the entire Mexican TG case/control sample ($P = 9.51 \times 10^{-7}$ and $P = 1.46 \times 10^{-10}$). These data show that the iHS scores remain large when the analysis is performed only on the Amerindian background, further supporting natural selection of *SIK3* in Mexicans.

Response to oral fat tolerance test in Mexicans. To examine if the Mexican-specific *SIK3* risk variant, rs139961185 affects postprandial TG metabolism, we carried out an oral fat tolerance test in a Mexican cohort. Briefly, the Mexican participants ate a fatty meal at the baseline and their TG levels were measured over a period of 8 h postprandially to calculate the postprandial TG response as an area under the curve (AUC) (see Methods for details of the diet study). Figure 5 demonstrates that both in the low TG (fasting baseline TG < 1.69 mmol l^{-1}) and high TG (fasting baseline TG > 1.69 mmol l^{-1}) groups (Fig. 5a) and in the combined Mexican study sample (Fig. 5b), the Mexican rs139961185 risk allele carriers consistently retained a significantly higher TG levels throughout the time course in contrast to non-carriers ($P = 0.03$ for TG AUC), suggesting that this TG-associated *SIK3* risk variant may delay TG clearance after a fatty meal in Mexicans.

Discussion

Admixed populations provide unprecedented opportunities to understand human demographic history and genetic diversity, and moreover, to uncover variants of different ancestral origin and frequency that may contribute to variations in disease prevalence between populations[26,27]. However, genetic studies in recently admixed populations have proven difficult due to the confounding effects of population substructure and the reliance on an ancestral population reference panel that might not be readily available[15,16]. To this end, we designed a CPAS-GWAS approach that restricts GWAS to include only those variants that differ in frequency between the two ancestral populations. We performed the first CPAS-GWAS to discover Amerindian variants associated with dyslipidemia and obesity in Mexicans.

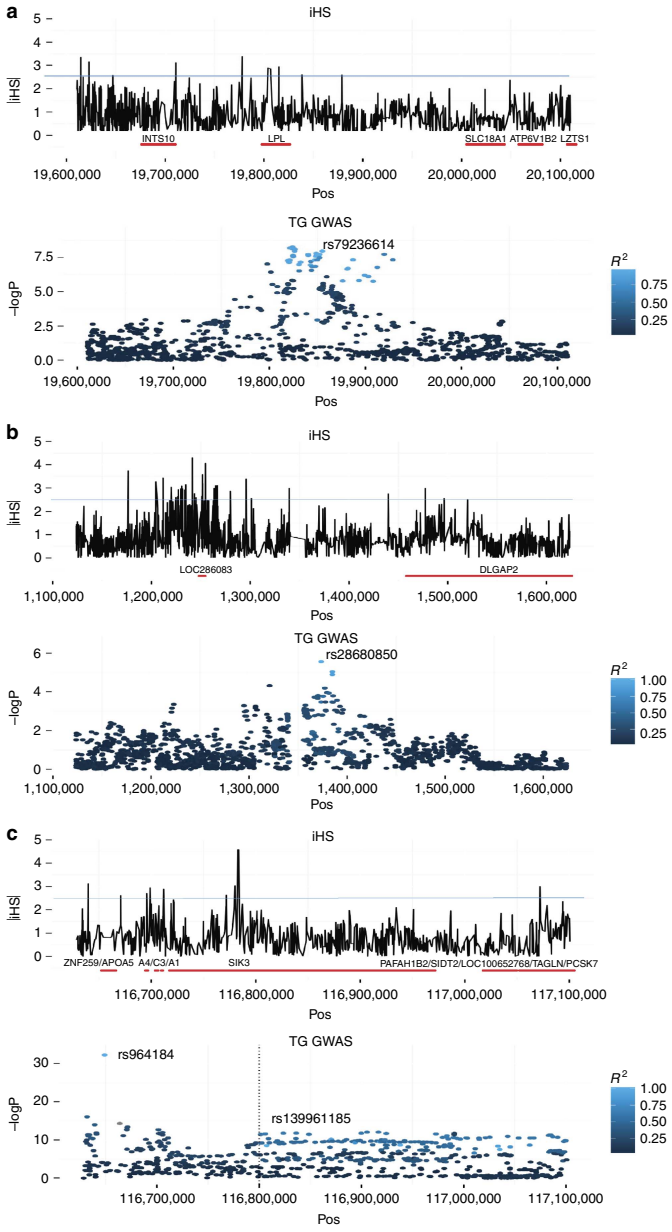


Figure 4 | Analysis of natural selection in the three Mexican TG risk regions. The absolute |iHS| were plotted across the three TG risk loci in the upper panel. A blue line indicates the top 1% chromosome-wide |iHS| threshold (>2.56). For comparison, the lower panel shows the logistic regression results of the Mexican TG case/control sample for the same SNPs (MAF>5%) in each region. LD (in R^2) is plotted against the regional lead SNP. All 3,701 Mexican individuals were included in the iHS analysis. (**a**) The |iHS| results on chr8p21. The highest peak was observed in the *LPL* promoter region although no extreme |iHS| scores (>4) were observed. (**b**) The |iHS| results on chr8p23.3. A region harbouring a lincRNA, *LOC286083* shows signs of positive selection with peaks of extreme |iHS| values. (**c**) The |iHS| results of chr11q23. Clusters of extreme |iHS| scores in the *SIK3* region suggests that it underwent positive selection pressure in Mexicans. In the lower panel, LD is measured against rs964184 or rs139961185, respectively, before or after the 168 MB bp position, indicated by the vertical line.

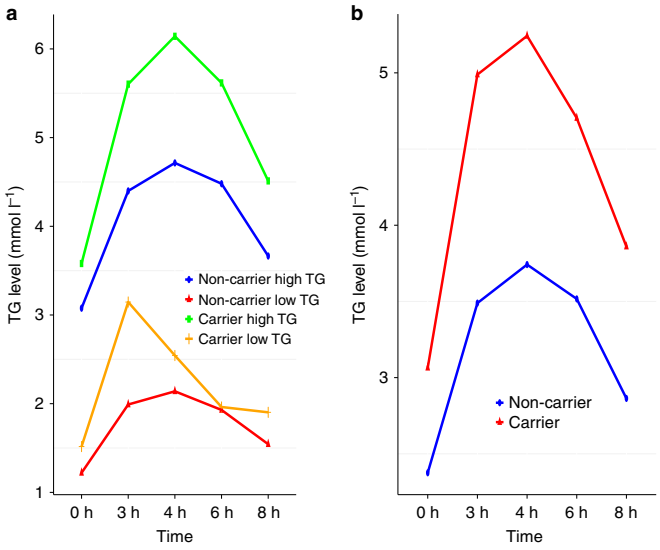


Figure 5 | Difference in postprandial TG clearance rate between rs139961185 risk allele carriers and non-carriers. The individuals carrying the rs139961185 risk allele in *SIK3* demonstrated a slower TG clearance rate ($P = 0.03$ for AUC TG from linear regression) when compared with the non-carriers consistently (**a**) in Mexicans with low and high fasting TG levels at baseline and (**b**) in the combined study sample, suggesting that *SIK3* is implicated for delayed postprandial TG clearance in Mexicans. There were 57 participants of which 3 and 9 were risk allele carriers (A/A and A/G) in the low and high TG group; and 17 and 28 were non-risk allele carriers (G/G) in the low and high TG group, respectively.

Hypoalphaproteinemia, hypertriglyceridemia and hypercholesterolaemia are more prevalent in Amerindian-origin populations than in Europeans, with 60.5% Mexicans suffering from hypoalphaproteinemia (HDLC $< 1.03\,mmol\,l^{-1}$); 43.6% from hypercholesterolaemia (TC $> 5.17\,mmol\,l^{-1}$); and 31.5% from hypertriglyceridemia (TG $> 1.69\,mmol\,l^{-1}$), respectively[1,2,5,28,29]. Clinical significance of dyslipidemia derives from the fact that patients with these lipid disorders are predisposed to CHD and often exhibit type 2 diabetes (T2D). CHD and T2D emerged as the two leading causes of death in Mexico in a recent national survey[30], and more than 65% of the Mexican diabetics have hypertriglyceridemia[31]. Furthermore, recent evidence demonstrate a causal role of TGs in CHD[3,32–34]. Thus, it is critical to focus efforts and resources on the identification of the population-specific genetic components that make hypertriglyceridemia so prevalent in Mexicans.

In contrast to other methods used to analyse admixed populations, CPAS-GWAS is able to achieve single-variant resolution uncovering susceptibility variants or their proxies instead of wider ancestry-enriched chromosomal regions identified using other approaches[15,16]. For example, our TG CPAS-GWAS identified eight Amerindian hypertriglyceridemia variants and one Amerindian-specific risk haplotype, of which all but one reside in genomic regions enriched for Amerindian ancestry in Mexican high TG cases as shown by local ancestry analysis. A two-step tree-based approach evaluating selection on a set of SNPs from several populations has previously been proposed that examines frequency difference among populations[35]. First, Bhatia *et al.*[35] built an unrooted tree utilizing Fst to identify divergence between populations followed by selection estimation at each marker common to all populations. To identify the potential traits under selection, they cross-referenced selected variants with GWAS catalogues. While CPAS and the tree-based method share similarity, they do not follow the same assumption and principle. CPAS does not assume variants to be under selection, rather we first screen for population-specific variants by comparing phenotypically matched distinct populations and then test their association with a trait directly. As a result, we can also identify population-enriched risk variants that correlate with a phenotype but are not necessarily under selection pressure, as is the case for instance with the *LPL* locus. Overall, our data demonstrate that CPAS-GWAS can effectively screen for ancestry-specific susceptibility variants in admixed populations.

CPAS-GWAS is not restricted to a single admixed population or trait, and in fact, it can easily be tailored for other populations or diseases as shown by our qualitative TG and quantitative HDLC, TC and BMI CPAS-GWAS analyses. Moreover, CPAS-GWAS is not vulnerable to estimation of local ancestry that can be nontrivial if the appropriate parental populations are unknown or unavailable, as is often the case for admixed Latino populations[16]. Accordingly, false positives due to incorrect ancestry calculations are major concerns of local ancestry inference[15,16]. However, CPAS-GWAS does not face the same challenge as this step is eliminated. One limitation of the CPAS-GWAS approach is that its resolution and accuracy rely on the density of the genotyping arrays and the quality of imputation, but both will likely be circumvented in the near future as whole genome or exome sequencing become common practice as the price of sequencing continues to drop. We utilized Finns as surrogates of Europeans in CPAS, because Finns are the single largest population group investigated in extensive European lipid GWAS studies[11,13], suggesting that Latino comparisons against Finns should sufficiently screen against the European lipid signals.

Chromosome 11q23 harbours a well-known TG-associated APOA1C3A4A5 gene cluster, and the variant rs964184 has been

implicated for TGs in multiple populations[8–14]. In this key TG region, CPAS-GWAS identified Amerindian TG risk variants and haplotype signatures, of which the most striking example is HT1 with zero frequency in Europeans and 20% frequency in Mexican TG cases. Of variants tagged by the haplotypes, rs11820589 and rs662799 explain ~6% of variability of TGs in Mexicans. Rs11820589 is in strong LD with a non-synonymous SNP (S19W), rs3135506, a known TG-increasing variant that resulted in a three-fold lower plasma Apo A-V levels when introduced in the mouse genome[24]. Rs662799, previously associated with both TGs and CHD[23], resides in the promoter or enhancer region of *APOA5*. It is worth noting that these TG risk variants rs3135506 and rs662799 are >2 and ~4 times more prevalent in the Mexican TG controls and Mexican TG cases than in the Finnish TG controls, respectively.

APOA5 is a potent regulator of serum TG levels, as knockout mice lacking *apoa5* have four times higher TG levels; mice expressing a human *APOA5* transgene have one-third lower plasma TG levels; and overexpression of *APOA5* reduces TG levels in mice[36–38]. In addition, *APOA5* stimulates the *LPL*-mediated VLDL-TG hydrolysis via interaction with proteoglycan-bound *LPL*[38,39]. The variants rs662799 and rs3135506 likely affect the function of *APOA5*, which in turn regulates *LPL* that is reflected as elevated TG levels in Mexicans. Targeted sequencing of the chr11q23 haplotype region that has substantial Amerindian ancestry in Mexican TG cases is bound to identify additional functional variants that influence TG levels in Amerindian-origin populations.

We also identified two TG loci on chr8p21 and chr8p23 with a significant Amerindian ancestry in the Mexican TG cases. Rs79632214 is located downstream of the key TG gene, *LPL*, previously associated with TGs and CHD[11,40,41]. In Mexicans rs79632214 is in tight LD with rs328 (S474X), resulting in an early stop in *LPL*. Interestingly, our SKAT-C data implicated the presence of multiple Amerindian rare risk variants in the *LPL* region contributing significantly to TGs in Mexicans. Variant rs28680850 on chr8p21 is intergenic and the region has not previously been implicated for lipids in other populations. Our initial data show that this novel TG variant influences differential methylation of a CpG site, suggesting that allele-specific methylation contributes to the underlying biological mechanism.

CPAS-GWAS also identified two novel replicated HDLC loci and one BMI locus that reside near or within genes that have never been associated with either trait in human. Interestingly, the new HDLC variant rs148533712 on chr15 is located in an intron of the retinoic acid *RORA* gene, and it is an independent signal of *LIPC*. *RORA* is a known transcriptional activator of *APOA5*, *APOA1* and *APOC3*[42–44], all residing in the Mexican risk haplotype region on chr11, suggesting distinct converging lipid pathways underlying dyslipidemia in Mexicans. At the chr20 BMI locus, protein phosphatase 1, regulatory subunit 3D (*PPP1R3D*) was recently identified for obesity in mice[45]. Thus, additional genes affecting BMI likely exist at this locus.

To the best of our knowledge, we carried out the first study examining positive selection of GWAS loci for metabolic traits in an admixed population. TG is the most plausible trait under selection at these loci since our diet study implicates *SIK3* in delayed TG clearance after a fatty meal; the chr11 locus displays the strongest association signal with TGs both in Mexicans and Europeans; and the novel chr8p23.3 region does not have significant associations with any other traits we tested ($P > 0.0003$). Furthermore, converging evidence from our selection analysis and diet study; TG and HDLC CPAS-GWAS; as well as a previous mouse model all support the role of *SIK3* in metabolic functions. Interestingly, these Mexican-specific TG and HDLC CPAS variants in *SIK3* are not present, and thus have not previously been identified in extensive European lipid GWAS studies[11,13], suggesting that there are Amerindian-specific genetic lipid pathways involving *SIK3*. Notably, recent data on a *Sik3* knockout mouse identified *SIK3* as a novel energy regulator, altering cholesterol and bile acid metabolism by coupling with retinoid metabolism[46]. We also searched the Gene Expression Omnibus[47] database at the NCBI and ArrayExpress[48] database at the European Bioinformatics Institute to verify that *SIK3* is expressed in human liver and adipose tissues, the most relevant tissues in lipid metabolism. Furthermore, the iHS analysis suggests that *SIK3* has been under positive selection pressure, pointing to an advantageous role for *SIK3* in reproductive survival. However, whether selection pressure was acting on Amerindians prior to or after admixture requires further investigation. One possible explanation is that the ability to retain sufficiently high serum lipid levels could have contributed to the survival when resources were scarce during the early period of human habitation in the America continent. As a result, this genetic background was preferentially retained in the population. Additionally, in line with the selection results, our fatty diet study demonstrated that the Mexican-specific rs139961185 TG risk allele is significantly associated with delayed postprandial TG clearance in Mexicans, further supporting the role of *SIK3* in TG metabolism and its candidacy for future functional studies. Individually, these findings do not stand alone as evidence of selection on TGs. However, taken together, they suggest that the *SIK3* gene, associated with TGs in modern Mexicans, has undergone selection at some point during the Amerindian lineage. *SIK3* may thus be a genetic responder to the Western diet that was recently introduced to Latinos, contributing to increased susceptibility to metabolic diseases in modern Mexicans. Additional future studies with whole-genome sequence data will help more comprehensively evaluate selection of lipid traits across the genome in Mexicans.

In summary, we developed the CPAS-GWAS approach to uncover Amerindian variants in Mexicans that contribute to their greater susceptibility to dyslipidemia and obesity when compared with Europeans. Of the novel lipid genes we identified, *RORA* and *SIK3* are of major interest. *RORA* is a transcriptional ligand-regulated mediator of multiple key lipid genes[42–45]. Furthermore, selective inhibition of the retinoic-acid-receptor-related orphan receptors via synthetic ligands has been suggested as a viable therapeutic approach for metabolic disorders[49]. Based on our findings from CPAS-GWAS, local ancestry, selection analysis, and oral fat tolerance test, we hypothesize that *SIK3* may have played an important role in maintaining high plasma TG level that was historically critical for Amerindian survival but led to a higher rate of dyslipidemia and obesity in modern Hispanics after the adaption of Western diet. Our results suggest *SIK3* as a strong candidate for future functional investigation to elucidate the molecular basis of the high prevalence of dyslipidemia in Mexicans.

Methods

Human subjects. A total of 19,273 participants from Finnish ($n = 9,791$) and Mexican ($n = 9,482$) cohorts were included in the study (see Supplementary Table 2 for clinical characteristics). All studies were approved by local research ethic committees: the Institutional Review Boards (IRB) of the Helsinki, Turku and Tampere University Hospitals; IRB of the National Institute for Health and Welfare; IRB of the Instituto Nacional de Ciencias Médicas y Nutrición, Salvador Zubiran; and IRB of UCLA), and all participants gave informed consent.

We screened six Finnish population-based cohorts with GWAS data available[50–52] (total $n = 14,217$) for individuals with low serum TG levels (TGs < 1.69 mmol l^{-1}) and not taking lipid-lowering medication. Fasting TG values were used to determine the low TG status, except for the FINRISK cohort. However, since non-fasting does increase and does not decrease serum TG levels, the use of non-fasting TGs in that cohort should not influence the results. A subset of 9,791 Finnish individuals with low TGs were included in the cross-population

screening step from the Northern Finland Birth Cohort 1966 (NFBC66) ($n=4,427$), the Cardiovascular Risk in Young Finns Study ($n=1,428$), Helsinki Birth Cohort Study ($n=991$), Health2000 GenMets Study ($n=1,301$), FinnTwin12 and FinnTwin16 cohort studies (Twins) ($n=421$; one randomly selected twin in each twin pair was selected to investigate only unrelated subjects), and FINRISK ($n=1,223$). The Finnish GWAS data on the NFBC1966 Study has been previously deposited in the NIH dbGAP data repository under the accession code phs000276.v1.p1.

Two Mexican cohorts ascertained for hypertriglyceridemia[14] or T2D[53] were combined and screened for low TG controls (fasting TGs $<1.69\,mmol\,l^{-1}$) ($n=1,645$) and high TG cases (fasting TGs $>2.26\,mmol\,l^{-1}$) ($n=1,678$), excluding individuals on lipid-lowering medication. The Mexican participants were recruited at the Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico City.

In the replication stage, we investigated 6,159 additional Mexican individuals for replication of 15 SNPs using the same criteria for the hypertriglyceridemia status as in the cross-population allele screen, which resulted in 2,129 high TG cases, 2,985 low TG controls and 903 family members from 73 Mexican dyslipidemic families[14,54,55]. To utilize all individuals with lipid phenotypes available in these cohorts ($n=6,159$), we also analysed log-transformed serum TGs as a quantitative trait.

Serum TGs, HDLC and TC were measured using enzymatic and enzymatic colorimetric methods with commercial reagents in the Finnish and Mexican cohorts[50–54]. The cut-points for TG cases (TGs $>2.26\,mmol\,l^{-1}$) and TG controls (TGs $<1.69\,mmol\,l^{-1}$) are based on the American Heart Association TG guidelines. The general population means of HDLC, TC and BMI in Finns and Mexicans were used as cut-points in the two populations for the CPAS stage to screen for controls. The thresholds of the three traits for controls in Finns and Mexicans were as follows: HDLC $>1.15\,mmol\,l^{-1}$ and HDLC $>1.54\,mmol\,l^{-1}$; TC $<5.17\,mmol\,l^{-1}$ for both populations; and BMI $<25\,kg\,m^{-2}$ and BMI $<27\,kg\,m^{-2}$, respectively.

The Mexican participants ($n=57$) included in the fatty meal diet study were recruited at the Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico City.

Genotyping and imputation. In the CPAS, Illumina genotyping platforms were used for all cohorts, as described in detail previously[14,50–52]. The NFBC cohorts were genotyped with the HumanHap CNV 370k array: GenMets and FINRISK with the HumanHap 610 k array: and Young Finns Study, Helsinki Birth Cohort Study and Twins with the HumanHap 670 k array, respectively. The Mexican cohorts were genotyped using Human 610 BeadChip and Human Omni 2.5 BeadChip array, respectively. Genotype quality control was performed on each cohort separately using the following inclusion criteria: SNP and sample genotyping success rate $\geq 95\%$, MAF $\geq 1\%$, Hardy–Weinberg equilibrium (HWE) $P \geq 1\times10^{-6}$, and individual heterozygosity rate <4s.d. Samples with gender discrepancies or closely related individuals were removed.

In the replication stage, SNPs were genotyped using Sequenom and TaqMan platform. These SNPs had a genotype call rate $\geq 90\%$, and they passed a Bonferroni corrected HWE P-value >0.05 for the number of tested SNPs. In addition, the family data were checked for Mendelian errors using the Mendel[56] mistyping option.

Imputation was carried out separately in Mexicans and Finns. To reduce imputation runtime, we first pre-phased the Mexican and Finnish cohorts separately using SHAPEIT with the 1000 Genomes Project reference panel[57,58]. Subsequently, imputation was carried out using IMPUTE2 utilizing the 1000 Genomes Project reference panel as well[59,60]. Following the IMPUTE2 guideline and results from a previous study, we employed a cosmopolitan imputation strategy that included all populations from the 1000 Genomes Project to maximize accuracy and the number of imputed SNPs[16,61]. Imputed data were filtered using the following quality control criteria: info ≥ 0.8, probability ≥ 0.9, MAF $\geq 1\%$ and HWE ($P>0.0001$).

Bisulphite pyrosequencing. The methylation status of the CpG site containing the SNP, rs28680850, was measured using bisulphite pyrosequencing with custom-designed kit from EpigenDx according to the standard protocol for bisulphite treatment and pyrosequencing by the manufacturer.

Association analyses. Association testing at the CPAS step and the subsequent GWAS was carried out for the binary TGs status with logistic regression using an additive genetic model, including age, sex and BMI as covariates to control for their potential confounding effects on serum TGs at the allele screen step. For the quantitative CPAS-GWAS analysis of HDLC and TC levels, HDLC and TC levels were first log-transformed to approximate normal distribution, and multiple linear regression was used with age, sex, BMI, global ancestry estimates and the high TG status as covariates. For the quantitative CPAS-GWAS analysis of BMI, age and sex were used as covariates in linear regression, as no inflation was observed (Supplementary Figs 11–14). Imputed SNPs were analysed using SNPTEST v2.4 (ref. 62) and the score method was used to incorporate the imputation uncertainties into the regression model. Redundant SNPs with a MAF $>5\%$ were pruned based on LD with $R^2 \geq 0.5$ in Mexican controls. In the CPAS step for qualitative TGs, a Bonferroni correction for 1,584,455 tested SNPs ($P<3.16\times10^{-8}$) was used to identify variants that have different allele frequencies in Mexicans and Finns, resulting in 967,056 SNPs that were significantly different and carried forward to the TG GWAS (Supplementary Fig. 1). The set of SNPs ($n=694,185$) that were variable in Mexicans but were monomorphic in the Finnish cohorts were also included in the GWAS to capture additional Amerindian-specific TG-associated variants. A total of 1,661,241 SNPs were analysed in the TG GWAS (Supplementary Fig. 1). We also performed CPAS for the three additional traits, HDLC, TC and BMI in a similar way (Table 2). The quantile–quantile plots (Supplementary Figs. 11–14) of the all GWAS results with the CPAS SNPs demonstrate that most of the distribution behaves as the expected null, ruling out major confounders.

Haplotype logistic regression, step-wise logistic regression and McKelvey and Zavoina pseudo-R^2 analysis, and Mantel–Haenszel test were all performed in R statistical package (http://www.r-project.org/). Conditional association analysis on rs964184 was carried out using SNPTEST2.4 with the SNP genotype as a covariate.

Association analyses of the 15 TG SNPs genotyped in the replication stage were performed employing the same logistic regression model as in the GWAS using PLINKv1.08 package[63]. In the replication stage, we also performed a quantitative trait analysis on log-transformed TG levels including sex and age as covariates using PLINK. For the two HDLC SNPs, linear regression was carried out using PLINK as well, including sex, age, BMI, high TG status and global ancestry as covariates. Part of the independent cohort ($n=2,121$) was used for HDLC replication as these samples have global ancestry estimates available. The family cohort was analysed using the quantitative trait locus association option of Mendel[64]. After taking into account multiple testing using Bonferroni correction, P-values of 0.0033 (15 tested SNPs), 0.025 (two tested SNPs) and 0.05 (one tested SNP) were considered as statistically significant in the replication stage for TG, HDLC and BMI SNPs, respectively, when combining the P-values of the two replication cohorts by weighting by sample size using METAL[20] or the subset of independent cohort for HDLC.

Analysis of combined rare and common variant effects was carried out using SKAT-C implemented in R with a window size of 50 kb and a sliding window of 40 kb. To increase the number of rare variants in SKAT, we used a 5% frequency cutoff. Alternatively, we also calculated the rare variant frequency as $1/\sqrt{2N}$ where N is the sample size ($N=3,701$).

Local ancestry inference. To investigate whether variants identified utilizing the cross-population allele screen approach reside in chromosomal genomic regions enriched for Amerindian ancestry in the Mexican high TG cases, we carried out local ancestry estimation utilizing Local Ancestry in adMixed Populations using LD (LAMP-LD)[17]. A three-population mixed model was assumed to estimate proportions of the three ancestral populations (European, Amerindian and African) in the modern Mexicans[65]. The parental population reference panels were constructed from individuals in the Genetics of Asthma in Latino Americans[66]study as described in detail previously[18] and LAMP-LD was run with default parameters, window size 300 and 15 hidden Markov models states, on each chromosome separately. To identify Amerindian enriched regions associating with TG, the standard scores of the difference in local Amerindian ancestry between the Mexican TG cases and controls were calculated for each region. A significance threshold of z-score >2 was used to call ancestral enrichment. To calculate the percent difference between cases and controls for each ancestral population, the proportion of all parental populations was estimated for every window in cases and controls separately, and the difference was calculated between the cases and controls for individual ancestry.

Analysis of positive natural selection. To examine if the 8p21, 8p23.3 and chr11q23 TG risk regions have undergone partial selective sweeps, we searched for haplotypes that were unusually long, given the frequency of the focal variant[67]. Specifically, we first estimated extended haplotype homozygosity using the 'rehh' R package[68]. Next, we calculated the integrated extended haplotype homozygosity for both ancestral and derived alleles for each genotyped SNP with MAF $>5\%$ and then calculated the standardized natural log ratio of integrated extended haplotype homozygosity between ancestral and derived alleles (iHS)[25]. Similarly, we also calculated the iHS scores for imputed variants only in the two chr8 TG risk regions and chr11 risk haplotype region due to computing time. All calculations were performed in the entire Mexican GWAS study sample and including all variants (MAF $>5\%$) without any ascertainment or CPAS screening to avoid a potential bias. We used the top 1% chromosome-wide absolute iHS ($|$iHS$|$) score (>2.56) as a cutoff to identify SNPs showing extremely large values of iHS.

Fatty meal study in Mexican cohort. The 57 Mexican participants underwent an oral fat tolerance test after a 12-hour overnight fast. The fatty meal contained 1,000 kcal; 72 g fat (saturated fat 65%, monounsaturated fat 30%, polyunsaturated fat 5%) with polyunsaturated:saturated fat ratio of 0.08, 490 mg cholesterol, 50 g carbohydrate and 38 g protein, as described in detail earlier[69]. In this diet study, blood samples were drawn at the baseline and at 3, 4, 6 and 8 h postprandially. Postprandial TG response was calculated as an AUC, as described in detail

earlier[70]. The intronic *SIK3* variant rs139961185 was genotyped in the 57 participants of which 20 had fasting TG levels $<1.7\,mmol\,l^{-1}$ at the baseline (the low TG group) and 37 had fasting TG levels $>1.7\,mmol\,l^{-1}$ at the baseline (the high TG group). To test for association between rs139961185 and postprandial TG clearance rate, a linear regression for TG AUC was performed using an additive genetic model and adjusting for the baseline TG status.

References

1. Tóth, P. P., Potter, D. & Ming, E. E. Prevalence of lipid abnormalities in the United States: The National Health and Nutrition Examination Survey 2003–2006. *J. Clin. Lipidol.* **6,** 325–330 (2012).
2. LaRosa, J. C. & Brown, C. D. Cardiovascular risk factors in minorities. *Am. J. Med.* **118,** 1314–1322 (2005).
3. Do, R. *et al.* Common variants associated with plasma triglycerides and risk for coronary artery disease. *Nat. Genet.* **45,** 1345–1352 (2013).
4. Nichols, M., Townsend, N., Scarborough, P. & Rayner, M. Trends in age-specific coronary heart disease mortality in the European Union over three decades: 1980-2009. *Eur. Heart J.* **34,** 3017–3027 (2013).
5. Aguilar-Salinas, C. A. *et al.* Hypoalphalipoproteinemia in populations of Native American ancestry: an opportunity to assess the interaction of genes and the environment. *Curr. Opin. Lipidol.* **20,** 92–97 (2009).
6. Rosenberg, N. A. *et al.* Genome-wide association studies in diverse populations. *Nat. Rev. Genet.* **11,** 356–366 (2010).
7. Bustamante, C. D., Burchard, E. G. & La Vega, De, F. M. Genomics for the world. *Nature* **475,** 163–165 (2011).
8. Bryant, E. K. *et al.* A multiethnic replication study of plasma lipoprotein levels-associated SNPs identified in recent GWAS. *PLoS ONE* **8,** e63469 (2013).
9. Dumitrescu, L. *et al.* Genetic determinants of lipid traits in diverse populations from the population architecture using genomics and epidemiology (PAGE) study. *PLoS Genet.* **7,** e1002138 (2011).
10. Elbers, C. C. *et al.* Gene-centric meta-analysis of lipid traits in African, East Asian and Hispanic populations. *PLoS ONE* **7,** e50198 (2012).
11. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466,** 707–713 (2010).
12. Wu, Y. *et al.* Trans-ethnic fine-mapping of lipid loci identifies population-specific signals and allelic heterogeneity that increases the trait variance explained. *PLoS Genet.* **9,** e1003379 (2013).
13. Global Lipids Genetics Consortium *et al.* Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45,** 1274–1283 (2013).
14. Weissglas-Volkov, D. *et al.* Genomic study in Mexicans identifies a new locus for triglycerides and refines European lipid loci. *J. Med. Genet.* **50,** 298–308 (2013).
15. Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* **11,** 459–463 (2010).
16. Seldin, M. F., Pasaniuc, B. & Price, A. L. New approaches to disease mapping in admixed populations. *Nat. Rev. Genet.* **12,** 523–528 (2011).
17. Baran, Y. *et al.* Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics.* **28,** 1359–1367 (2012).
18. Pasaniuc, B. *et al.* Analysis of Latino populations from GALA and MEC studies reveals genomic loci with biased local ancestry estimation. *Bioinformatics* **29,** 1407–1415 (2013).
19. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
20. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26,** 2190–2191 (2010).
21. Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J. D. & Lin, X. Sequence kernel association tests for the combined effect of rare and common variants. *Am. J. Hum. Genet.* **92,** 841–853 (2013).
22. Schunkert, H. *et al.* Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat. Genet.* **43,** 333–U153 (2011).
23. Triglyceride Coronary Disease Genetics Consortium and Emerging Risk Factors Collaboration *et al.* Triglyceride-mediated pathways and coronary disease: collaborative analysis of 101 studies. *Lancet* **375,** 1634–1639 (2010).
24. Ahituv, N. N., Akiyama, J. J., Chapman-Helleboid, A. A., Fruchart, J. J. & Pennacchio, L. A. L. *In vivo* characterization of human APOA5 haplotypes. *Genomics* **90,** 6–6 (2007).
25. Voight, B. F., Kudaravalli, S., Wen, X. Q. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol.* **4,** 446–458 (2006).
26. Hancock, A. M. *et al.* Adaptations to climate-mediated selective pressures in humans. *PLoS Genet.* **7,** e1001375 (2011).
27. Corona, E. *et al.* Analysis of the genetic basis of disease in the context of worldwide human relationships and migration. *PLoS Genet.* **9,** e1003447 (2013).
28. Stevens, G. *et al.* Characterizing the epidemiological transition in Mexico: national and subnational burden of diseases, injuries, and risk factors. *PLoS Med.* **5,** e125 (2008).
29. Aguilar-Salinas, C. A. *et al.* Prevalence of dyslipidemias in the Mexican National Health and Nutrition Survey 2006. *Salud. Publica. Mex.* **52**(Suppl 1): S44–S53 (2010).
30. González-Pier, E. *et al.* Priority setting for health interventions in Mexico's System of Social Protection in Health. *Salud. Publica. Mex.* **49**(Suppl 1): S37–S52 (2007).
31. Rull, J. A. *et al.* Epidemiology of type 2 diabetes in Mexico. *Arch. Med. Res.* **36,** 188–196 (2005).
32. Cullen, P. Evidence that triglycerides are an independent coronary heart disease risk factor. *Am. J. Cardiol.* **86,** 943–949 (2000).
33. Emerging Risk Factors Collaboration *et al.* Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. *Lancet* **375,** 2215–2222 (2010).
34. Keenan, T. E. & Rader, D. J. Genetics of lipid traits and relationship to coronary artery disease. *Curr. Cardiol. Rep.* **15,** 396 (2013).
35. Bhatia, G. *et al.* Genome-wide comparison of African-ancestry populations from CARe and other cohorts reveals signals of natural selection. *Am. J. Hum. Genet.* **89,** 368–381 (2011).
36. Pennacchio, L. A. L. *et al.* An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. *Science* **294,** 169–173 (2001).
37. van der Vliet, H. N., Schaap, F. G. & Levels, J. Adenoviral overexpression of apolipoprotein AV reduces serum levels of triglycerides and cholesterol in mice. *Biochem. Biophys. Res. Commun.* **295,** 1156–1159 (2002).
38. Merkel, M. *et al.* Apolipoprotein AV accelerates plasma hydrolysis of triglyceride-rich lipoproteins by interaction with proteoglycan-bound lipoprotein lipase. *J. Biol. Chem.* **280,** 21553–21560 (2005).
39. Nilsson, S. K. S., Heeren, J. J., Olivecrona, G. G. & Merkel, M. M. Apolipoprotein A-V; a potent triglyceride reducer. *Atherosclerosis* **219,** 15–21 (2011).
40. Johansen, C. T. *et al.* Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nat. Genet.* **42,** 684–687 (2010).
41. Middelberg, R. P. S. R. *et al.* Genetic variants in LPL, OASL and TOMM40/APOE-C1-C2-C4 genes are associated with multiple cardiovascular-related traits. *BMC Med. Genet.* **12,** 123–123 (2011).
42. Genoux, A. *et al.* ApoA-V: the regulation of a regulator of plasma triglycerides. *Arterioscler. Thromb. Vasc. Biol.* **25,** 1097–1099 (2005).
43. Lind, U. *et al.* Identification of the human ApoAV gene as a novel RORalpha target gene. *Biochem. Biophys. Res. Commun.* **330,** 233–241 (2005).
44. Jakel, H., Nowak, M., Helleboid-Chapman, A., Fruchart-Najib, J. & Fruchart, J.-C. Is apolipoprotein A5 a novel regulator of triglyceride-rich lipoproteins? *Ann. Med.* **38,** 2–10 (2006).
45. Morton, N. M. *et al.* A stratified transcriptomics analysis of polygenic fat and lean mouse adipose tissues identifies novel candidate obesity genes. *PLoS ONE* **6,** e23944 (2011).
46. Uebi, T. *et al.* Involvement of SIK3 in glucose and lipid homeostasis in mice. *PLoS ONE* **7,** e37803 (2012).
47. Edgar, R., Domrachev, M. & Lash, A. E. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30,** 207–210 (2002).
48. Brazma, A. *et al.* ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* **31,** 68–71 (2003).
49. Solt, L. A. & Burris, T. P. Action of RORs and their ligands in (patho)physiology. *Trends Endocrinol. Metab.* **23,** 619–627 (2012).
50. Kettunen, J. *et al.* Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nat. Genet.* **44,** 269–U65 (2012).
51. Vartiainen, E. *et al.* Thirty-five-year trends in cardiovascular risk factors in Finland. *Int. J. Epidemiol.* **39,** 504–518 (2010).
52. Soranzo, N. *et al.* A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nat. Genet.* **41,** 1182–1190 (2009).
53. Alberto Gamboa-Melendez, M. *et al.* Contribution of common genetic variation to the risk of type 2 diabetes in the Mexican Mestizo population. *Diabetes* **61,** 3314–3321 (2012).
54. Weissglas-Volkov, D. *et al.* Common hepatic nuclear factor-4alpha variants are associated with high serum lipid levels and the metabolic syndrome. *Diabetes* **55,** 1970–1977 (2006).
55. Barquera, S. *et al.* Methodology of the fasting sub-sample from the Mexican Health Survey, 2000. *Salud. Publica. Mex.* **49,** s421–s426 (2007).
56. Lange, K. *et al.* Mendel: the Swiss army knife of genetic analysis programs. *Bioinformatics* **29,** 1568–1570 (2013).
57. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat. Meth.* **9,** 179–181 (2012).
58. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Meth.* **10,** 5–6 (2013).
59. 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** (2009).

60. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3 (Bethesda)* **1,** 457–469 (2011).
61. Gao, X. *et al.* Genotype imputation for Latinos using the HapMap and 1000 genomes project reference panels. *Front. Genet.* **3** (2012).
62. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39,** 906–913 (2007).
63. 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).
64. Lange, K., Sinsheimer, J. S. & Sobel, E. Association testing with Mendel. *Genet. Epidemiol.* **29,** 36–50 (2005).
65. Price, A. L. *et al.* A genomewide admixture map for Latino populations. *Am. J. Hum. Genet.* **80,** 1024–1036 (2007).
66. Burchard, E. G. *et al.* Lower bronchodilator responsiveness in Puerto Rican than in Mexican subjects with asthma. *Am. J. Respir. Crit. Care Med.* **169,** 386–392 (2004).
67. Sabeti, P. C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419,** 832–837 (2002).
68. Gautier, M. & Vitalis, R. rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics* **28,** 1176–1177 (2012).
69. Matikainen, N. *et al.* Vildagliptin therapy reduces postprandial intestinal triglyceride-rich lipoprotein particles in patients with type 2 diabetes. *Diabetologia* **49,** 2049–2057 (2006).
70. Matthews, J. N., Altman, D. G., Campbell, M. J. & Royston, P. Analysis of serial measurements in medical research. *BMJ* **300,** 230–235 (1990).

Acknowledgements

We thank the Finns and Mexicans who participated in the study. We also thank Cindy Montes, Salvador Ramírez-Jiménez and Anu Loukola for technical assistance. This study is funded by the NIH grants HL-095056 and HL-28481 (P.P., R.M.C., D.W.-V., J.S.S.) and GM-053275 (B.P. and J.S.S.); by the NIH training grant in Genomic Analysis and Interpretation T32HG002536 (A.K.); and by the grants CONACyT 1288877 and 138826; and DGAPA, UNAM IT214711-3 (T.T.L. and L.R.). We also thank everybody involved in the Helsinki Birth Cohort Study, supported by grants from the Academy of Finland, the Finnish Diabetes Research Society, Samfundet Folkhälsan, Novo Nordisk Foundation, Finska Läkaresällskapet, Signe and Ane Gyllenberg Foundation and Wellcome Trust (grant WT089062). The Young Finns Study has been supported by the Academy of Finland: grants 126925, 121584, 124282, 129378 (Salve), 117787 (Gendi) and 41071 (Skidi); the Social Insurance Institution of Finland; Kuopio, Tampere and Turku University Hospital Medical Funds (grant 9N035 for T.L.); Juho Vainio Foundation; Paavo Nurmi Foundation; Finnish Foundation of Cardiovascular Research; Finnish Cultural Foundation; Tampere Tuberculosis Foundation; and Emil Aaltonen Foundation (T.L). V.S. was supported by the Academy of Finland, grant 139635 and the Finnish Foundation for Cardiovascular Research; M.J. by the Academy of Finland, grant 257545; and M.A. by the NIGMS of the NIH under award R25GM055052. A.R. is a recipient of the Eugene V. Cota-Robles Fellowship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Data collection and genotyping of the twin cohorts has been supported by National Institute of Alcohol Abuse and Alcoholism (grants AA-12502, AA-00145, and AA-09203 to R.J. Rose and AA15416 and K02AA018755 to D.M. Dick), the Academy of Finland (grants 100499, 205585, 118555, and 141054 to J.K.), and the Wellcome Trust Sanger Institute. The NFBC1966 Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with the Broad Institute, UCLA, University of Oulu, and the National Institute for Health and Welfare in Finland. This manuscript was not prepared in collaboration with investigators of the NFBC1966 Study and does not necessarily reflect the opinions or views of the NFBC1966 Study Investigators, Broad Institute, UCLA, University of Oulu, National Institute for Health and Welfare in Finland and the NHLBI. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

Study design: A.K. and P.P. Methods development and statistical analysis: A.K., P.P., R.M.C., J.S.S. and D.W.-V. Imputations and computational analysis: A.K. Analysis of local ancestry and/or natural selection: A.K., B.P., P.P., K.E.L., J.S.S., R.M.C., and R.B. Replication stage genotyping and quality control: A.K., E.N., P.M.V.L.R., M.A., A.R. and P.P. Data collection and GWAS genotyping: R.R.-G., I.C.B., O.A.-C., L.L.M.-H., V.S., J.K., A.J., M.J., M.H., O.R., T.L., J.G.E., M.P., M.-R.T, N. M., L.R., T.T.-L., C.A.-S, D.W.-V., E.N., and P.P. Manuscript: A.K. and PP wrote the manuscript and all authors read, reviewed and/or edited the manuscript.

Additional information

Accession codes: The Mexican hyperTG case-control GWAS data have been deposited in NIH dbGAP database under the accession code phs000618.v1.p1.

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

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

How to cite this article: Ko, A. *et al.* Amerindian-specific regions under positive selection harbour new lipid variants in Latinos. *Nat. Commun.* 5:3983 doi: 10.1038/ncomms4983 (2014).

Chapter 3

Integrative approaches for large-scale transcriptome-wide association studies

nature genetics

Integrative approaches for large-scale transcriptome-wide association studies

Alexander Gusev[1–3], Arthur Ko[4,5], Huwenbo Shi[6], Gaurav Bhatia[1–3], Wonil Chung[1], Brenda W J H Penninx[7], Rick Jansen[7], Eco J C de Geus[8], Dorret I Boomsma[8], Fred A Wright[9], Patrick F Sullivan[10–12], Elina Nikkola[4], Marcus Alvarez[4], Mete Civelek[13], Aldons J Lusis[4,13], Terho Lehtimäki[14], Emma Raitoharju[14], Mika Kähönen[15], Ilkka Seppälä[14], Olli T Raitakari[16,17], Johanna Kuusisto[18], Markku Laakso[18], Alkes L Price[1–3], Päivi Pajukanta[4,5] & Bogdan Pasaniuc[4,6,19]

Many genetic variants influence complex traits by modulating gene expression, thus altering the abundance of one or multiple proteins. Here we introduce a powerful strategy that integrates gene expression measurements with summary association statistics from large-scale genome-wide association studies (GWAS) to identify genes whose *cis*-regulated expression is associated with complex traits. We leverage expression imputation from genetic data to perform a transcriptome-wide association study (TWAS) to identify significant expression-trait associations. We applied our approaches to expression data from blood and adipose tissue measured in ~3,000 individuals overall. We imputed gene expression into GWAS data from over 900,000 phenotype measurements to identify 69 new genes significantly associated with obesity-related traits (BMI, lipids and height). Many of these genes are associated with relevant phenotypes in the Hybrid Mouse Diversity Panel. Our results showcase the power of integrating genotype, gene expression and phenotype to gain insights into the genetic basis of complex traits.

Although a large proportion of variability in complex human traits is due to genetic variation, the mechanistic steps between genetic variation and traits are generally not understood[1–7]. Many genetic variants influence complex traits by modulating gene expression, thus altering the abundance of one or multiple proteins[8–12]. Such relationships between expression and traits could be investigated through association scans in individuals for whom both measurements are available[8,13,14]. Unfortunately, studies that measure gene expression have been hampered by specimen availability and cost, with the few published studies of expression and complex traits being orders of magnitude smaller than studies of traits alone. Consequently, many expression-trait associations cannot be detected, especially those with small effects. To mitigate the reduced power from small sample size, alternative approaches have examined the overlap of genetic variants that influence gene expression (expression quantitative trait loci, eQTLs) with trait-associated variants identified in large, independent GWAS[5,6,8,9,11–13,15]. However, this approach is also likely to miss expression-trait associations of small effect.

We developed a new approach to identify genes whose expression is significantly associated with complex traits in individuals without directly measured expression levels (Online Methods). We leveraged a relatively small set of reference individuals for whom both gene expression and genetic variation (SNPs) were measured to impute the *cis* genetic component of expression into a much larger set of phenotyped individuals using their SNP genotype data (**Fig. 1**). The imputed expression data can be viewed as a linear model of genotypes with weights based on the correlation between SNPs and gene expression in the training data while accounting for linkage disequilibrium (LD) among SNPs. We then correlated the imputed gene expression with traits to perform a TWAS and identify significant expression-trait associations (Online Methods). Work in parallel to ours has also proposed to find expression-trait associations through imputation

[1]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA. [2]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA. [3]Program in Medical and Population Genetics, Broad Institute, Cambridge, Massachusetts, USA. [4]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California, USA. [5]Molecular Biology Institute, University of California, Los Angeles, Los Angeles, California, USA. [6]Bioinformatics Interdepartmental Program, University of California, Los Angeles, Los Angeles, California, USA. [7]Department of Psychiatry, VU University Medical Center, Amsterdam, the Netherlands. [8]Department of Biological Psychology, VU University, Amsterdam, the Netherlands. [9]Bioinformatics Research Center, Department of Statistics, Department of Biological Sciences, North Carolina State University, Raleigh, North Carolina, USA. [10]Department of Genetics, University of North Carolina, Chapel Hill, North Carolina, USA. [11]Department of Psychiatry, University of North Carolina, Chapel Hill, North Carolina, USA. [12]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [13]Department of Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California, USA. [14]Department of Clinical Chemistry, Fimlab Laboratories and University of Tampere School of Medicine, Tampere, Finland. [15]Department of Clinical Physiology, Pirkanmaa Hospital District and University of Tampere School of Medicine, Tampere, Finland. [16]Research Centre of Applied and Preventive Cardiovascular Medicine, University of Turku, Turku, Finland. [17]Department of Clinical Physiology and Nuclear Medicine, Turku University Hospital, Turku, Finland. [18]Department of Medicine, University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland. [19]Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California, USA. Correspondence should be addressed to A.G. (agusev@hsph.harvard.edu) or B.P. (pasaniuc@ucla.edu).

Received 22 June 2015; accepted 14 January 2016; published online 8 February 2016; doi:10.1038/ng.3506

of gene expression when GWAS data at an individual level are available[16]. However, a critical limitation is that large-scale GWAS data are typically only publicly available at the level of summary association statistics (for example, SNP effect sizes)[2–4]. To capitalize on the largest GWAS studies performed thus far (typically with data available only at the summary level), we extended our approach to impute the expression-trait association statistics directly from GWAS summary statistics (Online Methods). In contrast to expression imputation from individual-level data[16], imputation of expression-trait associations from GWAS summary statistics can exploit publically available data from hundreds of thousands of samples. Linear predictors naturally extend to indirect imputation of the standardized effect of *cis* genetic components of expression on traits starting from only GWAS association statistics[2–4] (Online Methods). This allowed us to increase the effective sample size for expression-trait association testing to hundreds of thousands of individuals. By focusing only on the genetic component of expression, we avoid instances of expression-trait association that are not a consequence of genetic variation but are driven by variation in traits (**Fig. 2**). Our approach can be conceptualized as a test for significant *cis* genetic correlation between expression and traits.

We applied our approaches to expression data from blood and adipose tissue measured in ~3,000 individuals overall. Through extensive simulations and analyses of real data, we show that our proposed approach increases performance over standard GWAS and eQTL-guided GWAS. Furthermore, we reanalyzed a 2010 lipids GWAS[17] to find 25 new expression-trait associations in those data. Among these associations, 19 of 25 contained genome-wide significant SNPs in the more recent and expanded lipids study[5], thus showcasing the power of our approach to find robust associations. We imputed gene expression into GWAS data from over 900,000 phenotype measurements[5–7] to identify 69 new genes significantly associated with obesity-related traits (body mass index (BMI), lipids and height). Many of these genes were associated with relevant phenotypes in the Hybrid Mouse Diversity Panel (HMDP). Overall, our results showcase the power of integrating genotype, gene expression and phenotype to gain insights into the genetic basis of complex traits.

RESULTS

SNP heritability of gene expression

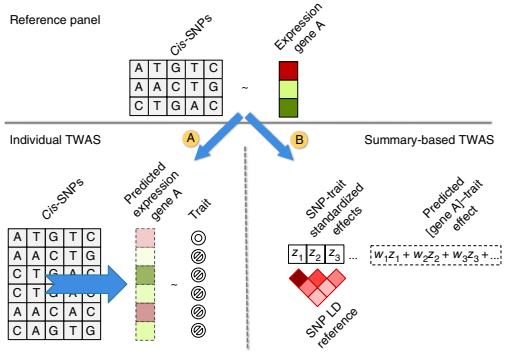
To investigate the potential use of a TWAS based on imputed gene expression, we first estimated the *cis* (1-Mb window around a gene) and *trans* (rest of the genome) SNP heritability (*cis*- and *trans*- h_g^2) for each gene in our data[18,19]. These metrics quantify the maximum possible accuracy (in terms of R^2) of a linear predictor from the corresponding set of SNPs[20,21] (Online Methods). We used 3,234 individuals for whom genome-wide SNP data and expression measurements were available from the Metabolic Syndrome in Men (METSIM; adipose), Young Finns Study (YFS; blood) and Netherlands Twins Registry (NTR; blood) data sets[22–24] (Online Methods and **Supplementary Table 1**). All expression measurements were adjusted for batch confounders, and array probes were merged into a single expression value for each gene, where possible (Online Methods).

Figure 1 Schematic of the TWAS approach. Top, estimate gene expression effect sizes in the reference panel either directly (eQTL), modeling all SNPs (BLUP), or modeling SNPs and effect sizes (BSLMM). Path A: predict expression directly for genotyped samples using the effect sizes from the reference panel and measure the association between predicted expression and a trait. Path B: indirectly estimate association between predicted expression and a trait as the weighted linear combination of SNP-trait standardized effect sizes while accounting for LD among SNPs.

Consistent with previous work[24,25], we observed significantly nonzero ($P < 1 \times 10^{-16}$) estimates of heritability across all three studies, with mean *cis*-h_g^2 values ranging from 0.01 to 0.07 and mean *trans*-h_g^2 values ranging from 0.04 to 0.06 in genes where estimates converged (**Supplementary Fig. 1** and **Supplementary Table 1**). Although we observed large differences in the average *cis*-h_g^2 estimates between the two blood cohorts, the estimates were strongly correlated across genes (Pearson $\rho = 0.47$ for YFS-NTR, as compared to $\rho = 0.15$ and $\rho = 0.26$ for METSIM-NTR and METSIM-YFS, respectively). This is consistent with a common but not identical genetic architecture. The *cis*-h_g^2 estimate was significantly nonzero (by likelihood-ratio test) for 6,924 genes after accounting for multiple hypotheses (1,985 for METSIM, 3,836 for YFS and 1,103 for NTR) (**Supplementary Fig. 1**), whereas current sample sizes were too small to detect individually significant *trans*-heritable genes. As expected, we also observed a high overlap of genes with significant *cis*-h_g^2 estimates across cohorts (**Supplementary Fig. 2** and **Supplementary Table 2**). We focused subsequent analyses on the 6,924 *cis*-heritable genes, as such genes are typically enriched for trait associations[7,9,13,24–29].

TWAS performance in simulation and cross-validation

We evaluated whether the expression levels of the 6,924 highly heritable genes could be accurately imputed from *cis*-SNP genotype data alone in these three cohorts. In each tissue, we used cross-validation to compare predictions from the best *cis*-eQTL to those from all SNPs at the locus either in a best linear unbiased predictor (BLUP) or a Bayesian model[30,31] (Online Methods). On average, the Bayesian linear mixed model (BSLMM)[31], which uses all *cis*-SNPs and estimates the underlying effect size distribution, attained the best performance, with a 32% gain in prediction R^2 over a prediction computed using only the top *cis*-eQTL (**Fig. 3** and **Supplementary Fig. 3**). BSLMM exhibited a long tail of increased accuracy, more than doubling the prediction R^2 for 25% of genes (**Supplementary Fig. 4**). In contrast to complex traits, where hundreds of thousands of training samples are required for accurate prediction[32,33], a substantial portion of variance in expression can be predicted at current sample sizes because of the much smaller number of independent SNPs in the *cis* region[21]. Furthermore, larger training sizes will continue to increase the total number of genes that can be accurately predicted (**Fig. 4**). We further evaluated cross-cohort prediction of these genes in the YFS and NTR cohorts, which were roughly equally sized and had expression measured in whole blood by microarray but were genotyped on different platforms and were from different Scandinavian populations. After accounting for *cis* heritability in the test cohort,



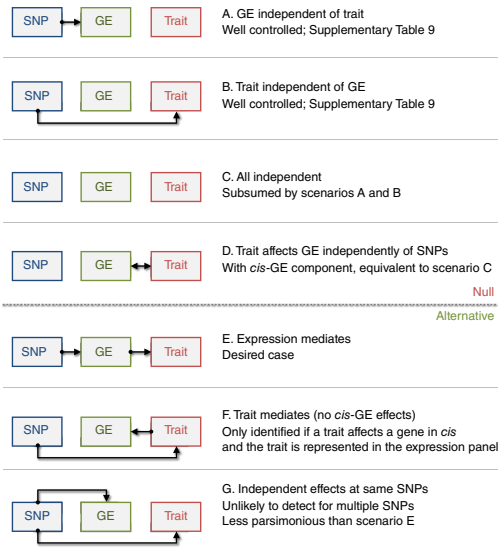
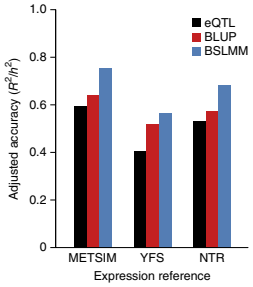


Figure 2 Modes of expression causality. Diagrams are shown for the possible modes of causality for the relationship between genetic markers (SNPs; blue), gene expression (GE; green) and traits (red). Scenarios A–D would be considered null by the TWAS model. Scenarios E–G could be identified as significant.

our cross-cohort standardized accuracy ($R^2/cis\text{-}h^2_g$) was broadly consistent with in-cohort cross-validation accuracy (**Supplementary Table 3**). BSLMM was again the most accurate predictor, with an average cross-cohort $R^2/cis\text{-}h^2_g$) value of 72%, outperforming the best eQTL by an average of 1.17×.

Next, we focused on evaluating the power of the TWAS approach to detect significant expression-trait associations using GWAS summary data from complex traits (equivalent to TWAS from individual-level data; Online Methods and **Supplementary Fig. 5**). For comparison, we also measured power to detect significant SNP-trait associations through standard GWAS (testing each SNP individually) and eQTL-based GWAS (eGWAS; where the best eQTL in each gene is the only variant tested for association), with all three tests corrected for their genome-wide testing burdens. Using real genotype data, we simulated a causal SNP-expression-trait model with realistic effect sizes and measured the power of each strategy to identify genome-wide significant variants (accounting for 1 million SNPs for GWAS and 15,000 expressed genes using family-wise error rate control). Over many diverse disease architectures, TWAS substantially increased power when the expression-causing variants were untyped or poorly tagged by an individual SNP (**Fig. 5** and **Supplementary Figs. 6–11**). The greatest gains in power were observed in the case of multiple causal variants: 92% power for TWAS as compared to 18% and 25% power for GWAS and eGWAS, respectively. This scenario would correspond to expression caused by allelic heterogeneity[9,34,35], or 'apparent' heterogeneity, at common variants (due to tagging of an unobserved causal variant)[36]. TWAS was comparable to the other approaches when a single causal variant was directly typed, in which case combining the effects of neighboring SNPs does not add signal. Under the null hypothesis where expression was completely independent of phenotype (with either being heritable; scenarios A–D in **Fig. 2**), the TWAS false positive rate was well controlled (**Supplementary Table 4**). As expected, all methods were confounded in the case where the same causal variants had independent effects on traits and expression (scenarios F and G in **Fig. 2** and **Supplementary Figs. 8** and **12**).

Our approach can be conceptually viewed as a test for the correlation between the genetic component of expression and the genetic component of a trait (Online Methods). Because several recent methods have been proposed that measure genetic correlation between summary statistics[37], we sought to evaluate this relationship empirically. We compared TWAS to the recently proposed cross-trait LD score regression (LDSC) that estimates genome-wide genetic correlation between traits[37]. Although LDSC is not intended for local analyses because of model assumptions on polygenicity and use of block jackknife across loci for estimating standard errors, we performed the evaluation using expression and phenotype (height) data from the YFS cohort, using the results over individual data as the 'gold standard' (Online Methods). We found that the LDSC estimate of genetic correlation between height and expression from summary data was highly correlated with the gold standard (correlation = 0.7; **Supplementary Fig. 13**), but the relationship was much noisier than that of TWAS (correlation = 0.99; **Supplementary Figs. 5** and **13**). This suggests that TWAS attains more power than LDSC in relating expression to complex traits.



TWAS is also conceptually similar to a test for colocalization of signal between expression and a complex trait[38,39], and we compared it to a recently proposed method, COLOC[38], that evaluates colocalization of expression at known GWAS risk loci. After matching the false discovery rate of the two methods in simulations (Online Methods), TWAS and COLOC had similar power under the scenario with a single typed causal variant (with slightly lower COLOC power at small GWAS sizes), but TWAS had superior performance when the causal variant was untyped or in the presence of allelic heterogeneity (**Supplementary Fig. 10**). This is likely due the fact that TWAS explicitly models LD to better capture untyped variants.

Finally, we investigated the effect of the size of the expression reference panel on performance of TWAS (**Supplementary Fig. 9**). In general, TWAS always outperformed eGWAS when multiple variants were causal. Interestingly, power for either approach did not increase substantially beyond 1,000 expression samples, suggesting that the expression panels analyzed in this manuscript nearly saturate the

Figure 3 Accuracy of individual-level expression imputation algorithms. Adjusted accuracy was estimated using cross-validation R^2 between predicted and true expression and normalizing by corresponding $cis\text{-}h^2_g$. Bars show the mean estimate across three cohorts and three methods: eQTL, single best *cis*-eQTL in the locus; BLUP, using all SNPs in the locus; and BSLMM, using all SNPs in the locus and noninfinitesimal priors.

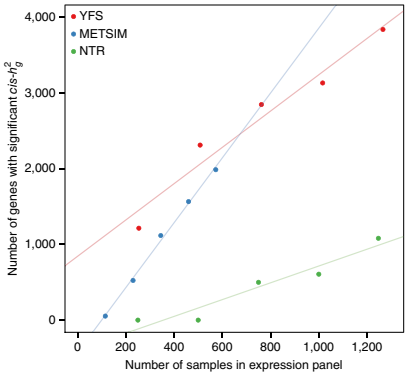


Figure 4 The number of genes with significant *cis* heritability observed at varying sample sizes. The number of genes with significant *cis* heritability was estimated by downsampling each data set (YFS, METSIM and NTR) into quintiles.

available imputation accuracy. This was further reflected in an analysis of real data, where merging expression data sets did not substantially change the distribution of TWAS statistics for the same gene set (**Supplementary Fig. 14**). Although these results come with caveats (for example, standard assumptions of additive effects and normal residuals), they suggest that the main benefit of larger expression reference panels is in increasing the total number of significant *cis*-heritable genes available for imputation (**Fig. 4**).

TWAS performance in GWAS summary data

To further validate our approach, we employed TWAS to identify expression-trait associations at the 697 known GWAS risk loci for height[7] using the YFS data for which height was also measured. At each locus, we considered three strategies for selecting a single causal gene: selection of (i) the gene nearest to the top GWAS SNP; (ii) the gene for which the index SNP was the strongest eQTL in the training data; and (iii) the most significant TWAS gene. For each strategy, we then constructed a risk score using the genetic value of expression for the selected genes and correlated the risk score with height measurements in the YFS individuals (an independent sample from the original height GWAS; **Supplementary Note**). R^2 between the risk score and height was 0.038 (nearest), 0.031 (eQTL) and 0.054 (TWAS), with the TWAS estimate significantly higher than the others in a joint model (Online Methods and **Supplementary Table 5**). When we recomputed the risk scores using TWAS values for expression from the NTR data (which introduces additional noise as a result of heterogeneity between the data sets), the TWAS estimate remained significantly higher than that from the eQTL strategy but was comparable to selecting the nearest gene (**Supplementary Table 5**). This indicates that using expression from a different study to select genes still significantly explains trait variance but is complementary (rather than superior) to selecting the nearest gene. Working from the assumption that genes with a higher *cis* genetic correlation to phenotype are more likely to be causal, these results motivate the use of TWAS to prioritize putative risk genes at known GWAS loci.

Across all known risk loci in our data, 77% of genome-wide significant loci (defined as lead SNP ±500 kb) overlapped at least one gene with significant *cis*-h^2_g and 36% overlapped at least one significant TWAS association (**Supplementary Table 6**). These results suggest that *cis* regulation of expression in blood and adipose tissue is an important mechanism through which genetic variation at known risk loci alters obesity-related traits. We expect that expression studies from other tissues relevant to obesity-related traits will further increase the overlap. Focusing specifically on the 282 TWAS-identified genes that were within 500 kb of the lead SNP, 187 (66%) were not the nearest gene, with many residing more than 100 kb away from the lead GWAS SNP (**Supplementary Fig. 15**). Because GWAS usually report the nearest gene, these 187 genes can be considered new candidates for follow-up at known risk loci. We note that gene-trait associations at known risk loci will not be found by TWAS if the causal mechanism does not involve *cis* expression of the tested genes or if there is insufficient power to identify and detect all *cis*-heritable genes at the locus.

Next, we employed TWAS to identify new expression-trait associations using summary association statistics from a 2010 lipid GWAS[17] (~100,000 samples), that is, associations more than 500 kb away from any genome-wide significant SNPs in that study. We used all three studies (METSIM, YFS and NTR) as separate SNP-expression training panels. We then looked for genome-wide significant SNPs at these loci in the larger 2013 lipid GWAS[5] (expanded to ~189,000 samples). We identified 25 such expression-trait associations in the 2010 study (**Supplementary Table 7**), of which 19 contained genome-wide significant SNPs in the 2013 study ($P = 1 \times 10^{-24}$ by hypergeometric test; Online Methods) and 24 contained a more significant SNP ($P = 1 \times 10^{-4}$), constituting a highly significant validation of the identified loci. The validation remained significant after conservatively accounting for sample overlap across the studies (binomial $P = 3 \times 10^{-16}$; Online Methods and **Supplementary Table 7**). As a sanity check, we compared

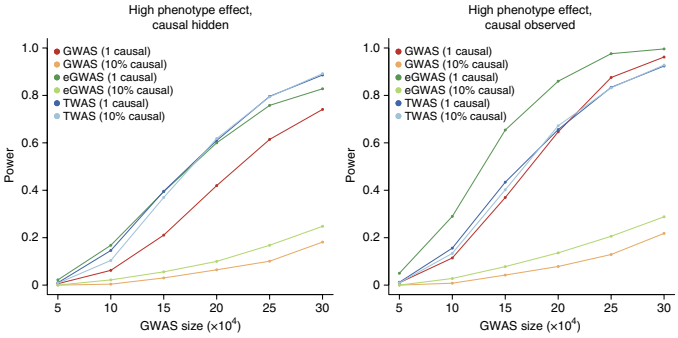


Figure 5 Power of summary-based expression imputation algorithms. Realistic disease architectures were simulated, and power to detect a genome-wide significant association was evaluated across three methods (accounting for 15,000 eGWAS or TWAS tests and 1 million GWAS tests). Colors correspond to the number of causal variants simulated and the methods used: GWAS where every SNP in the locus is tested; eGWAS where only the best *cis*-eQTL is tested; and TWAS computed using summary statistics. The expression reference panel was fixed at 1,000 out-of-sample individuals, and simulated GWAS sample size is designated on the *x* axis. Power was computed as the fraction of 500 simulations where significant association was identified.

direct and summary-level TWAS in the METSIM data and found the two sets of imputed expression-trait z scores to be nearly identical, with summary-level TWAS slightly underestimating the effect (Pearson $\rho = 0.96$; **Supplementary Fig. 16**). Overall, we find the TWAS approach to be highly predictive of robust phenotypic associations.

TWAS identifies new expression-trait associations

Having established the usefulness of TWAS, we applied the approach to identify new expression-trait associations using summary data from three recent GWAS over more than 900,000 phenotype measurements: lipid measures (high-density lipoprotein (HDL) cholesterol, low-density lipoprotein (LDL) cholesterol, total cholesterol (TC) and triglycerides (TG))[5], height[7] and BMI[6]. Significantly *cis*-heritable genes across the three expression data sets were tested individually (6,924 tests) and together in an omnibus test that accounts for predictor correlation (1,075 tests; Online Methods), and we conservatively corrected for the 8,000 total tests performed for each trait. Overall, we identified 665 significant gene-trait associations (**Supplementary Table 8**). Of these, 69 gene-trait associations did not overlap a genome-wide significant SNP in the corresponding GWAS, residing in 60 physically non-overlapping *cis* loci (**Table 1** and **Supplementary Table 9**). Averaging over the new genes, the z^2 statistics from TWAS were 1.5× higher than the strongest eQTL SNP for the same gene (although this may be slightly inflated because of winner's curse). Our previous simulations suggest that the substantial gain over testing the *cis*-eQTL is an indication of pervasive allelic heterogeneity[40] at these loci, and analyses of expression showed strong evidence for allelic heterogeneity at the TWAS genes (**Supplementary Fig. 17**).

We further sought to quantify the significance of the expression-trait associations conditional on the SNP-trait effects at the locus with a permutation test (Online Methods). Comparing to this null assesses how much signal is added by the expression given the specific GWAS architecture of the locus. For the 69 genes, this permutation test was significant for 54 (after accounting for 69 tests). After excluding these individually significant genes, the P values were still substantially elevated with λ_{GC} of 19 (ratio of the median χ^2 value to the expected null). For these 54 genes, we can confidently conclude that integration of expression data significantly refined the association with the trait. As before, more evidence of allelic heterogeneity in expression was observed at the loci that passed permutation (**Supplementary Fig. 17**). Our results are consistent with a model of causality where these genes harbor inherited causal variants that modulate expression, which in turn has a complex effect on the cell and downstream impact on complex traits[6].

Next, we evaluated the contribution to heritability of all expression-trait associations, including those that were not genome-wide significant (Online Methods)[29,30]. We estimated the variance in a trait explained by all METSIM and YFS imputed genes (h^2_{GE}) to be 3.4% averaged over six traits (**Supplementary Table 10**). We assumed independence of the two data sets and did not include the NTR genes because of the strong correlation of these data with YFS. Height had the most variance attributable to heritable genes at $h^2_{GE} = 7.1\%$. These combined estimates were consistently higher than those from a corresponding analysis using predictions from permuted expression (**Supplementary Table 10**). For the four traits with individual-level genotype and phenotype data in METSIM (BMI, TG, waist-hip ratio (WHR) and fasting insulin levels (INS)), we estimated h^2_{GE} directly using variance components over the imputed expression values (Online Methods). On average, all significantly heritable genes in adipose and blood explained 4–6% of the trait variance (16–19% of the total trait h^2_g) and were largely orthogonal between the two predictions (**Supplementary Table 11**). The imputed expression consistently explained more trait variance than the best *cis*-eQTL in each gene and did not strongly depend on the size of the *cis* window (**Supplementary Table 12**).

Reevaluation using other expression data sets

To replicate the 69 new expression-trait associations, we reevaluated the GWAS summary statistics with expression data from two external studies: eQTLs from ~900 samples in the MuTHER study[25] of fat, lymphoblastoid cell line (LCL) and skin cells and separate eQTL data from 5,311 samples in whole blood[11] (Online Methods). These expression studies only consist of summary-level associations and are expected to be much noisier as reference. In the relatively smaller MuTHER sample, 20 of 55 available genes replicated significantly in at least one tissue (after accounting for 55 tests; **Supplementary Table 9**). This is substantial given the apparent heterogeneity between expression data sets we previously observed (Online Methods). Notably, the correlations between discovery and replication z scores were strongest for associations found in the corresponding tissue ($\rho = 0.60$, $P = 1.5 \times 10^{-5}$ for blood and LCLs; $\rho = 0.66$, $P = 0.05$ for adipose; **Supplementary Table 13**), constituting significant aggregate replication and further evidence for the tissue-specific nature of our findings. Using the larger but heterogeneous training sample from ref. 11, 24 of 37 available genes replicated significantly (**Supplementary Table 9**). Although these replications are not strictly independent (they use the same GWAS data), they demonstrate that many of the newly identified loci are consistently significant across diverse expression cohorts.

Functional analysis of the new associations

To better understand their functional consequences, we evaluated the 69 new genes in the Hybrid Mouse Diversity Panel (HMDP) for correlation with multiple obesity-related traits. This panel includes 100 inbred mice strains with an extensive collection of obesity-related phenotypes from ~12,000 genes. Of the 69 new TWAS genes previously identified, 40 were present in the panel and could be evaluated for effect on phenotype. Of these, 26 were significantly associated with at least one obesity-related trait (after accounting for genes tested) and 14 remained significant after accounting for 36 phenotypes tested (very conservatively assuming that the phenotypes were independent) (**Supplementary Table 14**). Of the genes, 77% with an association were associated with multiple phenotypes. For example, expression of *Ftsj3* was significantly correlated with fat mass, glucose-to-insulin ratio and body weight in both liver and adipose tissue, with R^2 estimates ranging from 0.20 to 0.28. Another candidate, *Iih4*, was significantly correlated with LDL cholesterol and TC levels in liver. In humans, the corresponding gene is also linked to hypercholesterolemia in Online Mendelian Inheritance in Man (OMIM) and was previously associated with BMI in East Asians[41]. Because of complex correlation of phenotypes, it is difficult to assess whether this gene set is significant in aggregate and genes in the HMDP are typically expected to have strong effects. We could not perform enough random selections of genes to establish significance for this set. However, we consider the 26 individually significant genes to be fruitful targets for follow-up studies.

The BMI and height GWAS evaluated functional enrichment at identified loci, and we performed similar analyses for the new genes that we identified. We tested the ten new BMI-associated genes and 33 new height-associated genes for tissue-specific enrichment using DEPICT[42], a method based on large-scale gene coexpression analyses, following the protocol of the original GWAS[6,7]. Analysis of BMI identified significant enrichment for hypothalamus and neurosecretory systems ($P = 2.6 \times 10^{-4}$, significant at false discovery rate <5%). This enrichment is consistent with the landmark finding in the original

Table 1 TWAS significant genes with no known GWAS risk variants within 500 kb

GWAS	Training expression	Gene	Chr.	Locus start	Locus end	*P* value			
						TWAS	Permuted	Best *cis*-SNP	
LOCKE.BMI	Omnibus	*INO80E*	16	29,506,615	30,517,114	3×10^{-9}	3×10^{-7}	3×10^{-6}	*
LOCKE.BMI	Omnibus	*FTSJ3*	17	61,396,793	62,407,372	1×10^{-7}	5×10^{-6}	9×10^{-7}	*
LOCKE.BMI	Omnibus	*PAM*	5	101,589,685	102,866,809	3×10^{-7}	4×10^{-9}	3×10^{-3}	*
LOCKE.BMI	YFS	*GGNBP2*	17	34,400,737	35,446,278	1×10^{-6}	7×10^{-5}	3×10^{-6}	*
LOCKE.BMI	YFS	*MYO19*	17	34,351,477	35,399,284	3×10^{-6}	4×10^{-5}	3×10^{-6}	*
LOCKE.BMI	Omnibus	*OPRL1*	20	62,211,526	63,231,996	3×10^{-6}	9×10^{-5}	2×10^{-5}	*
LOCKE.BMI	NTR	*RABGAP1*	9	125,203,287	126,367,145	4×10^{-6}	3×10^{-5}	1×10^{-5}	*
LOCKE.BMI	YFS	*SMARCD2*	17	61,409,444	62,420,425	5×10^{-6}	9×10^{-5}	9×10^{-7}	*
LOCKE.BMI	METSIM	*AL049840.1*	14	103,677,607	104,679,149	5×10^{-6}	5×10^{-5}	1×10^{-7}	*
LOCKE.BMI	Omnibus	*LPAR2*	19	19,234,477	20,239,739	6×10^{-6}	3×10^{-5}	4×10^{-6}	*
WILLER.HDL	YFS	*MRPS18B*	6	30,085,486	31,094,172	1×10^{-7}	2×10^{-2}	4×10^{-6}	
WILLER.LDL	Omnibus	*PAM*	5	101,589,685	102,866,809	4×10^{-15}	9×10^{-12}	2×10^{-3}	*
WILLER.LDL	Omnibus	*ITIH4*	3	52,346,991	53,365,495	8×10^{-9}	4×10^{-6}	3×10^{-5}	*
WILLER.LDL	Omnibus	*WARS*	14	100,300,125	101,343,142	1×10^{-8}	6×10^{-7}	3×10^{-5}	*
WILLER.LDL	Omnibus	*MAN2C1*	15	75,148,133	76,160,971	1×10^{-8}	1×10^{-5}	6×10^{-5}	*
WILLER.LDL	YFS	*DHRS13*	17	26,724,799	27,730,089	6×10^{-7}	4×10^{-4}	2×10^{-6}	*
WILLER.LDL	YFS	*ERAL1*	17	26,681,956	27,688,085	8×10^{-7}	9×10^{-4}	2×10^{-6}	
WILLER.LDL	YFS	*HCG27*	6	30,665,537	31,671,745	2×10^{-6}	7×10^{-3}	7×10^{-8}	
WILLER.LDL	YFS	*VARS2*	6	30,376,019	31,394,236	3×10^{-6}	2×10^{-2}	1×10^{-5}	
WILLER.LDL	Omnibus	*PEX6*	6	42,431,608	43,446,958	5×10^{-6}	3×10^{-5}	4×10^{-4}	*
WILLER.LDL	Omnibus	*CSK*	15	74,574,398	75,595,539	6×10^{-6}	1×10^{-4}	1×10^{-5}	*
WILLER.TC	Omnibus	*PAM*	5	101,589,685	102,866,809	9×10^{-15}	3×10^{-13}	5×10^{-3}	*
WILLER.TC	Omnibus	*WARS*	14	100,300,125	101,343,142	2×10^{-8}	4×10^{-6}	2×10^{-5}	*
WILLER.TC	Omnibus	*MAN2C1*	15	75,148,133	76,160,971	3×10^{-7}	7×10^{-5}	2×10^{-6}	*
WILLER.TC	Omnibus	*ITIH4*	3	52,346,991	53,365,495	6×10^{-7}	2×10^{-5}	5×10^{-5}	*
WILLER.TC	NTR	*CDK2AP1*	12	123,245,552	124,256,687	6×10^{-7}	2×10^{-4}	5×10^{-6}	*
WILLER.TC	YFS	*TBKBP1*	17	45,271,447	46,289,416	9×10^{-7}	3×10^{-4}	2×10^{-7}	*
WILLER.TC	METSIM	*RPP25*	15	74,746,757	75,749,805	2×10^{-6}	2×10^{-4}	9×10^{-7}	*
WILLER.TC	Omnibus	*CSK*	15	74,574,398	75,595,539	2×10^{-6}	9×10^{-5}	9×10^{-7}	*
WILLER.TC	YFS	*MPI*	15	74,682,346	75,691,798	2×10^{-6}	2×10^{-4}	5×10^{-6}	*
WILLER.TC	Omnibus	*DAGLB*	7	5,948,757	7,023,821	2×10^{-6}	8×10^{-5}	7×10^{-6}	*
WILLER.TC	NTR	*TOM1*	22	35,195,267	36,243,985	2×10^{-6}	2×10^{-6}	1×10^{-6}	*
WILLER.TC	METSIM	*HMGXB4*	22	35,153,445	36,191,800	3×10^{-6}	4×10^{-5}	4×10^{-7}	*
WILLER.TC	NTR	*C17orf68*	17	7,628,139	8,651,413	3×10^{-6}	6×10^{-6}	2×10^{-7}	*
WILLER.TG	YFS	*PABPC4*	1	39,526,488	40,542,462	1×10^{-8}	1×10^{-4}	8×10^{-8}	*
WILLER.TG	Omnibus	*PACS1*	11	65,337,834	66,512,218	5×10^{-8}	3×10^{-4}	1×10^{-5}	*
WOOD.HEIGHT	Omnibus	*INO80E*	16	29,506,615	30,517,114	2×10^{-10}	2×10^{-5}	1×10^{-7}	*
WOOD.HEIGHT	NTR	*INPP5B*	1	37,825,435	38,912,729	2×10^{-9}	1×10^{-6}	1×10^{-6}	*
WOOD.HEIGHT	Omnibus	*MEGF9*	9	122,863,091	123,976,748	3×10^{-9}	2×10^{-4}	1×10^{-7}	*
WOOD.HEIGHT	Omnibus	*ATF1*	12	50,657,493	51,714,905	6×10^{-9}	4×10^{-5}	1×10^{-6}	*
WOOD.HEIGHT	Omnibus	*PAM*	5	101,589,685	102,866,809	2×10^{-8}	8×10^{-6}	1×10^{-5}	*
WOOD.HEIGHT	Omnibus	*CNIH4*	1	224,044,552	225,067,161	3×10^{-8}	2×10^{-5}	6×10^{-8}	*
WOOD.HEIGHT	Omnibus	*PLEKHA1*	10	123,634,212	124,691,867	1×10^{-7}	3×10^{-5}	6×10^{-7}	*
WOOD.HEIGHT	NTR	*PDXDC1*	16	14,568,832	15,632,186	1×10^{-7}	3×10^{-3}	7×10^{-6}	
WOOD.HEIGHT	YFS	*MSRB2*	10	22,884,435	23,910,942	2×10^{-7}	2×10^{-4}	3×10^{-6}	*
WOOD.HEIGHT	YFS	*ZNF213*	16	2,679,778	3,692,806	2×10^{-7}	1×10^{-3}	6×10^{-7}	
WOOD.HEIGHT	NTR	*YWHAB*	20	43,014,185	44,037,354	5×10^{-7}	3×10^{-4}	4×10^{-6}	*
WOOD.HEIGHT	NTR	*ITM2B*	13	48,307,273	49,336,451	5×10^{-7}	4×10^{-5}	1×10^{-6}	*
WOOD.HEIGHT	Omnibus	*WDSUB1*	2	159,592,304	160,643,310	7×10^{-7}	3×10^{-4}	5×10^{-6}	*
WOOD.HEIGHT	NTR	*STAT6*	12	56,989,190	58,005,129	8×10^{-7}	3×10^{-3}	8×10^{-6}	
WOOD.HEIGHT	Omnibus	*PLCL1*	2	198,169,426	199,937,305	9×10^{-7}	4×10^{-3}	6×10^{-7}	
WOOD.HEIGHT	YFS	*H2AFJ*	12	14,427,270	15,430,936	1×10^{-6}	4×10^{-4}	6×10^{-7}	*
WOOD.HEIGHT	YFS	*FAM8A1*	6	17,100,586	18,111,950	1×10^{-6}	2×10^{-3}	1×10^{-7}	
WOOD.HEIGHT	METSIM	*AC016995.3*	2	38,133,861	39,242,882	1×10^{-6}	3×10^{-4}	4×10^{-5}	*
WOOD.HEIGHT	Omnibus	*CDA*	1	20,415,441	21,445,401	1×10^{-6}	3×10^{-4}	6×10^{-7}	*
WOOD.HEIGHT	YFS	*ECHDC2*	1	52,861,656	53,892,884	1×10^{-6}	3×10^{-3}	1×10^{-5}	
WOOD.HEIGHT	YFS	*NFATC3*	16	67,618,654	68,763,162	2×10^{-6}	4×10^{-3}	4×10^{-7}	
WOOD.HEIGHT	YFS	*SH3YL1*	2	−282,270	766,398	2×10^{-6}	1×10^{-4}	6×10^{-7}	*
WOOD.HEIGHT	Omnibus	*PABPC4*	1	39,526,488	40,542,462	3×10^{-6}	4×10^{-4}	6×10^{-6}	*

(continued)

Table 1 (continued)

GWAS	Training expression	Gene	Chr.	Locus start	Locus end	*P* value			
						TWAS	Permuted	Best *cis*-SNP	
WOOD.HEIGHT	METSIM	*RP11-473M20.14*	16	2,666,043	3,684,883	3×10^{-6}	4×10^{-4}	1×10^{-7}	*
WOOD.HEIGHT	Omnibus	*HEBP1*	12	12,627,798	13,653,207	3×10^{-6}	7×10^{-4}	1×10^{-6}	*
WOOD.HEIGHT	YFS	*KBTBD2*	7	32,407,784	33,433,743	3×10^{-6}	4×10^{-3}	1×10^{-7}	
WOOD.HEIGHT	METSIM	*LRRC69*	8	91,614,060	92,731,464	3×10^{-6}	6×10^{-4}	6×10^{-7}	*
WOOD.HEIGHT	YFS	*RAB23*	6	56,553,607	57,587,078	4×10^{-6}	4×10^{-3}	6×10^{-7}	
WOOD.HEIGHT	YFS	*PPP4C*	16	29,587,299	30,596,698	5×10^{-6}	1×10^{-3}	1×10^{-7}	
WOOD.HEIGHT	NTR	*B3GALNT2*	1	235,110,442	236,167,884	5×10^{-6}	5×10^{-4}	1×10^{-6}	*
WOOD.HEIGHT	YFS	*PSRC1*	1	109,322,178	110,325,808	5×10^{-6}	3×10^{-4}	1×10^{-7}	*
WOOD.HEIGHT	YFS	*ACSS1*	20	24,486,868	25,539,616	5×10^{-6}	6×10^{-4}	1×10^{-6}	*
WOOD.HEIGHT	Omnibus	*GGPS1*	1	234,990,665	236,007,847	5×10^{-6}	4×10^{-4}	3×10^{-7}	*

*Significant after permutation. WILLER, LOCKE and WOOD correspond to GWAS data from refs. 5–7, respectively.

study[6] showing enrichment in these and other central nervous system tissues. Notably, we recapitulated this result using only new genes that did not overlap any genome-wide significant SNPs. In the analysis of height, DEPICT did not identify any tissue-specific enrichment.

DISCUSSION

In this work, we present methods that integrate genetic and transcriptional variation to identify genes with expression associated with complex traits. Using imputed gene expression to guide GWAS has three potential advantages. First, the gene is a more interpretable biological unit than an associated locus, which often contains multiple significant SNPs in LD that may not lie in genes and/or tag variants in multiple genes. Second, the lower total number of genes (or *cis*-heritable genes) means that the multiple-testing burden is substantially reduced relative to all SNPs. Lastly, combining *cis*-SNPs into a single predictor may capture heterogeneous signal better than individual SNPs or *cis*-eQTLs. Focusing prediction on the genetic component of expression also avoids confounding from environmental differences caused by the trait that may influence expression. Our approach builds upon the wealth of GWAS data in massive cohorts to directly implicate the gene-based mechanisms underlying complex traits.

Our proposed method has conceptual similarities with two-sample Mendelian randomization approaches that aim to identify causal relations between traits using genetic variation predictions as a randomizer[43–45]. However, whereas Mendelian randomization is intended to quantify the total causal effect, our method has the less strict goal of identifying significant associations and can operate on summary GWAS data. Notably, our approach maintains the attractive feature of not being confounded by effects on expression and a trait that are independent of the SNPs. Other recent work has proposed to leverage summary statistics to estimate the underlying genetic correlation between traits at the genome-wide level[37] but cannot be applied locally as it requires multiple loci to estimate standard errors (Online Methods). Recent work in parallel to ours also proposes gene expression imputation from individual-level data to find expression-trait associations and observes benefits from a reduced multiple-testing burden and increased interpretability[16]. In contrast, our approach does not require individual-level GWAS data and is applicable directly to GWAS summary data of very large sample sizes, thus increasing discovery power.

Unlike current methods, which focus on individually significant eQTL and SNP associations[5,6,8,9,11,13,26,29], our approach captures the full *cis*-SNP signal and does not require any individual marker to be significant. This is underscored by the fact that TWAS substantially outperformed its *cis*-eQTL analog both in imputing expression and in association with a trait. Our results show that the imputation approach is especially effective when multiple variants influence expression (which in turn influences a trait). The large number of new associations we identified in real data supports this phenomenon and suggests that it may be a strong contributor to common phenotypes[46]. Therefore, our approach can be seen as complementary to GWAS by identifying expression-trait associations that are not well explained by individual tagging SNPs. Future work could leverage the difference in performance of TWAS and GWAS to explicitly detect allelic heterogeneity. We note that it is still possible for some loci to have an independent SNP-phenotype and SNP-expression association driven by the same underlying variant, although we consider this to be an infrequent biological model.

We conclude with several limitations of our approach. First, variants influencing disease that are independent of *cis* expression—in general or in the training data—will not be identified. Second, as with any prediction, the number of genes that can be accurately imputed is still limited by the training cohort size and the quality of the training data. In particular, we found that prediction accuracy did not correspond with theoretical expectations and is likely driven by data quality. The impact of these weaknesses could be better quantified as expression data from larger sample sizes and a more diverse set of tissues become available. Although in this work we used both microarray and RNA sequencing as a measure of gene expression, thus showcasing the applicability of our approach to diverse data sets, the accuracy of our method intrinsically depends on the quality of the expression measurements. For the associated genes, it remains possible that the effect is actually mediated by phenotype (SNP → phenotype → *cis* expression; scenario F in **Fig. 2**). We attempted to quantify this in the YFS data by conditioning the heritability analyses on all the evaluated phenotypes (height, BMI and lipid concentrations) but observed no significant change at individual genes or in the mean *cis*-h^2_g. These results suggest that confounding from phenotype does not substantially affect the tested *cis* expression, although at the current sample size we cannot completely rule out such confounders for individual genes. An alternative confounder arises from independent effects on phenotype and expression at the same SNP or tag (**Fig. 2g** and Online Methods). Such instances could be indistinguishable from the desired causal model (Online Methods) without analyzing individual-level data, although we believe that they are still biologically interesting cases of colocalization. Both types of confounding could potentially be quantified by training the SNP-expression relationships in control individuals where phenotype is fixed or by interrogating the gene experimentally. Lastly, the summary-based TWAS cannot account for rare variants that are poorly captured by the LD reference panel or optimally capture nonlinear relationships between SNPs and expression. Additional sources of information could

potentially be incorporated to improve prediction, including significant *trans* associations[11,28], allele-specific expression[47,48], splice-QTLs affecting individual exons[10], haplotype effects and SNP-specific functional priors[20,49–51].

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. The predictor weights computed from the three expression studies as well as software to perform individual- and summary-level prediction are available at http://bogdan.bioinformatics.ucla.edu/software/twas/.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank the individuals who participated in the study. We also acknowledge L. Yang for helpful discussions that have improved the quality of this manuscript. We also thank K. Mohlke, M. Boehnke and F. Collins for help with the METSIM data. This work was funded in part by US National Institutes of Health (NIH) grants F32 GM106584 (A.G.), R01 GM053725 (B.P.), R01 GM105857 (A.L.P., A.G. and G.B.), HL-28481 (P.P., A.J.L. and M.C.) and HL-095056 (P.P. and B.P.) and by the US NIH training grant in Genomic Analysis and Interpretation T32 HG002536 (A.K.).

AUTHOR CONTRIBUTIONS

A.G. and B.P. conceived and designed the experiments. A.G., A.K. and H.S. performed the experiments and analyzed the data. G.B., W.C., B.W.J.H.P., R.J., E.J.C.d.G., D.I.B., F.A.W., P.F.S., E.N., M.A., M.C., A.J.L., T.L., E.R., M.K., I.S., O.T.R., J.K. and M.L. generated data, reagents, materials and analysis tools. A.G., A.L.P., P.P. and B.P. wrote the manuscript. All authors reviewed, revised and wrote feedback for the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Visscher, P.M., Brown, M.A., McCarthy, M.I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).
2. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375, S1–S3 (2012).
3. Lee, D., Bigdeli, T.B., Riley, B.P., Fanous, A.H. & Bacanu, S.A. DIST: direct imputation of summary statistics for unmeasured SNPs. *Bioinformatics* **29**, 2925–2927 (2013).
4. Pasaniuc, B. *et al.* Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics* **30**, 2906–2914 (2014).
5. Global Lipids Genetics Consortium. Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1283 (2013).
6. Locke, A.E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
7. Wood, A.R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46**, 1173–1186 (2014).
8. Musunuru, K. *et al.* From noncoding variant to phenotype via *SORT1* at the 1p13 cholesterol locus. *Nature* **466**, 714–719 (2010).
9. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
10. Zhang, X. *et al.* Identification of common genetic variants controlling transcript isoform variation in human whole blood. *Nat. Genet.* **47**, 345–352 (2015).
11. Westra, H.J. *et al.* Systematic identification of *trans* eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).
12. Albert, F.W. & Kruglyak, L. The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.* **16**, 197–212 (2015).
13. Raj, T. *et al.* Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. *Science* **344**, 519–523 (2014).
14. Letourneau, A. *et al.* Domains of genome-wide gene expression dysregulation in Down's syndrome. *Nature* **508**, 345–350 (2014).
15. Davis, L.K. *et al.* Partitioning the heritability of Tourette syndrome and obsessive compulsive disorder reveals differences in genetic architecture. *PLoS Genet.* **9**, e1003864 (2013).
16. Gamazon, E.R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).
17. Teslovich, T.M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
18. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
19. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
20. Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95**, 535–552 (2014).
21. Wray, N.R. *et al.* Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* **14**, 507–515 (2013).
22. Nuotio, J. *et al.* Cardiovascular risk factors in 2011 and secular trends since 2007: the Cardiovascular Risk in Young Finns Study. *Scand. J. Public Health* **42**, 563–571 (2014).
23. Raitakari, O.T. *et al.* Cohort profile: the cardiovascular risk in Young Finns Study. *Int. J. Epidemiol.* **37**, 1220–1226 (2008).
24. Wright, F.A. *et al.* Heritability and genomics of gene expression in peripheral blood. *Nat. Genet.* **46**, 430–437 (2014).
25. Grundberg, E. *et al.* Mapping *cis*- and *trans*-regulatory effects across multiple tissues in twins. *Nat. Genet.* **44**, 1084–1089 (2012).
26. Nicolae, D.L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* **6**, e1000888 (2010).
27. Torres, J.M. *et al.* Cross-tissue and tissue-specific eQTLs: partitioning the heritability of a complex trait. *Am. J. Hum. Genet.* **95**, 521–534 (2014).
28. Buil, A. *et al.* Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. *Nat. Genet.* **47**, 88–91 (2015).
29. Nica, A.C. *et al.* Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.* **6**, e1000895 (2010).
30. Robinson, G.K. That BLUP is a good thing: the estimation of random effects. *Stat. Sci.* **6**, 15–32 (1991).
31. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genet.* **9**, e1003264 (2013).
32. Dudbridge, F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet.* **9**, e1003348 (2013).
33. Chatterjee, N. *et al.* Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat. Genet.* **45**, 400–405, e1–e3 (2013).
34. Brown, C.D., Mangravite, L.M. & Engelhardt, B.E. Integrative modeling of eQTLs and *cis*-regulatory elements suggests mechanisms underlying cell type specificity of eQTLs. *PLoS Genet.* **9**, e1003649 (2013).
35. Wen, X., Luca, F. & Pique-Regi, R. Cross-population joint analysis of eQTLs: fine mapping and functional annotation. *PLoS Genet.* **11**, e1005176 (2015).
36. Wood, A.R. *et al.* Another explanation for apparent epistasis. *Nature* **514**, E3–E5 (2014).
37. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
38. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
39. Lee, D. *et al.* JEPEG: a summary statistics based tool for gene-level joint testing of functional variants. *Bioinformatics* **31**, 1176–1182 (2015).
40. Pritchard, J.K. & Cox, N.J. The allelic architecture of human disease genes: common disease–common variant...or not? *Hum. Mol. Genet.* **11**, 2417–2423 (2002).
41. Wen, W. *et al.* Meta-analysis of genome-wide association studies in East Asian–ancestry populations identifies four new loci for body mass index. *Hum. Mol. Genet.* **23**, 5492–5504 (2014).
42. Pers, T.H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6**, 5890 (2015).
43. Smith, G.D. & Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* **32**, 1–22 (2003).
44. Pickrell, J. Fulfilling the promise of Mendelian randomization. *bioRxiv* doi:10.1101/018150 (16 April 2015).
45. Pierce, B.L. & Burgess, S. Efficient design for Mendelian randomization studies: subsample and 2-sample instrumental variable estimators. *Am. J. Epidemiol.* **178**, 1177–1184 (2013).
46. Gusev, A. *et al.* Quantifying missing heritability at known GWAS loci. *PLoS Genet.* **9**, e1003993 (2013).
47. Dimas, A.S. *et al.* Common regulatory variation impacts gene expression in a cell type–dependent manner. *Science* **325**, 1246–1250 (2009).
48. Montgomery, S.B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773–777 (2010).
49. Pickrell, J.K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* **94**, 559–573 (2014).
50. Kichaev, G. *et al.* Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet.* **10**, e1004722 (2014).
51. Finucane, H.K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).

ONLINE METHODS

Data sets. In this study, we included 11,484 participants from two Finnish population cohorts, the Metabolic Syndrome in Men (METSIM; $n = 10{,}197$)[52,53] and the Young Finns Study (YFS; $n = 1{,}414$)[22,23]. 1,400 randomly selected individuals from the 10,197 METSIM participants underwent a subcutaneous abdominal adipose biopsy, of which 600 RNA samples were analyzed using RNA-seq (**Supplementary Note**). BMI, TG, WHR and INS were inverse rank transformed and adjusted for age and age^2. INS was additionally adjusted for T1D and T2D. 1,414 individuals (638 men with a median age of 43 years and 776 women with a median age of 43) with gene expression, phenotype and genotype data available were included in the blood expression analysis. Height, BMI, TG, TC, HDL and LDL were inverse rank transformed and adjusted for age, age^2 and sex. TC was also adjusted for statin intake. The biochemical lipid, glucose and other clinical and metabolic measurements of METSIM and YFS were performed as described previously[22,52,54]. Blood expression array data from the Netherlands Twins Registry (NTR; $n = 1{,}247$)[24,55] was processed as described in the original paper, followed by removal of any individuals with GRM values >0.05. Complete details on the pipeline and quality control procedures can be found in the **Supplementary Note**.

Heritability estimation with individual data. *Cis* and *trans* variance components were estimated using the REML algorithm implemented in GCTA[19]. As in previous studies, estimates were allowed to converge outside the expected 0–1 bound on variance to achieve unbiased mean estimates across all genes[24]. Standard error across gene sets was estimated by dividing the observed standard deviation by the square root of the number of genes that converged (this will lead to underestimation because of correlated genes but is presented for completeness). Genome-wide h^2_g values for the four traits in the GWAS cohort were estimated with GCTA from a single relatedness matrix constructed over all post–quality control SNPs in the strictly unrelated individuals. For estimating expression-wide h^2_{GE}, each predicted expression value was standardized to mean = 0 and variance = 1, and sample covariance across these values was used to define the relatedness matrix. The h^2_{GE} value was then estimated from this component with GCTA, with *P* values for difference from zero computed using a the likelihood-ratio test. Twenty principal components were always included as fixed effects to account for ancestry. Genetic correlation between traits in the GWAS cohort was estimated from all post–quality control SNPs in the full set of 10,000 individuals with GEMMA[31] (**Supplementary Table 15**). For YFS, we quantified the mediating effects of a trait on *cis* expression by separately re-estimating *cis*-h^2_g with all analyzed traits (height, BMI, TC, TG, HDL cholesterol and LDL cholesterol) included as fixed effects in addition to principal components. We did not observe significant differences in any individual gene (after accounting for 3,836 genes tested) nor in the mean estimate of *cis*-h^2_g.

Heritability estimation with summary data. As shown previously[51,56], for an association study of n independent samples, the expected χ^2 statistic is $E[\chi^2] = 1 + nlh^2_{GE}/m$, where l is the LD score accounting for correlation, m is the number of markers and h^2_{GE} is the variance in the trait explained by imputed expression. We estimated l directly from the genetic values of expression to be close to independence (1.4 and 1.5 for METSIM and YFS, respectively), allowing us to solve for h^2_{GE} from the observed distribution of χ^2 (or, asymptotically equivalent z^2) statistics. We did not compute this value for the BMI GWAS because the conservative multiple-GC correction applied in that study would yield a severe downward bias[6].

Imputing expression into genotyped samples. We evaluated three prediction schemes: (i) *cis*-eQTL, where the single most significantly associated SNP in the training set was used as the predictor; (ii) BLUP[30], which estimates the causal effect sizes of all SNPs in the locus jointly using a single variance component; and (iii) BSLMM[31], which estimates the underlying effect size distribution and then fits all SNPs in the locus jointly. For BLUP and BSLMM, prediction was done over all post–quality control SNPs using GEMMA[31]. We note that BLUP and BSLMM both perform shrinkage of the SNP weights but not variable selection, so all SNPs are included in the predictor. Recent work in parallel to ours also evaluated expression imputation using polygenic risk scores, LASSO and elastic net[16].

Evaluating prediction accuracy. Within-study prediction accuracy was measured by fivefold cross-validation in a random sampling of 1,000 of the highly heritable genes (genes with significant nonzero *cis* heritability) for each study. Cross-study prediction accuracy was measured by merging the YFS and NTR genotyped individuals and predicting from all individuals in one cohort into all individuals in the other cohort. In all instances, the R^2 between predicted and true expression across all predicted folds was used to evaluate accuracy (**Supplementary Fig. 18** and **Supplementary Note**).

Imputing expression into GWAS summary statistics. Summary-based imputation was performed using the ImpG-Summary algorithm[4] extended to train on the *cis* genetic component of expression. Let Z be a vector of standardized effect sizes (z scores) of SNP for a trait at a given *cis* locus (Wald statistics $\beta/\mathrm{se}(\beta)$). We impute the z score of the expression and trait as a linear combination of elements of Z with weights W (these weights are precompiled from the reference panel as $\Sigma_{e,s}\Sigma^{-1}_{s,s}$ for ImpG-Summary or directly from BSLMM). $\Sigma_{e,s}$ is the covariance matrix between all SNPs at the locus and gene expression, and $\Sigma_{s,s}$ is the covariance among all SNPs (LD). Under null data (no association) and a multivariate normal assumption, $Z \sim N(0,\Sigma_{s,s})$. It follows that the imputed z score of expression and trait (WZ) has variance $W\Sigma_{s,s}W^t$; therefore, we use $WZ/(W\Sigma_{s,s}W^t)^{1/2}$ as the imputation Z score of the *cis* genetic effect on the trait. In practice, for each gene, all SNPs within 1 Mb of the gene present in the GWAS were selected, and $\Sigma_{s,s}$ and $\Sigma_{e,s}$ and were computed in the reference panel (expression and SNP data). To account for finite sample size and instances where $\Sigma_{s,s}$ was not invertible, we adjusted the diagonal of the matrix using a technique similar to ridge regression with $\lambda = 0.1$ (as evaluated in Pasaniuc *et al.*[4]). This regularization, as well as noise in the estimation of W, can translate to lower power for association but yield conservative imputed Z statistics.

We used the YFS samples that were assayed for SNPs, phenotype and expression to assess the consistency of individual-level and summary-based TWAS. We first computed GWAS association statistics between phenotype (height) and SNPs and used them in conjunction with the expression data to impute summary-based TWAS statistics. The TWAS statistics were compared to those from the simple regression of (height ~expression) in the YFS data. We observed a correlation of 0.415 (**Supplementary Fig. 5**), consistent with an average *cis*-h^2_g of 0.17 ($\approx 0.415^2$) observed for these genes. When restricting to a regression of (height ~*cis* component of expression), we observed a correlation of 0.998 to the summary-based TWAS, demonstrating the equivalence of the two approaches when using in-sample LD.

Power analysis of the summary-based method. Simulations to evaluate the summary-based method were performed in 6,000 unrelated METSIM GWAS individuals. One hundred genes and the SNPs in the surrounding 1 Mb were randomly selected for testing. For each gene, normally distributed gene expression was simulated as $E = X\beta + \varepsilon$, where X is a matrix of the desired number of causal genotypes, sampled randomly from the locus; β is a vector of normally distributed effect sizes for each causal variant; and ε is a vector of normally distributed noise to achieve a *cis*-h^2_g value of 0.17 (corresponding to the mean observed in our significant gene sets). One thousand individuals with SNPs and simulated expression were then withheld for training the predictors (**Supplementary Fig. 19**). For the remaining 5,000 individuals, normally distributed noise was applied to expression to generate a heritable phenotype where expression explained 0.10/180 or 0.20/180 of the phenotypic variance (with the former corresponding to the average effect sizes for associated genes observed in a large GWAS of height[57] and the latter corresponding to high-effect loci). Association between SNPs and a phenotype was estimated in the 5,000 individuals (standard Z score), and phenotype generation was repeated with different environmental noise (up to 60 iterations) to generate results from multiple GWAS substudies. Association statistics from each run were then subjected to meta-analysis to reach precision corresponding to a larger GWAS of the desired size (up to 300,000) (**Supplementary Note**).

Detecting a locus was defined as follows. The single most significant trait associated SNP was taken as the GWAS association, considered detected if GWAS significance was $< 5 \times 10^{-8}$. The single most significant eQTL in the training set was taken as the eQTL-guided association (eGWAS), and considered detected if GWAS significance was <0.05/15,000. The TWAS association

was measured by training the imputation algorithm on the 1,000 held-out samples with expression and imputing into the GWAS summary statistics, and considered detected if significance was <0.05/15,000. The entire procedure was repeated 500 times (5 per gene) and power was estimated by counting the fraction of instances where each method detected the locus. As in the cross-validation analysis, training on the genetic component of expression instead of the overall expression consistently increased TWAS power by ~10% (**Supplementary Fig. 7**). Two null expression models were tested by generating gene expression for the 1,000 held-out samples that was standard normal as well as heritable expression (*cis*-$h_g^2 = 0.17$) with GWAS *Z* scores drawn from the standard normal (**Supplementary Table 4**). See the **Supplementary Note** for detailed simulation setup.

Power comparison to COLOC. COLOC uses summary data from eQTL and GWAS studies and a Bayesian framework to identify the subset of GWAS signals that co-localize with eQTLs. We sought to compare TWAS to the COLOC-estimated posterior probability of association (PPA) being shared for both phenotypes (PP4 in the COLOC implementation). COLOC additionally evaluates the hypothesis of multiple independent associations (PP3), but this is more general than the proposed TWAS model and was not tested. Because COLOC relies on priors of association to produce posterior probabilities of co-localization, we sought to identify a significance threshold that would make a fair comparison to the TWAS *P* value–based threshold. Specifically, we ran both methods on a realistic null expression simulation (with the generative model described previously): the expression was sampled from a null standard normal for 1,000 individuals and eQTLs computed; the trait associations were derived from a simulated 300,000 GWAS with a single typed causal variant that explained 0.001 variance of the trait (high effect). We believe this scenario is both realistic and consistent with the GWAS assumptions of COLOC. We then empirically identified the statistical threshold for COLOC and TWAS that would yield a 5% false discovery rate: co-localization statistic PP4 > 0.17 for COLOC, and $P < 0.05$ for TWAS. We note that this empirical COLOC threshold is much less stringent than PP4 > 0.8 used in the COLOC paper (PP4 > 0.8 would yield lower power for COLOC in our simulations). These thresholds were subsequently used to evaluate the power to detect an expression-trait association in simulations with a true effect (**Supplementary Figs. 10** and **12**). The reported power is for a single locus, and we did not attempt to quantify genome/transcriptome-wide significance.

Individual-level analysis of METSIM GWAS. We imputed the significantly heritable genes into the METSIM GWAS cohort of 5,500 unrelated individuals with individual-level genotypes (and unmeasured expression). We then tested the imputed expression for obesity-related traits: body mass index (BMI), triglycerides (TG), waist-hip-ratio (WHR) and fasting insulin levels (INS). Overall, the evaluated traits exhibited high phenotypic and genetic correlation as well as highly significant genome-wide h_g^2 ranging from 23–36% (**Supplementary Table 15**), consistent with common variants having a major contribution to disease risk[1]. Association was assessed using standard regression as well as a mixed-model that accounted for relatedness and phenotypic correlation[31], with similar results. The effective number of tests for each trait was estimated by permuting the phenotypes 10,000 times and, for each permutation, re-running the association analysis on all predicted genes. For each trait, P_{perm}, the *P*-value in the lowest 0.05 of the distribution, was computed and the effective number of tests was $0.05/P_{perm}$ (reported in **Supplementary Tables 16** and **17**). All phenotypes were shuffled together, so any phenotypic correlation was preserved. The effective number of tests corresponded to 88–95% of the total number of genes, indicating a small amount of statistical redundancy (**Supplementary Note**). To evaluate the TWAS approach, we computed phenotype association statistics for the 5,500 unrelated individuals and re-ran the analysis using only these summary statistics and the same expression reference panels. The resulting TWAS associations were nearly identical to the direct TWAS associations across the four traits (Pearson $\rho = 0.96$). Reassuringly, the TWAS was generally more conservative than the direct estimates (**Supplementary Fig. 16**).

Refining trait-associated genes at known loci. We focused on GWAS data from height[7] that identified 697 genome-wide significant variants in 423 loci, and conducted the summary-based TWAS over all genes in these loci using YFS and NTR as expression training data. Because the YFS individuals had been phenotyped for height and not tested in the GWAS, we could directly evaluate whether selected genes were associated with phenotype. At each locus, we considered three strategies for selecting a single causal gene: (i) the gene nearest to the most significantly associated SNP; (ii) the gene for which the index SNP is the strongest eQTL in the training data; (iii) the most significant TWAS gene. For each strategy, we then constructed a risk-score using the genetic value of expression for the selected gene weighted by the corresponding TWAS *Z*-score. The same procedure was then re-evaluated using TWAS values trained in the NTR cohort (which introduces additional noise due to heterogeneity between the cohorts, **Supplementary Table 5**). We separately used GCTA to estimate the heritability of height explained by all of the genes selected by each algorithm by constructing a GRM from the selected genes. In contrast to the risk score, this does not assume predefined weights on each gene but allows them to be fit by the REML model. Results were comparable, with only the TWAS-selected genes explaining significantly nonzero heritability (**Supplementary Table 5**).

Validation analysis in lipid GWAS data. We evaluated the performance of TWAS by identifying significantly associated genes in the 2010 lipid study that did not overlap a genome-wide significant SNP, and looking for newly genome-wide significant SNPs in the expanded 2013 study. The *P*-value for the number of genes with increased significance and genome-wide significance in the 2013 study was computed by a hypergeometric test, with background probabilities estimated from the set of significantly heritable genes. Of the genes not overlapping a significant locus in the 2010 study, 70% had a more significant SNP in the 2013 study, and 3.5% overlapped a genome-wide significant SNP ($P < 5 \times 10^{-8}$).

Meta-analysis of imputed expression from multiple tissues. We proposed a novel omnibus test for significant association across predictions from all three cohorts. Because the imputation is made into the same GWAS cohort, correlation between predictors must be accounted for. For each gene *i*, we estimated a correlation matrix $\mathbf{C}_i$ by predicting from the three tissues into the ~5,500 unrelated METSIM GWAS individuals (though any large panel from the study population could be used). This correlation includes both the genetic correlation of expression as well as any correlated error in the predictors, thus capturing all redundancy. On average, a correlation of 0.01, 0.01 and 0.43 was observed between YFS:METSIM, NTR:METSIM, and YFS:NTR, highlighting the same tissue of origin the last pair. We then used the three-entry vector of TWAS predictions, Z_i, to compute the statistic $\mathrm{omnibus}_i = \mathrm{Z}_i' \, \mathbf{C}_i^{-1} \, \mathrm{Z}_i$ which is approximately χ^2 (3-dof) distributed and provides an omnibus test for effect in any tissue while accounting for correlation[58,59]. Though the correlation observed in our data was almost entirely driven by the YFS:NTR blood data sets, we expect this to be an especially useful strategy for future studies with many correlated tissues. An alternative approach would be to perform traditional meta-analysis across the three cohorts and then predict the TWAS effect. However, this would lose power when true eQTL effect sizes (or LD) differ across the cohorts, which we have empirically observed to be the case by looking at predictor correlations. The proposed omnibus test aggregates different effects across the studies, at the cost of additional degrees of freedom.

Gene permutation test. The standard TWAS *Z* score is a test against the null of no SNP-trait association; that is, $Z_{\mathrm{TWAS}} = WZ/(W\Sigma_{s,s}W')^{1/2}$ is well calibrated (has a mean of 0 and unit variance) only under the null model of $Z \sim N(0,\Sigma_{s,s})$. In the alternate model where *Z* is drawn from a nonzero mean distribution[60,61], Z_{TWAS} has a distribution that depends both on *Z* and the weights *W*. To quantify the impact of the weights on Z_{TWAS} regardless of whether *Z* is null or non-null, we conduct permutations conditional on the observed *Z* vector. For each gene, the expression labels were randomly shuffled, and the summary-based TWAS analysis was trained on the resulting expression to compute a permuted new null for Z_{TWAS}. Testing against this permuted null distribution is equivalent to testing for an expression-trait association (or genetic correlation between expression and a trait) conditional on the observed GWAS statistics at the locus (which may not be drawn from the null of no association).

The permutation test empirically computes this distribution of Z_{TWAS} values conditional on the observed Z and asks how extreme the observed Z_{TWAS} is among all possible W values coming from permuted expression data. Note that failing the permutation test may be an indication of lack of power to show that the expression significantly refines the direct SNP-trait signal. In practice, the permutation test was run 1,000 times for each TWAS gene, and a P value was computed by Z test against this null.

Relationship to genetic covariance/correlation. Our tests relate to previously defined estimators of genetic correlation and covariance between traits. We consider two definitions of genetic covariance at a locus: (i) the covariance between the genetic component of expression and the genetic component of a trait and (ii) the covariance between the causal effect sizes for expression and the causal effect sizes for a trait. Under assumptions of independent effect sizes, these definitions yield asymptotically identical quantities[37]. Assuming a substantially large training set where the genetic component of expression can be perfectly predicted, the direct TWAS tests for a significant association between the genetic component of expression and the trait—equivalent to testing definition (i) for a polygenic trait. Likewise, the summary-based TWAS tests for a significant sum of products of the causal expression effect sizes and the causal trait effect sizes—equivalent to definition (ii) up to a scaling factor. The TWAS approach therefore fits naturally with the broader study of shared genetic etiology of multiple phenotypes. At the sample sizes evaluated in this study, the TWAS approach is substantially better powered than an LD-based estimate of local genetic correlation (**Supplementary Note**).

52. Stancáková, 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).
53. Stancáková, A. *et al.* Changes in insulin sensitivity and insulin release in relation to glycemia and glucose tolerance in 6,414 Finnish men. *Diabetes* **58**, 1212–1221 (2009).
54. Turchin, M.C. *et al.* Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nat. Genet.* **44**, 1015–1019 (2012).
55. Boomsma, D.I. *et al.* Netherlands Twin Register: from twins to twin families. *Twin Res. Hum. Genet.* **9**, 849–857 (2006).
56. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
57. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).
58. Bolormaa, S. *et al.* A multi-trait, meta-analysis for detecting pleiotropic polymorphisms for stature, fatness and reproduction in beef cattle. *PLoS Genet.* **10**, e1004198 (2014).
59. Zhu, X. *et al.* Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. *Am. J. Hum. Genet.* **96**, 21–36 (2015).
60. Zaitlen, N., Paşaniuc, B., Gur, T., Ziv, E. & Halperin, E. Leveraging genetic variability across populations for the identification of causal variants. *Am. J. Hum. Genet.* **86**, 23–33 (2010).
61. Han, B., Kang, H.M. & Eskin, E. Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. *PLoS Genet.* **5**, e1000456 (2009).

Chapter 4

ASElux: an ultra-fast and accurate allelic reads counter

Bioinformatics, 34(8), 2018, 1313–1320
doi: 10.1093/bioinformatics/btx762
Advance Access Publication Date: 23 November 2017
Original Paper

OXFORD

Gene expression

ASElux: an ultra-fast and accurate allelic reads counter

Zong Miao[1,2], Marcus Alvarez[1], Päivi Pajukanta[1,2,3] and Arthur Ko[1,3,*]

[1]Department of Human Genetics, David Geffen School of Medicine at UCLA, [2]Bioinformatics Interdepartmental Program and [3]Molecular Biology Institute, UCLA, Los Angeles, CA 90024, USA

*To whom correspondence should be addressed.
Associate Editor: Ziv Bar-Joseph

Received on August 17, 2017; revised on October 25, 2017; editorial decision on November 18, 2017; accepted on November 22, 2017

Abstract

Motivation: Mapping bias causes preferential alignment to the reference allele, forming a major obstacle in allele-specific expression (ASE) analysis. The existing methods, such as simulation and SNP-aware alignment, are either inaccurate or relatively slow. To fast and accurately count allelic reads for ASE analysis, we developed a novel approach, ASElux, which utilizes the personal SNP information and counts allelic reads directly from unmapped RNA-sequence (RNA-seq) data. ASElux significantly reduces runtime by disregarding reads outside single nucleotide polymorphisms (SNPs) during the alignment.

Results: When compared to other tools on simulated and experimental data, ASElux achieves a higher accuracy on ASE estimation than non-SNP-aware aligners and requires a much shorter time than the benchmark SNP-aware aligner, GSNAP with just a slight loss in performance. ASElux can process 40 million read-pairs from an RNA-sequence (RNA-seq) sample and count allelic reads within 10 min, which is comparable to directly counting the allelic reads from alignments based on other tools. Furthermore, processing an RNA-seq sample using ASElux in conjunction with a general aligner, such as STAR, is more accurate and still $\sim$4$\times$ faster than STAR + WASP, and $\sim$33$\times$ faster than the lead SNP-aware aligner, GSNAP, making ASElux ideal for ASE analysis of large-scale transcriptomic studies. We applied ASElux to 273 lung RNA-seq samples from GTEx and identified a splice-QTL rs11078928 in lung which explains the mechanism underlying an asthma GWAS SNP rs11078927. Thus, our analysis demonstrated ASE as a highly powerful complementary tool to cis-expression quantitative trait locus (eQTL) analysis.

Availability and implementation: The software can be downloaded from https://github.com/abl0719/ASElux.

Contact: zmiao@ucla.edu or a5ko@ucla.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Allele specific expression (ASE) denotes the preferential allelic expression of a gene in the diploid genome. Integrating ASE with expression quantitative trait locus (eQTL) analysis improves fine-mapping accuracy and sensitivity (Kumasaka *et al.*, 2015), thus helping identify biologically meaningful regulatory signals such as imprinting and cis regulation. Although several methods have been developed to identify ASE events from RNA-sequencing (RNA-seq) data (Castel *et al.*, 2015; León-novelo *et al.*, 2014; Liu *et al.*, 2014; Li *et al*, 2012), mapping bias remains a major obstacle in ASE analysis (Degner *et al.*, 2009; Panousis *et al.*, 2014; Stevenson *et al.*, 2013). Therefore, there is an important scientific knowledge gap that motivates the development of fast and accurate allelic expression analysis tools.

Previously, simulations have been used to identify variant sites showing bias towards one allele (Buil *et al.*, 2015). However,

simulations perform sub-optimally in practice since they are largely based on single-end reads whereas most RNA-seq data are now paired-end reads. There are also methods that utilize available genotype information and build personal allelic reference genomes for an allele-aware alignment that are implemented in programs such as SNP-o-matic (Manske and Kwiatkowski, 2009) and GSNAP (Wu and Nacu, 2010). In these approaches, the aligners are aware of single nucleotide polymorphisms (SNP) and align reads against both alleles. Even though the SNP-aware methods are more accurate than simulation-based approaches, they are more time consuming and computationally intensive, which makes them impractical for large RNA-seq datasets. A recently developed allele-specific analysis method (WASP) (van de Geijn *et al.*, 2015) substitutes the SNP base with the alternative genotype in allelic reads and re-aligns those reads to correct for the reference bias. By excluding the allelic reads that are affected by different genotypes, WASP obtains extremely low false positive rate when identifying ASE SNPs. However, the process of generating reads with alternative genotypes in WASP takes a relatively long time (∼3.5 h) and many reads are excluded due to its stringent requirements.

To this end, we developed a new and more efficient approach, ASElux, which focuses on SNP-overlapping reads and combines the alignment and estimation of allelic expression into one step. Since accurate genotyping is essential for ASE analysis, the genotype information is usually obtained separately from the RNA-seq data using SNP array or genome/exome sequencing (Lonsdale *et al.*, 2013). ASElux builds a personal allelic reference genome by using the individual's existing genotype information to generate all possible ASE reads and pre-screen the RNA-seq data. This allows us to perform SNP-aware alignment and to efficiently identify only the reads that cover the unique set of SNPs present in each individual. Compared to all of the tested tools, ASElux is ultra-fast while achieving the closest allelic mapping accuracy to the benchmark SNP-aware aligner, GSNAP. Adding the time consumption to analyze an RNA-seq sample using a general-purpose aligner, such as STAR, the overall runtime of ASE analysis using ASElux is still ∼4 times faster than STAR (Dobin *et al.*, 2013) followed by WASP (STAR + WASP), which re-aligns the reads with SNPs to decrease the reference bias (van de Geijn *et al.*, 2015). We applied ASElux to 273 lung transcriptomes from the Genotype-Tissue Expression Project (GTEx) (Lonsdale *et al.*, 2013) to demonstrate the increased power of ASE analysis in detecting local gene regulation. The high speed and accuracy of this novel ASE software makes it possible to analyze ASE in large datasets, helping efficient transformative interrogation of variants.

2 Materials and methods

2.1 Workflow of ASElux

Since only ∼10% of sequencing reads can be identified as SNP-overlapping, ASElux saves time by focusing on aligning reads that overlap with an individual's SNPs obtained either from a genotype array, imputed SNPs based on a reference panel, or DNA sequencing. To implement this new alignment, we designed a hybrid index system that performs both genome-wide alignment and personal SNP-aware alignment (Supplementary Fig. S1A). The hybrid index system contains a static index that is built once for each reference genome. ASElux aggregates the genic regions in the reference genome to form a trimmed genome and uses a suffix array (Nong *et al.*, 2009; Manber and Myers, 1990) as the static index for a fast alignment. The other part of our hybrid index system is the dynamic index. We extract the flanking sequence on both sides of the exonic SNP and store that in the dynamic index. The dynamic index is generated before alignment and it takes only ∼3 min to build it for each individual. Supplementary Figure S1B shows the workflow of aligning paired-end reads. For a pair of reads, we follow the workflow twice to treat each read first as the main read and then as the mate read. To accommodate sequencing errors, ASElux by default allows up to two mismatches elsewhere than at the SNP site. The user can set the number of allowed mismatches to fit various read lengths. We first use the dynamic index to identify the allelic reads during the alignment. Only the reads that match the dynamic index would be mapped to the genome with the static index to locate their multi-alignment loci. Then we try to align the other read, known as the mate read, near the identified multi-alignment loci. Thus, we only align the mate read if the main read matches the dynamic index. If both reads are uniquely aligned to one gene, we count the reads for the allele they originated from.

2.2 Filtering candidate SNPs

Since exonic reads can provide the best estimation of gene expression, ASElux disregards non-exonic SNPs and alignments for ASE analysis. Within ASElux, we provide a fast and useful tool to select exonic SNPs using genome annotation. A standard genome annotation contains overlapped exons and transcripts due to alternative splicing, and the overlapping information is redundant for pruning SNPs. To facilitate the pruning process, we merge all overlapping exons from different transcripts within the same gene into one. Small indels are another mechanism of allelic expression, but they tend to cause an alignment error leading to bias in ASE estimation (Heap *et al.*, 2010; Stevenson *et al.*, 2013). Thus, most ASE analyses focus on SNPs alone rather than the combination of SNPs and indels for the better accuracy (David *et al.*, 2017). Therefore, we load the SNP and indel information from the vcf file and disregard all SNPs within one read length of an indel. The distance allowed between SNPs and indels varies according to the read length of the particular set of RNA-seq data. For example, if the read length is 50 bp, all SNPs within 50 bp of any indels in each individual would be disregarded from the further alignment. As shown in the 20 GTEx samples, only ∼0.9% of the SNPs were excluded by this process (Supplementary Table S1).

2.3 Hybrid index system

To perform a personalized SNP-aware alignment and maintain a high speed, we designed a hybrid index system that contains both static and dynamic indices. The static indices are built only once for each reference genome. Since only a small proportion of RNA-seq reads consist of intergenic reads (Mortazavi *et al.*, 2008), ASElux uses the genic regions as the reference genome to achieve the least compromised balance between the alignment accuracy and speed and relatively low memory usage. We locate the start and the end of each gene so that the sequence between them covers all the components of a gene (exons, introns, UTRs etc.). Then we aggregate the sequences of all genes to form a trimmed genome. In the human genome (hg19), ASElux generates a new genome that contains ∼1.5 billion bp out of the ∼3 billion bp. For a genome that contains N genes, we construct N suffix arrays for the N genes and 1 more general suffix array for the trimmed genome as the static index using the sais algorithm (Nong *et al.*, 2011). Although in theory searching globally in one suffix array is faster than using N suffix arrays for N genes, in practice combining local and global indices is faster due to the low-level memory management strategy in a modern computer

(Kim *et al.*, 2015). Briefly, the static index is built based on the trimmed reference genome and accordingly, the global alignment is not allele-specific. The suffix array indices of the trimmed genome and genes costs ∼30 GB of RAM (10 bytes for each base). We only use ∼15 GB with the trimmed genome, thus keeping our overall RAM usage at ∼ 20GB.

For each individual, we build a personalized dynamic index for SNP-aware alignment. We first prune the non-exonic SNPs to make sure that ASElux focuses on aligning only the expression-related reads. For each exonic SNP, we extract N-1 bp flanking sequence on both sides of the exonic SNP from the reference transcripts, where N is the read length, and replace the allele at the SNP location to generate reference sequences for all possible exonic reads that overlap with the SNP. To cover the SNPs adjacent to various splicing junctions, we extract the SNP flanking regions from all transcripts in each gene. Thus, each SNP has two 2N-1 bp long sequences for the reference and alternative alleles from each transcript. If the individual has additional known SNPs within the flanking sequence, we generate all possible haplotypes with alternative alleles of these adjacent SNPs to avoid misaligning reads with multiple variants. As there are regions with extremely high SNP density, ASElux only counts the first 10 heterozygous SNPs in each read. Noteworthy, as most indices are unique, we do not expect ambiguous indices to substantially bias the alignment of the ASE reads. To quickly locate SNP-overlapping reads, we aggregate all of the generated sequences as the dynamic index and build a suffix array for it. Then we save the generated sequences, SNPs and gene names for the dynamic index to query.

2.4 Alignment

Aligning only to the SNP-overlapping regions of the genome to identify the allelic reads is the key to the high speed of ASElux. For paired-end reads, we treat one read as the main read and the other as the mate read to help alignment. As shown in Supplementary Figure S1B and Algorithm 1 of Supplementary Methods, we check if the main read can be identified as an allelic read and use the mate read to properly align the whole read fragment. Only the ASE reads that are aligned to the dynamic index with up to two mismatches by default (not counting the SNP locus) will be aligned against the static index built on the trimmed genome (global alignment) to identify all of the multi-alignment loci. During the local alignment step, ASElux tries to locally align each main read's mate read to the static index of the same gene. Thus, both the global alignment and the local alignment are against the static index. Since the read fragment should come from the same gene, we require the read mates to be aligned to the same gene. In the case where the major read is multi-aligned, we count the major read towards the ASE estimate only if both the main and mate reads are aligned to the same gene and at least one of them is uniquely aligned. Finally, we exchange the roles of the main and mate read and align the main read again to identify all possible alignments for the read pair. Thus, each read is treated once as the main read of the paired end reads.

2.4.1 Alignment against the dynamic index

Similarly to STAR, ASElux uses a binary search strategy to identify the Maximal Mappable Prefix (MMP) of a read in a suffix array index. The alignment of the main read starts from the left end of the read and identifies the longest common sequence with the dynamic index. Since the suffix array is built in the forward genome direction, we also align the reverse complement of the read to cover both directions. Algorithm 2 in the Supplementary Methods shows the process of aligning reads against the dynamic index. For the reads with mismatches, the alignment process stops at the mismatched locus and restarts at the base after the mismatched locus. Thus, several regions divided by the mismatched loci in the main reads are aligned to different loci in the dynamic index. For the regions aligned by no less than 20 bp, ASElux compares the whole read against the sequence around the mapped loci to check if the main read can be aligned to the locus with no more than 2 mismatches (using the ASElux default setting) while not counting indels (Supplementary Fig. S2). Since ASElux aligns the main read while being aware of the individual's SNP loci, the mismatches are mainly caused by sequencing errors or unknown adjacent variants. Furthermore, we calculated that for a 100-bp read, allowing for up to 2 mismatches (using the default setting) covers 99.985% of the reads with the typical sequencing error rate of 0.1% per base expected for the Illumina platform (Schirmer *et al.*, 2016). Although ASElux allows 2 mismatches for ASE reads by default, users can adjust the number of allowed mismatches to fit for the various read lengths.

2.4.2 Local alignment

Using the static index, ASElux aligns the mate read against the same gene region that the main read aligns to. Therefore, the reads without mismatches, indels, or splice junctions are perfectly mapped to the reference genome in this step. Supplementary Figure S3 shows an example of aligning a junction read. For reads that are not identical to the reference genome, the MMP is a substring of the read that stops before a variant or splice site. As shown in Algorithm 3 of the Supplementary Methods, we skip eight bases to avoid mapping indels or SNPs and search the MMP again for the unmapped part of the read. We chose to skip 8 bp in line with STAR because in practice most indels would safely be skipped with this set-up and it still allows us to utilize the remaining read for alignment. Separate MMPs of a read indicate that mismatches or splicing occurs between MMPs. We repeatedly search for the MMP until all parts of the read are mapped or we have searched more than the default of four times, indicating that the read should not be mapped to the reference due to too many mismatches or splicing loci. We selected the default of four times since it provided the best balance between the alignment accuracy and speed. After identifying all MMPs, we reassemble the read and only accept the read alignment if the read was properly reconstructed such that the MMPs are in the same order in the query read and the reference.

2.4.3 Global alignment

The global alignment is similar to the local alignment but extends to the trimmed genome in the static index. Hence, the MMP can originate from multiple local indices, indicating that the read is aligned to multiple genes. Since the lengths of the perfectly aligned prefixes in multi-aligned reads vary and searching for the MMP requires a perfect alignment, the multi-aligned reads will only be aligned to the locus that has the longest prefix shared with the reference genome. Thus, if we only align a read once to the trimmed genome, the multi-aligned loci that have the shorter perfectly aligned prefixes would be missed. Since it is crucial to find all possible multi-alignment targets for the ASE reads, we developed a masked binary search strategy to align the read to additional possible loci by masking off the known alignment results (Algorithm 4 of the Supplementary Methods). To globally fast align the ASE read, we utilize the fact that the information about the one perfectly mapped locus is available for the ASE reads. To find all possible genes where

the read may be mapped to, we skip the locus that the read is already aligned to when searching for the MMP for the read. Since smaller MMPs have too many matches to the trimmed genome by chance alone, we only use MMPs longer than 20 bases and record the genes they reside in. Then we locally align the main read and the mate read in those genes to finish the alignment. ASElux can repeatedly align the reads with more and more masked genes. Therefore, the loci with smaller MMPs will not be missed due to the existence of the other loci with longer MMPs. In more detail, to find all MMPs within the read, we will start from the beginning of the read and search for the longest shared sequence between the particular read and the trimmed genome. We move along the read to find all MMPs longer than 20 bp in the read. Accordingly, there can be several MMPs which all must be longer than 21 bp. After the global alignment, we still locally align the mate read (Supplementary Fig. S1B), which ensures that a locus with only 20 bp match will not be identified as a properly aligned locus. As the next step, since the static index contains no SNP information, we align not only the ASE read but also the read that resides in the same locus with a different genotype. ASElux combines the alignment results of the two reads to make sure that we have the most comprehensive multi-alignment result.

The details of alignment with existing methods as well as the simulation data are described in the Supplementary Methods.

2.5 ASE and splice-QTL analyses in the GTEx project

We processed 273 RNA-seq samples from the GTEx project (Lonsdale *et al.*, 2013) with ASElux. We downloaded the RNA-seq data and the imputed genotype data from the dbGaP accession phs000424.v6.p1. We randomly selected 20 samples for the comparisons in this study. The samples have on average 40 million 50-bp paired-end reads. Reads were aligned to the human genome (hg19) with the four tested aligners. We used the default alignment parameters of all the tested methods. The uniquely aligned reads were then kept for the subsequent analyses.

The results of the cis-eQTL analysis (version 6) were obtained from the GTEx portal (Ardlie *et al.*, 2015). For each individual, we pruned out all SNPs aligned with less than 30 reads or less than 6 reads from one allele. To identify ASE SNPs across the population, we picked all SNPs that were heterozygous and passed the read count threshold in at least 30 individuals. We performed a paired *t*-test with the read counts of the reference allele and alternative allele from all individuals. The SNPs with Bonferroni corrected *P*-values less than 0.05 were identified as ASE SNPs. The GWAS SNPs ($P \leq 5 \times 10^{-8}$, two-sided) were obtained from the NHGRI GWAS Catalog (Welter *et al.*, 2014). We calculated the linkage disequilibrium (LD) between the ASE SNPs and GWAS SNPs within 1 Mb distance and obtained all of the SNPs in LD ($R2 \geq 0.8$) with the ASE SNPs using PLINK (Purcell *et al.*, 2007). Then we annotated the SNPs in LD with the ASE SNPs using ANNOVAR (Wang *et al.*, 2010).

For the splice-QTL analysis, we aligned the 273 GTEx lung samples with STAR and identified all the splice events using LeafCutter. Following the analytical guideline of LeafCutter, we then used MatrixeQTL (Shabalin, 2012) to identify whether rs11078928 is a significant splice-QTL of GSDMB. For the isoform level eQTL analysis, we used RSEM (Li and Dewey, 2011) to estimate the isoform expression of the 273 GTEx lung samples and calculated the proportional transcript expression as the transcript expression level over the total gene expression as the phenotype in the eQTL analysis performed by MatrixeQTL (Shabalin, 2012).

3 Results

3.1 Test on simulated RNA-seq dataset

We first tested ASElux and other alignment methods on a simulated RNA-seq dataset with ~180 M 2× 50 bp paired-end reads (see Supplementary Methods). Since comprehensively testing the alignment bias is important, we generated a high coverage simulated dataset. SNPs and junction reads were introduced to mimic real RNA-seq data. We added alternative alleles to the simulated reads based on imputed genotypes from a random GTEx sample and set both alleles to be equally expressed, which allowed us to accurately calculate the mapping bias of all methods. Besides ASElux, we also tested STAR 2.4.2a (Dobin *et al.*, 2013), GSNAP 2015-6-23 (Wu and Nacu, 2010), HISAT2 2.0.4 (Kim *et al.*, 2015) and WASP (van de Geijn *et al.*, 2015) on the simulated dataset using the default parameters during the alignment (see Supplementary Methods). Since we focus on the alignment bias, we only tested the mapping function of WASP. We used the reference genome hg19 for all aligners and the GENCODE v19 annotation if the gene annotation could be supplied. To utilize the power of SNP-aware alignment, we used GSNAP to build a SNP-integrated alignment index for GSNAP. The HISAT2 alignment index was downloaded from its website along with the SNPs and transcript information. We used the default parameters for each aligner.

Using the genome-wide SNP data (genotyped and imputed) from GTEx (Lonsdale *et al.*, 2013), we calculated read counts of each allele at exonic SNP sites to estimate ASE. The proportion of reference allele read counts when compared to the total read counts indicates the imbalance of allelic expression. Since the two alleles were equally expressed in the simulated dataset, the expected RACR of each SNP is 0.5. Accordingly, we measured the reference bias as the deviation of RACR from 0.5. Since each method aligns allelic reads differently, we performed the reference bias analysis using SNPs with enough aligned reads in all methods. ASElux, GSNAP, STAR and HISAT2 uniquely aligned ~10 M allelic reads whereas WASP aligned ~24% less reads than the other tested methods using the same simulated dataset (Supplementary Table S2). Figure 1 shows the proportion of SNPs in different bias categories. Although the majority of the SNPs displayed a bias less than 5% using all methods, ASElux achieved the highest accuracy by properly accounting for allele imbalance for ~90% of the SNPs. Among the biased SNPs identified by each method, ASElux and the SNP-aware GSNAP showed substantially fewer SNPs with reference allele bias (~70%) when compared to HISAT2 and STAR (~99%). Even though STAR + WASP identified the fewest SNPs with a bias more than 5%, still the majority (88%) of the SNPs identified by WASP showed a bias in the range of more than 0% but less than 5% (Fig. 1). STAR alone performed worst since no SNP information was used during the alignment. Even though HISAT2 considers all common SNPs (MAF > 1%), it performs better than STAR but not as well as WASP, GSNAP and ASElux.

To test the ability of identifying ASE SNPs by ASElux and other methods, we generated another simulated dataset with 20% of genes exhibiting imbalanced allelic expression. These imbalanced genes were randomly selected and one random allele from the selected genes was overexpressed. Compared to the less expressed allele, we generated 1.5–3.5× more reads from the overexpressed allele. To mimic real RNA-seq data, we introduced sequencing error in addition to SNPs and junction reads. To ensure a 50× coverage for each allele, we overexpress one allele by generating more reads when simulating the imbalanced allelic expression. Using the binomial test, we identified a SNP as an ASE SNP if the Bonferroni corrected

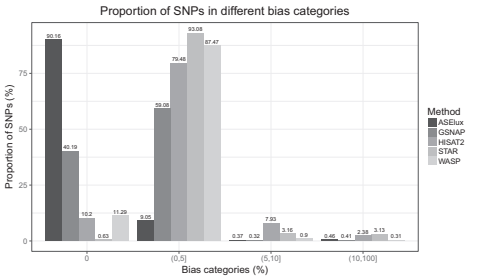


Fig. 1. Proportions of SNPs in different bias categories show that ASElux performs better than general RNA-seq aligners. Y axis shows the proportion of SNPs, and X axis shows the different bias categories. The bias is the absolute difference between the predicted proportion of the reference allele reads and the real proportion of reference allele reads (0.5)

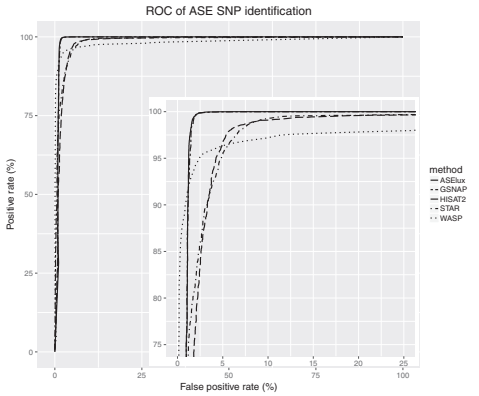


Fig. 2. The receiver operating characteristic (ROC) of ASE SNP identification shows that ASElux performs as well as GSNAP, and outperforms HISAT2 and STAR in a simulated dataset. The X axis is the false positive rate and the Y axis is the positive rate

P-value is less than 0.05. To ensure that the tested methods are fairly compared, we used the intersection of the SNPs identified by all of the tested methods. The receiver operating characteristic (ROC) curve (Fig. 2) indicates that ASElux outperforms HISAT2 and STAR alone on identifying ASE SNPs. Since GSNAP and ASElux both utilize personal SNP information, they identified all of the ASE SNPs (true positive rate = 100%) while maintaining a low false positive rate of ∼5%. Although ASElux performed better than GSNAP in the first simulation test (Fig. 1), ASElux and GSNAP showed a comparable number of SNPs showing more than 5% bias. Thus, GSNAP and ASElux performed similarly based on the ROC curve (Fig. 2). The false positive rate of WASP is the lowest among all the tested methods while the true positive is below 92%. In the first simulation test under the null condition (Fig. 1), WASP showed the smallest number of SNPs that have bias more than 5%, suggesting that WASP tends to be highly conservative in order to achieve a low false positive rate. However, since WASP filters out potentially falsely aligned reads by STAR, some SNPs might have insufficient coverage to pass a stringent threshold, which may contribute to the low positive rate of WASP.

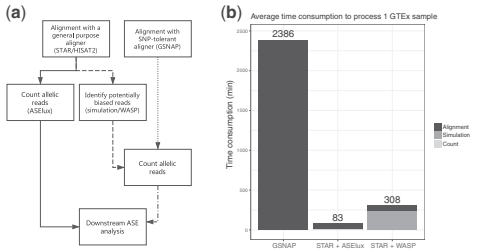


Fig. 3. ASElux is faster than the other tested methods (WASP and GSNAP). (**a**) The workflow of ASE analysis using different programs. (**b**) The estimated average time consumptions to count allelic reads from 10 GTEx RNA-seq samples. The X axis shows the tested method. The Y axis is the time needed for processing the dataset

3.2 Speed benchmarks

We performed the speed benchmark on a server with 64-bit Intel CPUs @2.66 GHz with ∼95GB RAM. Figure 3A shows the common workflow of ASE analysis using different methods. Researchers can first map reads using a RNA-seq aligner (STAR/HISAT2) and then count allelic reads with specialized tools, such as ASElux or WASP, or alternatively use a SNP-aware aligner (GSNAP) and count allelic reads directly based on the alignment. Figure 3B shows the average time consumption of a single thread to perform ASE analysis on 10 samples from the GTEx project using STAR + ASElux, STAR + WASP and GSNAP, respectively. Among the tested methods, GSNAP used ∼12 GB RAM, HISAT2 ∼8 GB RAM and ASElux ∼22 GB of RAM, respectively. WASP itself requires no more than 1GB of RAM but the actual RAM requirement of WASP depends on the alignment tool it uses, e.g. STAR would need additional ∼30 GB RAM. Counting allelic reads with ASElux is, however, ultra-fast since it only takes ∼20 min to process a GTEx RNA-seq sample. Therefore, STAR + ASElux can has a ∼33× faster processing speed than GSNAP. WASP requires ∼4 CPU hours for each GTEx sample, which makes STAR + WASP ∼4× slower than STAR + ASElux. The tests shown in Figure 3 were based on single thread mode. ASElux, HISAT2, STAR and GSNAP all have a multi-thread mode, however, WASP does not support multi-thread computing. Thus, we also tested the multithread mode of each tool except for WASP (Supplementary Fig. S4), which resulted in similar relative alignment speeds across the tools as in the single thread mode. As I/O often plays a significant factor in runtime, the system cache was cleared before each alignment run to avoid any bias due to pre-loaded reference index in the memory during the benchmarking. On average, index loading contributes up to 25% of the overall runtime of ASElux without caching.

3.3 Comparing GSNAP, ASElux, HISAT2 and STAR on 20 experimental samples

To evaluate the performance of ASElux, GSNAP, STAR, HISAT2 and WASP on real RNA-seq data, we processed 20 lung RNA-seq samples from the GTEx (Lonsdale *et al.*, 2013) cohort with the five methods. Each sample contains ∼40 million pairs of 76 bp reads. The genotype data of each sample consists of the imputed and genome-wide SNP array data from the GTEx study. For each sample, ∼120 000 exonic SNPs were obtained from the VCF file. We

built a personalized index for each sample for ASElux (see methods) and GSNAP, and provided the same indices as in the simulated analysis to STAR and HISAT2. After alignment, we extracted the allelic read counts on each heterozygous SNP for further analyses.

The level of imbalanced allelic expression represented by the RACR provides more information on ASE than the statistics by the binomial test. In an ASE analysis, the proportion of reference allelic reads closer to 0 or 1 often indicates stronger allele-specific gene expression. Therefore, we compared the allelic imbalance (AIB), which is the difference between 0.5 and RACR, derived by ASElux to AIBs by the other methods. Under the null hypothesis that most SNPs will not have an ASE effect, we expect equal expression from both the reference and alternative haplotypes. Consequently, the theoretical distribution of AIB should be centered at zero with a few outliers towards the two tails. If the reference bias hampers the alignment, the mean and median of AIB of all SNPs would shift up from 0, which is shown in Figure 4. GSNAP shows a minimal reference bias in the test. Although ASElux shows a higher average AIB when compared to GSNAP (Fig. 4), its average AIB is significantly lower than the AIBs obtained using WASP, HISAT2 and STAR. WASP, HISAT2 and STAR aligned significantly more reads to the reference allele, indicating a higher reference bias. Although WASP showed the lowest false positive rate in the simulation test, the majority of the WASP SNPs still had a bias more than 0% and less than 5%, which is similar to HISAT2 (Fig. 1). The AIBs derived from the 20 GTEx samples confirmed this similarity between WASP and HISAT2.

ASElux uniquely aligned ∼1.3 M allelic reads for each sample; whereas WASP aligned ∼1.5 M allelic reads; GSNAP and HISAT2∼1.7 M allelic reads; and STAR ∼2.8 M allelic reads for each sample, respectively (Supplementary Fig. S5a, Table S2). ASElux identified ∼15% fewer SNPs than GSNAP but ∼37% more than WASP (Supplementary Fig. S5b). It is worth noting, however, that not all SNPs identified by STAR and HISAT2 are suitable for downstream ASE analysis. Previous studies show that more than 10% of the heterozygous SNPs would be excluded when employing a simulation procedure to correct for the reference alignment bias, (Kukurba *et al.*, 2014; Panousis *et al.*, 2014) while using a general purpose aligner. Thus, overall ASElux would identify a similar number of heterozygous SNPs that are suitable for the downstream ASE analysis when compared to STAR, and HISAT2.

Although STAR uniquely aligned more reads than the other tested tools, it identified a similar number of SNPs with a coverage of ≥30 reads when compared to HISAT2 and GSNAP (Supplementary Table S3, Fig. S4b). The extra allelic reads aligned by STAR mainly overlap with the low coverage SNPs that do not contribute to the ASE analysis (Supplementary Fig. S6). Since WASP depends on STAR for the alignment, a large amount of reads in WASP also overlap with the low coverage SNPs (Supplementary Fig. S6). Thus, WASP identified less SNPs than the other tested tools with similar number of reads aligned (Supplementary Fig. S5).

3.4 ASE analysis strengthens cis-eQTL analysis in identifying local regulation of gene expression

Utilizing the ultra-fast speed of ASElux, we applied ASElux to a dataset of 273 lung RNA-seq samples and imputed SNP array data from the GTEx study. Figure 5 shows that ASElux has significantly less reference bias when compared to the allelic read counts reported by GTEx using their ASE analysis pipeline (Ardlie *et al.*, 2015) (P-value$< 2.2 \times 10^{-16}$, two-sided t-test). The distribution of RACR from ASElux is centered at 0.5 whereas the distribution reported by GTEx displays an upward bias. To verify whether ASElux has identified enough heterozygous SNPs for the ASE analysis, we compared the number of SNPs identified by ASElux and the GTEx study. In both analysis by ASElux and the GTEx study a heterozygous SNP must be covered by ≥30 reads to be counted for the downstream ASE analysis. A median of 6385 SNPs passed the simulation correction in the GTEx study (Panousis *et al.*, 2014) for each sample which is ∼20% less than the median SNP number identified by ASElux in the 273 GTEx lung samples, indicating that ASElux can identify more ASE SNPs than the GTEx ASE protocol.

In addition to ASE analysis, a cis-eQTL analysis is also widely used to detect allele-specific regulation of gene expression. We compared the ASE results from ASElux to the cis-eQTL results on the same set of SNPs publicly available at the GTEx website. Among the 273 lung samples we aligned with ASElux, we identified 21 550 heterozygous exonic SNPs covered by at least 30 reads in no less than 30 samples. Using a paired t-test, we identified 2765 SNPs residing in 1790 genes that showed ASE ($P < 2.32 \times 10^{-6}$, two-sided). Although not all ASE events are caused by exonic SNPs, the paired

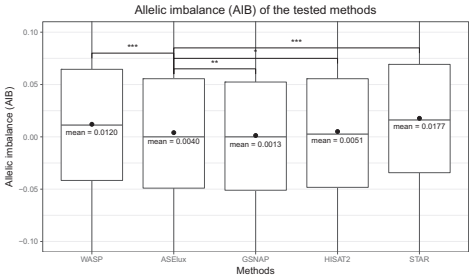


Fig. 4. For ASE analysis, ASElux has less reference bias than WASP, and the general aligners, STAR and HISAT, when testing the 20 real RNA-seq samples. * indicates a P-value of 1.05×10^{-3} (two-sided); ** indicates a P-value of 7.44×10^{-14} (two-sided); and *** indicates a P-value $< 2.2 \times 10^{-16}$ (two-sided). Y axis displays the allele frequency differences between the tested methods and 0.5 (i.e. the two alleles are expressed equally). The reference bias of ASElux is significantly smaller than that of HISAT2, STAR and WASP, but higher than GSNAP. Y axis was limited from -0.1 to 0.1 to show the distribution of most SNPs. The red dots indicate the mean values of each method

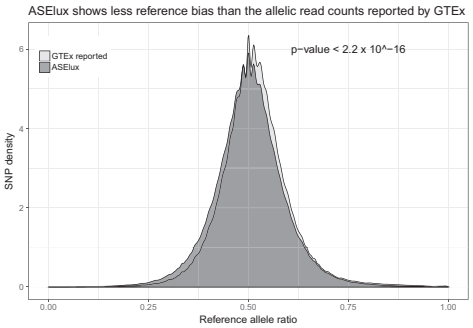


Fig. 5. Compared to the allelic read counts reported by GTEx, ASElux shows less reference bias in 273 lung samples (P-value $< 2.2 \times 10^{-16}$, two-sided). The X axis shows the reference allele ratio, and the y axis shows the density of all SNPs

t-test of the exonic SNPs should identify either the causal exonic ASE SNPs or the exonic SNPs tagged by non-coding causal variants. Next, we investigated whether these 2765 ASE SNPs were also identified as cis-eQTLs by GTEx. Using the gene-specific permutation threshold used by GTEx (Ardlie *et al.*, 2015), 1421 of the ASE SNPs were significant cis-eQTLs in lung for the genes they are located in. Overall, 1344 (48.61%) of the ASE SNPs were missed by the cis-eQTL analysis, and 965 (53.91%) of the ASE genes were not identified as cis-eQTL genes. Accordingly, the combination of ASE and cis-eQTL analysis increased the power to identify variants associated with local regulation of gene expression when compared to cis-eQTL analysis alone.

To further investigate the association between ASE and lung disorders, we calculated the linkage disequilibrium (LD) between ASE SNPs and 67 GWAS SNPs (Ardlie *et al.*, 2015) of lung disorders, such as smoking; asthma; lung cancer; chronic obstructive pulmonary disease (COPD); and pulmonary hypertension. There are 11 ASE SNPs in strong LD ($r2 > 0.8$) with the GWAS SNPs (Supplementary Table S4). Of the 11 ASE SNPs, 10 are identified as cis-eQTLs of the genes in which they reside. Both the cis-eQTL and ASE analysis indicate that the alternative genotype of the 10 ASE SNPs is associated with a lower gene expression. It is worth noting, however, that the ASE SNP rs2305480 located in Gasdermin B (GSDMB) is in LD ($R2 = 1$) with a GWAS SNP rs11078927 which is associated with the increased risk of asthma (Bouzigon *et al.*, 2008; Bønnelykke *et al.*, 2014). This SNP rs11078927 has never been identified as a significant cis-eQTL of GSDMB in the lung tissue before. Moreover, rs2305480 has also been identified as a GWAS SNP of another inflammatory disorder, ulcerative colitis (McGovern *et al.*, 2010), supporting the role of the GSDMB gene in several disorders with a known inflammatory component.

We further investigated the potential mechanism of the GWAS SNP rs11078927 and discovered rs11078928, which is a splice donor site variant previously identified in the whole blood and suggested to be involved in asthma (Morrison *et al.*, 2013). It is in tight LD ($R2 = 0.99$) with two asthma GWAS hits, rs2305480 (the ASE SNP) and rs11078927. To examine the splicing effect in a human tissue highly relevant for asthma, we performed a splice-QTL analysis in 273 GTEx lung RNA-seq samples using LeafCutter and identified rs11078928 as a significant splice-QTL of GSDMB in the lung (Fig. 6). The genotype of rs11078928 is significantly associated (P-value $= 4.63 \times 10^{-35}$, two-sided) with the proportional expression level of the junction reads overlapping exon 5 and exon 6 of GSDMB, which is consistent with the splicing event identified previously in the whole blood (Morrison *et al.*, 2013).

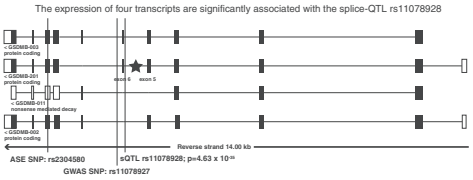


Fig. 6. Using the proportional transcript expression as the phenotype, four transcripts GSDMB-003, GSDMB-201, GSDMB-011 and GSDMB-002 are significantly associated with the genotype of the splice-QTL SNP, rs11078928 (P-value $< 9.43 \times 10^{-4}$, two-sided). The stars indicate that genotypes of rs11078928 are significantly associated with the splice junction reads between the exon 5 and exon 6 of GSDMB (P-value $= 4.63 \times 10^{-35}$, two-sided)

To determine which isoform expression of GSDMB is impacted by the splice variant, we used RSEM (Li and Dewey, 2011) to estimate the expression of isoforms in 273 GTEx lung samples and used the proportional transcript expression as the phenotype for an isoform eQTL analysis. The relative expression of four transcripts GSDMB-003, GSDMB-201, GSDMB-011 and GSDMB-002 are significantly associated with the genotype of the splice-QTL rs11078928 [P-value $< 9.43 \times 10^{-4}$ using linear regression via MatrixeQTL (Shabalin, 2012) with Bonferroni correction] (Fig. 6, Supplementary Table S5). Thus, the biological mechanism underlying the asthma risk SNPs, rs2305480 and rs11078927, is likely mediated by the SNP rs11078928 via splicing regulation on GSDMB in the human lungs.

We further functionally annotated the 52 460 SNPs ($R2 > 0.8$) tagged by the ASE SNPs, identified by ASElux, using ANNOVAR (Wang *et al.*, 2010) (Supplementary Fig. S7). There are 19 additional SNPs identified as splice variants by ANNOVAR and 7 of them were missed by the GTEx cis-eQTL analysis. Taken together, an ASE analysis provides substantially more power for analysis of local gene expression, complementing the regular cis-eQTL analysis.

4 Discussion

With growing interest in ASE analysis, mapping bias remains a critical barrier that hinders the accuracy of ASE analysis in RNA-seq. We provide a novel approach, ASElux that focuses solely on SNP-overlapping reads, allowing a fast and accurate SNP-aware alignment for ASE analysis. To ensure a high alignment accuracy, we used the whole gene body (50% of the reference genome) to build the alignment index. It is worth noting that this speed gain is largely due to the fact that ASElux first aligns all reads to the very small dynamic index to identify the allelic reads and then only aligns them to the large static index. The size of the static index will not affect the speed substantially because the time complexity of searching through suffix array is O(mlog(n)), where the n is the size of the reference and m is the size of the pattern. In addition, ASElux shows a minimal reference bias when compared with other methods based on both simulated and experimental RNA-seq data. ASElux aligns against both alleles by employing personal dynamic indices to minimize the reference bias. We demonstrated that ASElux works optimally with short reads currently generated by most RNA-seq studies.

Due to the complexity of RNA-seq alignment and variable expression of genes across tissues, SNP-calling from RNA-seq is often less accurate than from DNA-sequencing data (Quinn *et al.*, 2013). Thus, external genotype information from whole exome sequencing (WES), whole genome sequencing (WGS), or SNP-arrays are preferred for ASE or eQTL analysis (Ardlie *et al.*, 2015). ASElux and all of the tools tested here do not directly identify SNPs from RNA-seq reads and are therefore only applicable to RNA-seq cohorts that have genotype data available. Simultaneously calling SNPs and ASE from RNA-seq data will enable ASE analyses in additional RNA-seq cohorts, but it will require development of new methods in the future.

Multi-alignment also presents a serious challenge in ASE analysis. Reads generated from different regions might be falsely identified as ASE reads due to their similar sequences. ASElux tries to find all possible multi-alignment loci in addition to the optimal alignment even if the read has the best alignment quality as an ASE read to stringently remove possible false ASE reads. As ambiguously aligned reads are more stringently excluded, ASElux tends to align

less allelic reads than the other tested tools. However, not all SNPs are reliable for the ASE analysis due to the reference alignment bias when using a general-purpose aligner such as STAR and HISAT2, and in fact, the previous studies show a ∼10% loss in the number of SNPs during the simulation correction (Kukurba *et al.*, 2014; Panousis *et al.*, 2014). We have shown here that the high accuracy of ASElux has provided more reliable SNPs for the downstream ASE analysis than STAR did in the analyzed GTEx lung samples.

As an alignment tool exclusively designed for ASE analysis, ASElux outperforms most existing methods in speed and provides a better accuracy than the existing non-SNP-aware aligners for correcting the reference bias in alignment while also achieving the closest accuracy to GSNAP. ASElux is ultra-fast: it is able to process 40 million 2× 50 bp reads in 16 min. Combined with a general purpose aligner, such as STAR, STAR + ASElux is ∼33 times faster than the golden standard SNP-aware aligner GSNAP, and ∼4 times faster than the popular combination of STAR + WASP. The high speed and accuracy make ASElux an ideal tool to perform ASE analysis in large-scale RNA-seq studies. We demonstrated the usefulness of ASElux by performing the ASE analysis in lung RNA-seq data from 273 individuals of the GTEx project in two days (∼70 CPU hours using multi-CPU). By comparing the ASE SNPs and eQTLs from the same dataset, we also demonstrated that the combination of ASE and cis-eQTL analysis provides more power to detect local regulation of gene expression.

Acknowledgements

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 08/14/2016 and dbGaP accession number phs000424.v6.p1 on 08/11/2016.

Funding

This work was supported by the National Institutes of Health (NIH) [grant numbers HL-095056, HL-28481]. A.K. was supported by the NIH [grant number F31HL127921] and M.A. was supported by the NIH [grant number T32HG002536].

Conflict of Interest: none declared.

References

Ardlie,K.G. *et al.* (2015) The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, **348**, 648–660.

Bønnelykke,K. *et al.* (2014) A genome-wide association study identifies CDHR3 as a susceptibility locus for early childhood asthma with severe exacerbations. *Nat. Genet.*, **46**, 51–55.

Bouzigon,E. *et al.* (2008) Effect of 17q21 variants and smoking exposure in early-onset asthma. *N. Engl. J. Med.*, **359**, 1985–1994.

Buil,A. *et al.* (2015) Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. *Nat. Genet.*, **47**, 88–91.

Castel,S.E. *et al.* (2015) Tools and best practices for allelic expression analysis. *Genome Biol.*, **16**, 195.

David,A.K. *et al.* (2017) Allele-specific expression reveals interactions between genetic variation and environment. *Nat. Methods*, **14**, 699–702.

Degner,J.F. *et al.* (2009) Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, **25**, 3207–3212.

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

Heap,G.A. *et al.* (2010) Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Hum. Mol. Genet.*, **19**, 122–134.

Kim,D. *et al.* (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357–360.

Kukurba,K. *et al.* (2014) Allelic expression of deleterious protein-coding variants across human tissues. *PLoS Genet.*, **10**, e1004304.

Kumasaka,N. *et al.* (2015) Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat. Genet.*, **48**, 206–213.

León-Novelo,L.G. *et al.* (2014) A flexible Bayesian method for detecting allelic imbalance in RNA-seq data. *BMC Genomics*, **15**, 920.

Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.

Li,G. *et al.* (2012) Identification of allele-specific alternative mRNA processing via transcriptome sequencing. *Nucleic Acids Res.*, **40**, 1–13.

Liu,Z. *et al.* (2014) Comparing computational methods for identification of allele-specific expression based on next generation sequencing data. *Genet. Epidemiol.*, **38**, 591–598.

Lonsdale,J. *et al.* (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.

Manber,U. and Myers,G. (1990) Suffix string arrays: a new searches method for on-line. *Proc. first Annu. ACM-SIAM Symp. Discret. Algorithms*, 319–327.

Manske,H.M. and Kwiatkowski,D.P. (2009) SNP-o-matic. *Bioinformatics*, **25**, 2434–2435.

McGovern,D. *et al.* (2010) Genome-wide association identifies multiple ulcerative colitis susceptibility loci. *Nat. Genet.*, **42**, 332–337.

Morrison,F.S. *et al.* (2013) The splice site variant rs11078928 may be associated with a genotype-dependent alteration in expression of GSDMB transcripts. *BMC Genomics*, **14**, 627.

Mortazavi,A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.

Nong,G. *et al.* (2009) Linear suffix array construction by almost pure induced-sorting. In: 2009 Data Compression Conference, pp. 193–202.

Nong,G. *et al.* (2011) Two efficient algorithms for linear time suffix array construction. *IEEE Trans. Comput.*, **60**, 1471–1484.

Panousis,N.I. *et al.* (2014) Allelic mapping bias in RNA-sequencing is not a major confounder in eQTL studies. **15**, 467.

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

Quinn,E.M. *et al.* (2013) Development of strategies for SNP detection in RNA-Seq data: application to lymphoblastoid cell lines and evaluation using 1000 genomes data. *PLoS One*, **8**, e58815.

Schirmer,M. *et al.* (2016) Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics*, **17**, 125.

Shabalin,A.A. (2012) Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, **28**, 1353–1358.

Stevenson,K.R. *et al.* (2013) Sources of bias in measures of allele-specific expression derived from RNA-sequence data aligned to a single reference genome. *BMC Genomics*, **14**, 536.

van de Geijn,B. *et al.* (2015) WASP: allele-specific software for robust discovery of molecular quantitative trait loci. *Nat. Methods*, **12**, 1061–1063.

Wang,K. *et al.* (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.

Welter,D. *et al.* (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.*, **42**, 1001–1006.

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

Chapter 5

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

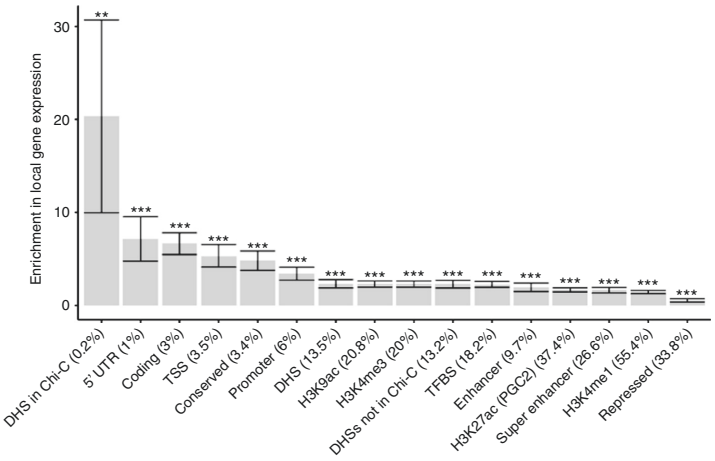


Fig. 1 Open chromatin sites (DHSs) within adipocyte promoter CHi-C chromosomal interactions show significant enrichment in *cis* expression. Enrichments in *cis* expression with error bars for different categories using LD score regression analysis (see Methods). For the horizontal axis labels, the value 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

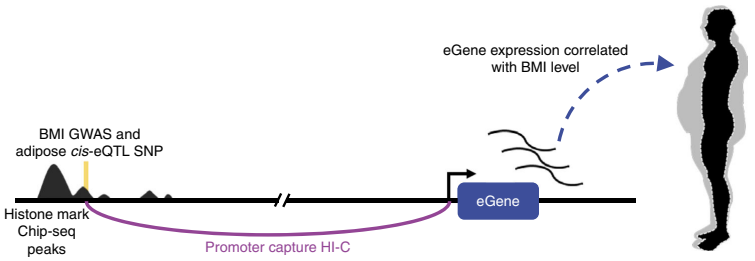


Fig. 2 Overview of the study design targeted to identify new genes for obesity and related metabolic traits. A schematic illustrating the integration of multi-omics data utilized in this study to elucidate genetics of obesity-related traits.

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 co-morbidities 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

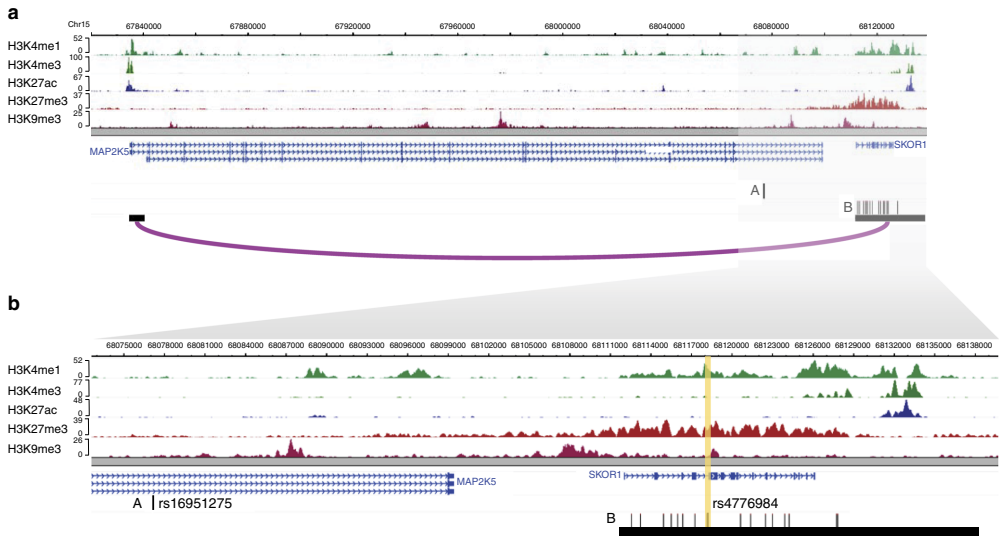


Fig. 3 Promoter Capture Hi-C enables refinement of the BMI GWAS locus that colocalizes with *cis*-eQTLs interacting with the target gene promoter of *MAP2K5*. Genomic landscape of the BMI locus, *MAP2K5* (panels **a**, **b**), modified from the WashU Genome Browser to show the histone mark calls from 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

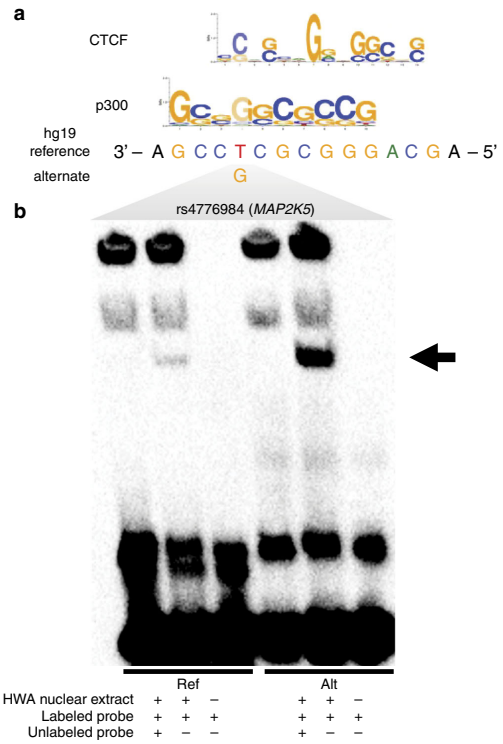


Fig. 4 Predicted TF motifs and electrophoretic mobility shift assay (EMSA) at the rs4776984 site indicate allele-specific binding. **a** Predicted TF motifs for CTCF and p300, as well as the hg19 reference genome sequence. **b** Biotinylated (labeled probe) 31-bp oligonucleotide complexes with ±15 bp 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

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 6

Obesity-induced reduction of Death Associated Protein Kinase 2 predisposes to non-alcoholic fatty liver disease

Abstract

The increasing prevalence of obesity gives rise to a pandemic of non-alcoholic fatty liver disease (NAFLD), which can escalate to non-alcoholic steatohepatitis (NASH), cirrhosis, and liver cancer. The mechanisms leading to dysregulation of body fat distribution and ectopic fat deposition in the liver of obese individuals remain elusive. To elucidate the unknown molecular mechanisms, we examined liver histology as well as liver and subcutaneous adipose tissue transcriptomes in morbidly obese individuals and re-profiled their adipose transcriptomes one year after bariatric surgery. The human adipose expression of 17 genes was downregulated in NASH and negatively correlated with liver steatosis and serum triglycerides (TGs). Mouse models validate that adipose expression is correlated with diet-induced liver steatosis for 5 of these genes, including Death Associated Protein Kinase 2 (*Dapk2*). Post-surgery, weight loss restores *DAPK2* adipose expression. Combining phenotype and longitudinal transcriptomes, mediation analyses demonstrated the causal effect of *DAPK2* adipose expression on NAFLD. Finally, *DAPK2* knockdown in human primary preadipocytes downregulates five genes involved in autophagy, of which two also function in adipocyte differentiation, highlighting a possible disease mechanism. Overall, our genomic and experimental data implicate *DAPK2* in NAFLD pathogenesis.

Introduction

Non-alcoholic fatty liver disease (NAFLD) is the hepatic manifestation of the metabolic syndrome[1] and is becoming the most common cause of advanced liver disease. NAFLD is independent of alcohol consumption and begins with ectopic fat accumulation in the liver, also known as steatosis, which can progress to non-alcoholic steatohepatitis (NASH), characterized by lobular inflammation, liver cell ballooning, and hepatic cell death[1]. Without intervention or treatment, NASH may further develop into more severe liver disorders such as cirrhosis, liver failure, or even hepatocellular carcinoma (HCC) that are often irreversible and require a liver transplant. The main risk factor of NAFLD is obesity, which continues to rise in prevalence, currently affecting more than 600 million adults globally[2,3]. As a consequence, the approximate 25% worldwide frequency of NAFLD[4,5] also increases. Therefore, acquiring a deeper understanding of NAFLD and developing better tests for prognosis, diagnosis, and treatment for NASH are imperative to relieve the associated health and economic burdens. Most NAFLD patients only have simple steatosis, which is generally benign[6]; however, roughly 20-30% of people with NAFLD will advance into NASH[6,7]. An important group of patients is those that progress directly from NAFLD to HCC without significant fibrosis or cirrhosis[8]. Despite the previous advances in identifying genes, such as *PNPLA3* and *TM6SF2*[9,10], for NAFLD, the key drivers of NASH still remain largely unknown. Accordingly, the only effective treatment is weight loss that is difficult to achieve and maintain for some patients.

In healthy individuals, excessive fat is canonically stored in the subcutaneous adipose tissue, which can expand if necessary. However, if the adipose function is compromised by obesity, hypertrophic dysfunctional adipocytes are promoted and TGs will accumulate as fat in ectopic body parts, including the liver and epicardium. Indeed, NAFLD is now viewed as a

metabolic, multisystem disorder[11]. Thus, uncovering the regulators that modulate the TG synthesis and storage as well as fatty acid synthesis and oxidation in the adipocytes may be a key to understanding the development of NAFLD. Furthermore, studying the subcutaneous adipose tissue requires less invasive methods than biopsy of the liver. Formulating better adipose-based biomarkers for NAFLD can improve its diagnosis and prognosis.

Endeavors to identify genes contributing to NAFLD and NASH via genome-wide association studies (GWASs) have thus far uncovered only a handful of candidates[12]. The gold standard to accurately diagnose NASH is by examination of the histology of a liver biopsy, which is difficult to collect in humans. The current non-invasive diagnostic methods to examine NAFLD range from liver enzyme measurements, abdominal ultrasound, ultrasound elastography (Fibroscan), magnetic resonance imaging (MRI), and NAFLD fibrosis scores to liver biopsy. Consequently, small sample sizes, diagnostic heterogeneity, and poor statistical power have hampered the success of GWAS. In addition to obesity, there are other potential environmental effects that can contribute to the development of NAFLD and NASH, making GWAS less likely to identify the underlying genes. On the other hand, mRNA expression is a highly dynamic process that can influence the pathogenesis or be perturbed by diseases. Furthermore, changes in expression often display much larger effect sizes than genetic variants. Here, we present our work analyzing the liver and adipose transcriptomes from high-risk, morbidly obese individuals that have been carefully characterized for multiple metabolic traits and biopsy-based liver histology, followed by functional experiments in human primary preadipocytes, to gain insight into the etiology of NAFLD and NASH.

Results

To better understand the roles of adipose and liver in the pathogenesis of obesity and NAFLD, we established the Kuopio OBesity Surgery (KOBS) cohort (Fig. 1a) by recruiting severely obese patients (mean BMI=43.03±5.27) undergoing a gastric bypass surgery. Metabolic phenotypes were measured during their visits, and we also collected their liver and subcutaneous adipose biopsies for RNA-sequencing (RNA-seq). After quality controls (QC), we profiled a total of 259 and 254 baseline liver and subcutaneous adipose transcriptomes, respectively, and 169 individuals were resampled for subcutaneous adipose transcriptomes one-year post surgery. Histological examination of the baseline liver was performed to diagnose NASH and analyze steatosis grade, lobular inflammation, hepatocyte ballooning, and fibrosis stage (Supplementary Table 1). We classified patients into three unambiguous groups based on the overall liver histological phenotypes: normal, simple steatosis, and NASH (Fig. 1b). Overall, 84 had normal liver histology, 40 steatosis, 43 NASH, and the remaining 92 patients had a varying degree of NAFLD, respectively.

We first examined associations between liver traits and age, sex, and several metabolic traits in KOBS (n=259, Kruskal-Wallis and χ^2 tests for quantitative and categorical traits, respectively, Fig. 1b). Age, fasting glucose, TGs, and alanine aminotransferase (ALT) levels differed significantly among the three groups (Bonferroni corrected $P<0.05$), where all were elevated in the NASH group (Fig. 1b and 1c). Similarly, NASH patients showed a significantly higher rate of T2D ($P=9.39\times10^{-8}$, Fig. 1b) since 67% of the subjects with NASH also exhibited T2D. To develop a trait representing the overall NAFLD status, we performed a non-linear PCA on all 259 individuals and took the first principal component (D1) as an omnibus meta-liver trait

for NAFLD. We note that D1 is negatively correlated with all original histological parameters, meaning that a higher D1 represents a healthier liver (Fig. 1d).

Preserved liver co-expression modules are associated with metabolic and hepatic traits

We performed a weighted co-expression network analysis (WGCNA) on the 259 KOBS liver transcriptomes and tested the association of co-expression modules with 13 metabolic and liver traits. Thirteen of the 19 modules are significantly associated (FDR<0.05), with at least one of the traits (Fig. 2a and Supplementary Table 2). We validated the module preservation in liver RNA-seq data from GTEx subjects whose causes of death were not liver diseases (n=96). Most trait-associated liver modules (9 of the 13) are highly preserved (Z-score>10, Fig. 2b and Supplementary Table 2), suggesting that gene co-regulation related to main liver functions is robust and consistent across cohorts. Notably, the statin-associated module (light cyan, Fig. 2a) includes *HMGCR*, the therapeutic target of statins, 19 known cholesterol pathway genes, and several potentially novel statin response genes (see Supplementary Note).

The omnibus NAFLD trait, D1, is significantly associated with the six preserved modules (Fig. 2b). We observed the strongest correlation between statin medication and light cyan module, which is highly enriched for metabolic and steroid biosynthesis pathways. Overall, the liver mRNA co-expression networks reflect the key role of liver in multiple metabolic functions as well as the strong immune responses associated with NAFLD (Supplementary Fig. 1a).

Adipose co-expression modules are enriched for immune and lipid pathways

In a parallel analysis, we performed WGCNA on 254 subcutaneous adipose RNA-seq KOBS samples to examine the expression profiles in the morbidly obese adipose tissue. We identified

13 co-expression modules of which 6 are significantly associated (FDR<0.05) with at least one liver or metabolic trait (Fig. 2c and Supplementary Table 3). The adipose modules were preserved (Z-summary>10) in another Finnish adipose RNA-seq cohort, METSIM (n=335) (Fig. 2d), with a consistent trend between the module preservation and association with the liver phenotype, such that robust modules are more likely to be associated with phenotypes. The preserved adipose WGCNA modules are enriched for multiple KEGG pathways (Supplementary Fig. 1b). Overall, the adipose transcriptomes of severely obese individuals are significantly associated with lipid metabolism; however, their strong inflammatory response likely reflects the impaired adipose tissue (Supplementary Fig. 1b).

Inflammatory response genes differentially expressed in NASH livers

To identify genes perturbed by NASH in the liver, we performed a differential expression (DE) analysis between healthy and NASH livers (n=69 and 43, respectively). We classified the liver as healthy if the sum of the NAFLD histology score is zero and the individual did not have T2D. There are 2,823 liver DE genes passing the genome-wide FDR<0.05 (Fig. 3a and Supplementary Table 4), predominantly enriched for ABC transporter, bile secretion, and multiple immune response pathways, verifying the known association between NASH and inflammation[1] and suggesting that NASH influences drug and bile acid metabolism in the liver (Supplementary Fig. 1c).

NASH perturbed adipose genes are associated with serum TGs

Next, we examined whether adipose gene expression is deregulated by NASH and performed DE analysis between individuals with healthy and NASH livers (n=68 and 41, respectively). Overall,

we observed a smaller number of differences for NASH in the adipose tissue than in the liver, as only 43 genes are significantly DE (FDR<0.05) (Fig. 3b and Supplementary Table 5) compared to the 2,823 DE genes in the liver (Fig. 3a and Supplementary Table 4). We did not observe any significantly enriched pathways for the 43 adipose DE genes, likely due to the small number of genes.

Both the liver and adipose tissue play major roles in TG synthesis and storage to regulate body fat distribution in humans, and TGs are transported in lipoprotein particles through the blood stream. Therefore, serum TG levels might be a key link between the two tissues that determine fatty liver progression. Indeed, adipose expression of 21 of the 43 adipose NASH DE genes was also significantly associated with serum TGs (Bonferroni corrected P<0.05) in the KOBS cohort (Fig. 3c and Supplementary Table 6). We further replicated the serum TG association of 17 of the 21 genes in the METSIM adipose RNA-seq data (Fig. 3c and Supplementary Table 6), thus supporting the critical relationship between serum TGs and the expression of these genes in the adipose tissue. Notably, all 17 TG-associated genes are also significantly associated with the liver steatosis grade (P<0.05/21, Bonferroni correction) (Fig. 3c), implicating a potential causal pathway among adipose gene expression, serum TGs, and hepatic TG content.

Mouse model validates correlations of adipose expression with hepatic steatosis

To validate the association between adipose gene expression and liver steatosis, we investigated a large panel of diverse inbred strains of mice, the Hybrid Mouse Diversity Panel[13], that were fed a high fat, high sucrose diet to induce high hepatic TG content, which is analogous to steatosis in humans. Of the 17 replicated human TG-associated genes, 12 corresponding mouse genes,

targeted by 17 probes, were available. Nine probes targeting five genes (*Pfkfb3*, *Hadh*, *Fdft1*, *Npr1*, and *Dapk2*) were replicated across species for the hepatic TG content ($P<0.05/17$, Bonferroni correction) (Supplementary Table 7), showing the same direction of association as in humans. In summary, we identified five genes that are down-regulated in the adipose tissue of individuals with NASH; display robust association between their adipose expression and serum TG levels; and show significant correlations with liver steatosis in both humans and mice (Fig. 3c). Of these, *FDFT1* is a known NASH GWAS gene[14].

Patients improve metabolically one year after the bypass surgery

A total of 173 patients returned for a follow-up examination one year post the surgery. We recorded their metabolic attributes and sampled a subcutaneous adipose biopsy for RNA-seq (n=169 after QC). Consistent with previous reports[15], most metabolic traits, except LDL-C and TC, markedly improved during the year after the surgery, as demonstrated by a significant difference in trait values tested using the paired Wilcoxon signed rank test (Fig. 4a). We compared the gene expression of the adipose tissue between the two timepoints and identified 3,797 genes (Fig. 4b and Supplementary Table 8) that were significantly changed at the follow-up (FDR<0.05). They are significantly enriched for the ribosome pathway ($FDR=2.66x10^{-4}$). Notably, adipose expression of two of the five replicated genes downregulated in NASH, *DAPK2* and *HADH*, are upregulated after the weight loss.

Adipose gene expression partially mediates the association between serum TGs and NAFLD

Given the correlation among serum TGs, NAFLD (represented by D1), and adipose gene expression, we carried out a 3-step mediation analysis (Fig. 5a) to delineate their causal

relationship using the 250 KOBS samples without missing data. Mediation analysis compares linear models to eliminate reverse causality or confounding as possible explanations for an association and thereby allows us to infer causality in an observational study[16]. We identified the appropriate genes for the analysis by performing a stepwise linear regression with D1, as the dependent variable, the five validated genes and serum TGs as predictors, and T2D, age, and sex as covariates in the model. Using alpha<0.05 as the inclusion/exclusion criterion, *NPR1* ($P= 2.96x10^{-4}$) and *DAPK2* ($P= 2.46x10^{-3}$) remain as significant genes in the model. In addition, the T2D status remains in the final model as a significant covariate ($P= 3.49x10^{-8}$), and it is significantly associated with serum TGs (P=0.012). Thus, T2D was included as a covariate in the mediation analysis to avoid any hidden confounding effect on D1 and serum TGs.

In each 3-step mediation analysis, we used the adipose expression of one of the two genes, *NPR1* or *DAPK2*, and serum TGs and D1 as either the mediator, causal, or dependent variable, as illustrated in Fig. 5a. There are 6 possible models in a 3-step mediation analysis (Fig. 5a), when the true causal relationship is unknown. However, the associations between serum TGs and *NPR1*/*DAPK2* adipose expression are consistent in METSIM, which should encompass only few NAFLD patients (mean BMI=26.82, and 3.0% T2D rate in METSIM versus mean BMI=43.03, and 37.0% T2D rate in KOBS, respectively). Consequently, NAFLD (D1) is unlikely to mediate the correlation between TG and *NPR1*/*DAPK2* adipose expression. We therefore excluded models 4 and 6 in the subsequent analyses *a priori*. After accounting for the 2x4 tests using a Bonferroni correction, both models 2 and 5 showed significant mediated effect for *NPR1* and *DAPK2*, indicating their adipose expressions as mediators between serum TGs and D1 (Fig. 5b).

We further expanded models 2 and 5 to jointly model both *NPR1* and *DAPK2* as mediators between serum TGs and D1 using structural equation modeling (SEM) (Fig. 5a). SEM can be generalized as a mediation analysis to infer causality in an observational study using linear model comparisons[17]. The combined effects of *NPR1* and *DAPK2* are highly significant in both models (Fig. 5b), accounting for ~61% of the total correlation between serum TGs and D1. On the other hand, their individual effects still remain significant, suggesting that they each at least partially mediate the association between serum TGs and D1 independently. Overall, our data suggest that *NPR1* and *DAPK2* gene expression in adipose plays the role of mediators in fatty liver development. We further examined the directionality in the next section. However, we caution that the other three validated genes or additional undiscovered genes could also mediate the effect between serum TG levels and NAFLD.

Serum TGs and DAPK2 adipose expression mediate a causal path from obesity to NAFLD

To further elucidate the directionality of the causal pathway, we leveraged the KOBS baseline and follow-up adipose RNA-seq data. We first observed that weight loss and decreased BMI leads to increase in *DAPK2* mRNA adipose expression in the follow-up when compared to its baseline adipose expression ($P=1.19x10^{-10}$), a recovery trend reported previously[15]. Based on this observation, we hypothesized that BMI is most likely the causal variable for NAFLD, and performed another mediation analysis using the population-cohort, METSIM, to determine the causal relationship among BMI, serum TGs, and *DAPK2* adipose expression. Additionally, BMI is supported as the causal variable because the bypass surgery primarily leads to weight loss via inducing calorie deficiency. We tested two models, in which BMI was the causal variable, and adipose expression of *DAPK2* and serum TGs are either the mediator or the dependent variable.

The METSIM data support serum TGs as the mediator between the correlation of BMI and *DAPK2* adipose expression ($P=1.4x10^{-5}$) (Fig. 5c). In conjunction with our previous mediation analysis, which suggests that adipose expression of *DAPK2* mediates the correlation between serum TGs and NAFLD, the most plausible causal pathway is BMI→TGs→*DAPK2*→D1. Taken together, these analyses in KOBS and METSIM support the conclusion that obesity, mediated via increase in serum TGs and decrease in adipose expression of *DAPK2*, has a causal effect on NAFLD in humans.

DAPK2 knockdown suppresses expression of autophagy pathways

To understand the consequences of *DAPK2* reduction in human adipose tissue and given the role of *DAPK2* in autophagy[18], we knocked down *DAPK2* mRNA expression by ~70% using siRNA in biologically independent human primary preadipocytes from two individuals (Fig. 6a). We then assayed the mRNA expression of key autophagy pathway genes and identified five autophagy genes, *ATG9A*, *BAD*, *DAPK1*, *TGM2*, *ULK2*, to be consistently downregulated in both independent experiments of human primary preadipocytes using triplicate wells (Fig. 6b). Notably, two of these genes *TGM2* and *ULK2* also function in adipocyte differentiation[19,20]. Thus, this reduction of autophagy capacity might impact proper adipocyte differentiation leading to compromised adipose tissue. Combining the results of mediation analyses and functional experiments, we propose a novel mechanism for NAFLD pathogenesis due to obesity mediated by *DAPK2* adipose expression (Fig. 6c).

Discussion

To understand the molecular pathogenesis of NAFLD, we utilized 1,113 transcriptomes of human liver and subcutaneous adipose tissues, statistical causal inference, mouse model, and human primary preadipocytes to uncover *DAPK2* as a potential causal gene for NAFLD. The adipose expression of *DAPK2* is significantly lower in individuals with NASH than with a healthy liver and is correlated with serum TGs and steatosis. We further replicated the association between *DAPK2* adipose expression and hepatic steatosis in a large panel of mouse strains (Fig. 3c). Leveraging the longitudinal adipose transcriptomes, we demonstrate the causal relationship among obesity, serum TGs, *DAPK2*, and NAFLD. *DAPK2* knockdown via siRNA in human primary preadipocytes reduced the expressions of five key autophagy genes, which elucidates a possible mechanism of NAFLD, mediated by *DAPK2* in adipose tissue. Based on converging evidence from multiple independent data and experiments, we showed that *DAPK2* is a key mediator linking obesity, lipid metabolism, and hepatic steatosis (Fig. 6c).

DAPK2 (Death Associated Protein Kinase 2) is a serine/threonine kinase, involved in multiple cellular signaling pathways that trigger cell survival, apoptosis, and autophagy[18]. It is also one of the most downregulated genes in the adipose tissue of morbidly obese individuals[21]. Soussi et al. showed previously that adipose *DAPK2* mRNA levels in obese patients (n=10) are inversely correlated with fat cell size and gradually recover after bariatric surgery-induced weight loss[15]. Our longitudinal adipose RNA-seq data from the morbidly obese individuals in the KOBS cohort (n=169) confirm this weight-loss induced recovery effect as *DAPK2* adipose expression significantly increases one year after the bariatric surgery. Taken together with our finding that *DAPK2* adipose expression is associated with liver steatosis and further reduced in the morbidly obese individuals with NASH, *DAPK2* expression in the adipose tissue might

contribute to the NAFLD development caused by obesity. We hence performed causal inference analyses[16,17] to support *DAPK2* adipose expression as the mediator between obesity and NAFLD. The first 3-step mediation analysis using the baseline KOBS data indicated that adipose expression of *DAPK2* and *NRP1* are mediators between NAFLD and serum TGs (Fig. 5a-b, Models 2 and 5). Leveraging the longitudinal cohort, KOBS, we anchored obesity (represented by BMI) as the causal variable due to the temporal precedence of weight loss on restoring the *DAPK2* adipose expression after the bariatric surgery. We then performed a second mediation analysis among BMI, serum TGs, and *DAPK2* adipose expression in an independent cohort, METSIM (Fig. 5c). These results helped us orient the causal direction between TGs and NAFLD and imply the causal pathway BMI→TGs→*DAPK2*→NAFLD. The novel mediator role of *DAPK2* adipose expression in this causal path between obesity and NAFLD is further supported by our functional experiments in human primary preadipocytes.

Earlier data showed that reduced *DAPK2* expression in morbidly obese individuals causes defective lysosomal autophagy in adipose tissue, contributing to their fat cell dysfunction[15]. Noteworthy, previous studies using autophagy-defective mouse models have implicated a pivotal role of autophagy in adipocyte differentiation[22-25]. Our knockdown experiments further illustrate the functional role of *DAPK2* in the key adipose tissue cell type, human primary preadipocytes that have the full capacity of adipogenesis. We demonstrate a consistent down-regulation of five known autophagy genes, *ATG9A*, *BAD*, *DAPK1*, *TGM2*, *ULK2*. Of the impacted autophagy genes, *TGM2* and *ULK2* have also been implicated in adipocyte differentiation[19,20]. Specifically, *Tgm2* deficient mouse embryonic fibroblasts display an accelerated lipid accumulation due to increased expression of major adipogenic transcription factors, PPAR□ and CEBPA[19]. *Ulk2* is required for adipocyte differentiation in mouse 3T3-L1

cells as knockdown of *Ulk2* suppresses adipogenesis in this cell line[20]. Collectively, our data and these knockdown experiments in mouse cell lines support the notion that *TGM2* and *ULK2* may impact adipogenesis in human primary preadipocytes due to defective autophagy. However, additional functional studies are warranted to establish this and the underlying functional mechanisms in humans.

In conclusion, our genomic analyses and functional experiments imply that the recovery of *DAPK2* adipose expression levels due to the weight loss after bariatric surgery may be beneficial for NAFLD/NASH prognosis. Additional future studies taking an adipose RNA sample and liver imaging during the bariatric surgery and one year after the surgery are warranted to assess the prognostic value of monitoring *DAPK2* adipose expression during weight loss in predicting the recovery from NAFLD. Ultimately, these results support the mechanistic model where obesity-induced reduction of *DAPK2* impairs subcutaneous adipose tissue and mediates the development of NASH.

Methods

Study cohort:

Overall, a total of 1,113 adipose and liver transcriptomes were analyzed in this study from the participants of the following three cohorts. The Kuopio OBesity Surgery (KOBS) study consists of a total of 356 Finnish patients that underwent bariatric bypass surgery and their 1-year follow-up at the University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland, as described in detail previously[26]. All patients are Finns of European ethnicity. Liver and subcutaneous adipose biopsies were taken at the baseline during the surgery, and a subcutaneous adipose biopsy was taken from a subset of patients at the one-year follow-up (see Supplementary Table 1 for a summary of the KOBS cohort). Whole blood samples were collected for genotyping and biochemical measurements after 12-hour fasting. Lipid, glucose, and anthropometric measurements were performed at the baseline and follow-up using standard methods, as described previously[26].

The METabolic Syndrome in Men (METSIM, n=10,197) [27] participants were recruited at the University of Eastern Finland and Kuopio University Hospital, Finland. All METSIM participants are Finns of European ethnicity. Subcutaneous abdominal adipose needle biopsy was taken from a subset of METSIM individuals that did not exhibit T2D at examination. Adipose RNA-seq data from 335 METSIM individuals were included in this study. Both METSIM and KOBS study designs were approved by local ethics committee and all participants gave a written informed consent.

We also included liver RNA-seq data on 96 samples from the GTEx project[28]. We downloaded the raw RNA-seq reads as FASTQ files from the dbGaP accession phs000424.v6.p1.

Liver biopsy histology and meta-liver trait:

The liver histological evaluation was performed and scored by one pathologist according to the NASH Clinical Research Network (CRN) criteria[29] for the following four attributes: steatosis grade (0-3), lobular inflammation (0-2), ballooning (0-2), and fibrosis stage (0-4). The diagnosis for NASH was also determined by the pathologist following the standard guidelines[30]. Based on the histological examination of the liver, we classified patients into three groups as previously described[26]: (i) normal liver, which has no steatosis, inflammation, ballooning, fibrosis, or NASH, (ii) simple steatosis, which has a steatosis score$\geq$1 but no ballooning, inflammation, fibrosis, or NASH, and (iii) NASH, which is diagnosed by the pathologist. Other patients that cannot distinctly fit into any of the above categories were excluded in the specific analyses that utilize this classification, but otherwise all samples were included in the analyses.

To assess the NAFLD status in all 259 individuals, we performed a non-linear principal component analysis (PCA) using the homals R package[31] on the four CNR liver histological phenotypes and used the first principal component (PC1) as the aggregated meta-liver trait (D1) for NAFLD (Fig. 1d).

Genotyping and QC:

All KOBS participants were genotyped using the Illumina OMNI Exome Express array with DNA isolated from whole blood. The QC on the genotype data was carried out with the following criteria: (i) sex concordance, (ii) Hardy-Weinberg Equilibrium (HWE) $P<1x10^{-6}$, (iii) genotype call rate >95%, and (iv) no extreme heterozygosity rate. Second-degree relatives (Pi_HAT>0.2) within the KOBS cohort were excluded.

RNA isolation and RNA-sequencing:

All RNA samples were isolated at the University of Eastern Finland using the miRNeasy (Qiagen) kit and were subsequently sequenced at the UCLA sequencing core. The KOBS and METSIM adipose RNA-seq libraries were prepared using TruSeq PolyA-select, whereas, the KOBS liver was prepared using Ribo-Zero gold. External RNA Controls Consortium (ERCC) spike-ins (ThermoFisher Scientific) were added to the libraries of both KOBS tissues. The target read depth is approximately 40-50 million read pairs per sample. On average, a sample library was divided into 3 flowcell lanes during the multiplexing step and each lane consisted of up to 12 samples. A summary of RNA-seq library type and platform is shown in Supplementary Table 9.

RNA-seq processing and QC:

We aligned the reads from the KOBS, METSIM, and GTEx RNA-seq data against the human reference genome, hg19, using STAR and its two-pass protocol[32]. No additional QC was performed for the GTEx dataset since we only included samples that already passed the previously published GTEx QC pipeline[28]. We assayed the KOBS and METSIM RNA-seq quality using QoRT [33] and generated data quality and alignment summary statistics with Picard (http://broadinstitute.github.io/picard/). To exclude or correct contaminated or mislabeled samples respectively, we tried to match the genotype array and RNA-seq data using exonic SNPs via VerifyBamID[34]. RNA-seq samples that cannot be exclusively linked to a genotype data were excluded, otherwise, we relabeled the RNA-seq sample to the corresponding genotyped individual. We required an RNA-seq sample to have at least 20 million uniquely mapped reads and the correct Library strandedness. Overall, a total of 1,017 new human RNA-seq samples

were included in the study as follows: The samples passing all QCs consisted of 259 KOBS baseline liver samples, 254 KOBS baseline adipose samples, 169 KOBS follow-up adipose samples, and 335 METSIM adipose samples, respectively (Supplementary Table 9). In addition, we included liver RNA-seq data of 96 GTEx samples[28], resulting in a total of 1,113 analyzed RNA-seq samples.

Transcript and gene quantification:

We estimated the transcript abundance as read counts and transcript per million (TPM) using Kallisto[35], based on GENCODE version 25 liftover to hg19 gene annotation. The gene read counts or TPM were calculated as the sum of read count or TPM respectively of all transcripts from a gene. We focused on autosomal protein-coding genes and lincRNAs in all subsequent analyses. To filter out lowly expressed genes, we required a gene to have at least 10 reads in 80% of samples resulting 15,670 genes in the liver RNA-seq and 14,622 genes in the adipose RNA-seq.

Hidden covariate estimation for RNA-seq:

As RNA-seq is prone to hidden confounders often due to sequencing and other technical factors that can obstruct true biological signal (Supplementary Fig. 2), we first examined if these technical attributes are associated biological traits (Supplementary Fig. 3), and then adjusted the data for the necessary technical and phenotypic covariates as follows. We performed a supervised surrogate variable analysis (sSVA)[36] on gene TPM quantification, utilizing 92 ERCC spike-in transcripts as invariable controls to estimate latent variables for downstream analyses in

both the liver and baseline adipose KOBS RNA-seq data. The null and full models provided to sSVA are:

$$Null\ model: Gene\ exp. \sim Alignment\ rate + MT\ \% + 3'bias + BMI + Sex + Age$$

$$Full\ model: Gene\ exp. \sim Alignment\ rate + MT\ \% + 3'bias + BMI + Sex + Age + D1$$

where alignment rate is the percent of uniquely aligned reads; MT % is the percent of mitochondrial reads; 3' bias is the three prime bias estimated by Picard; D1 is the PC1 from the liver histological score. We performed a separate sSVA combining the baseline and follow-up adipose data with RIN, alignment rate, MT%, 3' bias, and BMI as covariates and subject ID for blocking to account for the paired status. The surrogate variables are highly correlated with RNA-seq technical factors but not with biological phenotypes (Supplementary Fig. 4). Since ERCC spike-ins were not added to METSIM and GTEx libraries, sSVA was not performed for either cohort. All known covariates in the null model and sSVA factors were included as covariates for all downstream analyses.

Differential expression analysis:

Using read counts from Kallisto as gene quantification, we first performed TMM normalization using the *calcNormFactors* function from the edgeR package[37] in R and then carried out variance stabilization via voom[38] prior to the differential expression analysis. We built a linear model using LIMMA[39] with all samples and then contrasted the coefficients of the healthy and NASH livers to identify DE genes between the two groups. Individuals without T2D, steatosis, or any fatty liver related symptoms were considered healthy (n=69 in liver; n=68 in adipose). There are

43 and 41 NASH individuals in the liver and adipose RNA-seq data respectively. The significant threshold was set at FDR<0.05.

We used the same analytical pipeline for the DE analysis between the baseline and follow-up adipose tissue RNA-seq data, except that in this analysis we accounted for the repeated measurement by including the subject as the blocking factor.

WGCNA analysis:

We first transformed raw TPM to log2(TPM+1) and then performed empirical Bayes-moderated linear regression implemented in the WGCNA package[40] (function *empiricalBayesLM*) to correct for covariates while retaining the variation due to the trait of interest (the liver meta-trait, D1, when available). We calculated pairwise gene correlation using biweight correlation allowing a maximum of 5% outliers, and subsequently built a signed network using the soft threshold power of 12. The eigen-gene of each module was calculated and used for trait association tests. To test the module preservation in GTEx and METSIM, we re-processed the RNA-seq raw reads using our pipeline and the same QC. Only genes that were expressed in both KOBS and GTEx/METSIM were used in the analyses. We considered a module with a preservation summary Z-statistics >10 as strongly preserved based on the previous method and guidelines[41].

Statistical analysis for expression-trait association:

All statistical analyses were performed in R (https://www.R-project.org/). We used linear regression in all trait association tests where the expression and trait were treated as dependent and independent variables, respectively. The log2(TPM+1) values were adjusted for covariates

and/or sSVA surrogate variables using Bayes-moderated linear regression, and all continuous traits were first inverse rank transformed to normality.

Mouse microarray analysis:

The mouse microarray data were obtained and processed as previously described[13,42]. Briefly, a total of 112 mouse strains (3 males and 3 females per strain on average) were profiled using Affymetrix hT_MG430A arrays and array image data were processed using Affymetrix GCOS algorithm. The probe signal was normalized using either quantile normalization or robust multiarray method (RMA) for each probe[42]. The probe intensity and phenotypes were averaged across replicates for each strain and sex. The log2 transformed RMA probe intensity was used as a dependent variable, and the hepatic TG content was used as the predictor in a linear regression model, in which sex was included as a covariate.

Pathway enrichment analysis:

We performed the overrepresentation enrichment analyses using the WEB-based Gene SeT AnaLysis Toolkit (WebGestalt) (http://www.webgestalt.org/) using the KEGG pathway database. Only pathways passing FDR<0.05 are reported.

Mediation analysis:

We performed a 3-step mediation analysis by building the following two linear models:

$$M \sim aX + e_1$$

$$Y \sim cX + bM + e_2$$

where X is the causal variable; M is the mediator; Y is the dependent variable, and e_1 and e_2 are the error terms. The total effect of X on Y is estimated as c+a·b, whereas the mediated effect is a·b. Under the null hypothesis of no significant mediation (a·b=0), we calculated the mediated effect and its 95% confidence interval using the RMediation package[43] in R, and the p-value was calculated using a two-sided Z-score test. Since the true causal relationships among serum TGs, *NPR1*/*DAPK2* adipose gene expression, and the liver meta-trait, are unknown, we rotated their roles in the model and tested all possible combinations (Fig. 5a). We included the T2D status in both linear models as a covariate due to its association with serum TGs and D1. For the mediation analysis among BMI, serum TGs, and *DAPK2* adipose expression in METSIM, we anchored BMI as the causal variable as the KOBS patients were primarily treated for weight loss via the bypass surgery. Thus, we only rotated the roles of serum TGs and *DAPK2* adipose expression as the mediator and dependent variable.

We extended the 3-step mediation model to accommodate two mediators as follows:

$$M_1 \sim a_1 X + e_1$$

$$M_2 \sim a_2 X + e_2$$

$$Y \sim cX + b_1 M_1 + b_2 M_2 + e_3$$

where X and Y are the causal and dependent variables; M_1 and M_2 are the mediators, and e_1, e_2, and e_3 are the error terms. We carried out the analysis using a structural equation model (SEM) framework implemented in the lavaan R package[44], and we also included the T2D status as a covariate and modeled the correlation between the mediators. Similar to the single mediation

model, the mediating effects by M_1 and M_2 are $a_1 \cdot b_1$ and $a_2 \cdot b_2$, respectively, and the combined mediated effect is consequently $a_1 \cdot b_1 + a_2 \cdot b_2$. The total effect of X on Y is thus $c + a_1 \cdot b_1 + a_2 \cdot b_2$. The 95% confidence intervals were estimated using the adjusted bootstrap percentile method ("bca.simple" option). P-values of the mediated effects were calculated using a two-sided Z-score test. All variables were inverse rank transformed to normality to avoid outlier effects and standardize all effect sizes. The adjusted *NPR1*/*DAPK2* adipose TPM value from Bayes-moderated linear regression was used in the models.

***DAPK2* functional assays in primary preadipocytes:**

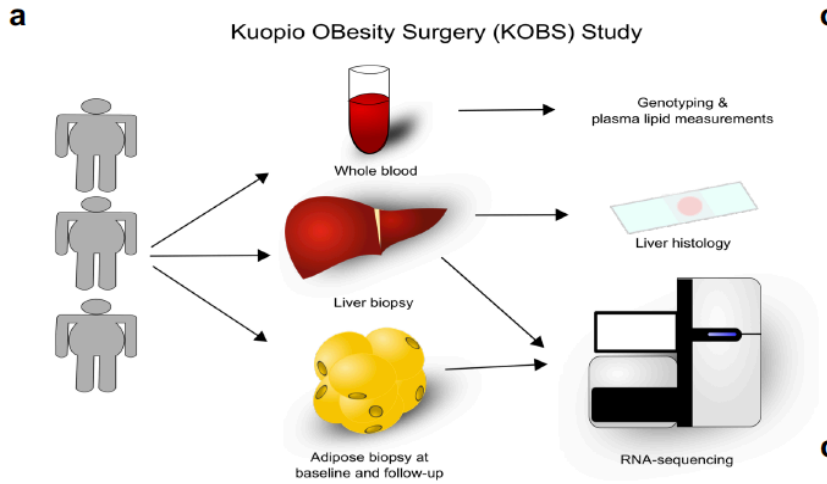
We performed two biologically independent experiments using human primary white preadipocytes (PromoCell, lot # 395Z024 and 419Z023) derived from two different Caucasian women (age 44 and BMI 22.0; and age 56 and BMI 23.8, respectively). The cells were maintained in a monolayer culture at 37^{o}C and 5% CO^2 using preadipocyte growth medium (PromoCell C-27410) with 1% HyClone penicillin/streptomycin (GE Healthcare Life Sciences SV30010) and following PromoCell preadipocyte culturing protocols. For siRNA transient transfections, we plated 0.3 million primary human white preadipocytes in a 6-well plate in triplicates and grew to ~80% confluency. We transfected the cells in Opti-MEM (Gibco 31985062) with Lipofectamine RNAiMax (Invitrogen 13778100) at the appropriate siRNA concentration, 20 pmoles per manufacturer's protocol. Control siRNA (IDT #51-01-14-03) was used as negative control (NC). After 24 hours of transfection, the medium was removed and the cells were washed with PBS once prior to being treated with Trizol (Invitrogen 15596026). We performed RNA extraction per manufacturer's protocol using Direct-zol RNA Mini-Prep (Zymo

Research R2061). We synthesized the cDNA from 500 ng of RNA using Maxima First Strand cDNA Synthesis Kit after DNAse treatment (Thermo Scientific K1642).

RT-qPCR experiment and analysis:

For the siRNA knockdown experiment, we measured relative gene expression by RT-qPCR using an Applied Biosystems QuantStudio 5 detector. Expression levels were normalized to *36B4*. Primers for *DAPK2* and *DAPK2* siRNA were obtained from Soussi et al[15]. Autophagy gene expression panel was conducted using PrimePCR Assay autophagy (SAB Target List, Bio-Rad 10034452). Ct values were normalized to *ACTB*, and relative expression was calculated as ΔCt. We compared the ΔCt values between knockdowns and controls using an unpaired t-test and reported the two-tailed p-values for all RT-PCR values.

Gene	Primer Pair	Primer Sequence
DAPK2	Forward	5'- ACGTGGTGCTCATCCTTGA-3'
	Reverse	5'- TGGCCTCCTCCTCACTCA-3'
DAPK2 siRNA		5'-UGUCUGGAGGAGAGCUCUU-3'
36B4	Forward	5'-CCACGCTGCTGAACATGCT-3'
	Reverse	5'-TCGAACACCTGCTGGATGAC-3'



b

Trait	NASH	Simple steatosis	Normal liver	P
N	43	40	84	-
Sex (M/F)	21/22	11/29	21/63	0.019*
T2D (control/case)	14/29	29/11	69/15	9.39e-08*
Statin (non-user/user)	19/22	34/6	61/22	0.00042*
Age (years)	51.91 (8.5)	47.49 (8.2)	45.88 (9.9)	0.0026*
BMI (kg/m^2)	43.75 (5.1)	43.15 (4.7)	42.58 (5.3)	0.34
ALT (IU/L)	56.12 (43.6)	43.75 (28.0)	35.55 (23.1)	0.0034*
TC (mmol/L)	4.08 (0.9)	4.14 (0.8)	4.12 (0.8)	0.71
HDL-C (mmol/L)	1.08 (0.3)	1.04 (0.2)	1.15 (0.3)	0.23
LDL-C (mmol/L)	2.16 (0.8)	2.36 (0.8)	2.35 (0.7)	0.26
TG (mmol/L)	1.86 (0.9)	1.66 (0.7)	1.37 (0.5)	0.0028*
Glucose (mmol/L)	7.84 (2.7)	6.32 (2.0)	5.90 (1.2)	4.20e-07*

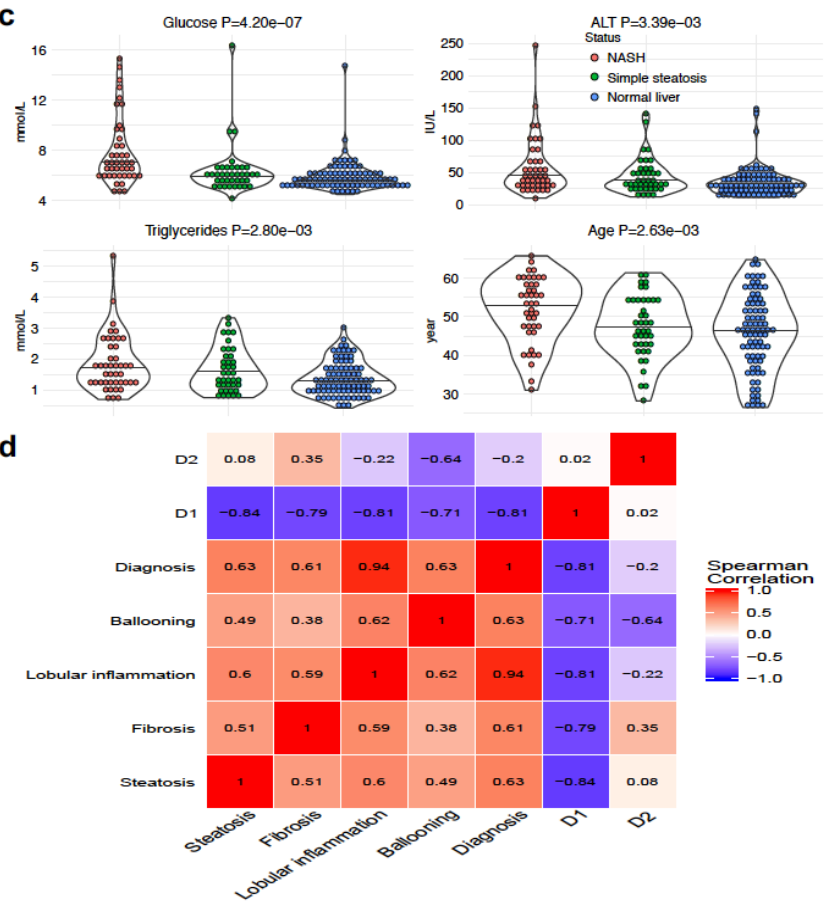


Figure 1. An overview of the KOBS cohort and summary of clinical and histological phenotypes. a, The KOBS participants underwent a gastric bypass surgery to treat obesity and biopsies were collected from the liver and subcutaneous adipose tissue. Adipose tissue biopsy and clinical phenotypes were also collected at one-year of follow-up. **b**, A comparison of the three liver phenotype groups in the KOBS cohort, summarizing the metabolic traits. The numbers in parenthesis are the percentage of affected patients and standard deviation for categorical and quantitative variables, respectively. P-values were calculated using Kruskal-Wallis and χ^2 tests for quantitative and categorical traits respectively. **c**, Violin plots of the four continuous metabolic traits that are significantly different among the three liver phenotype groups. **d**, Spearman correlations between the histology-based liver traits and their first two non-linear principal components, D1 and D2. We used D1 as the omnibus liver trait to represent the NAFLD status in all 259 individuals. A lower D1 implies a healthier liver.

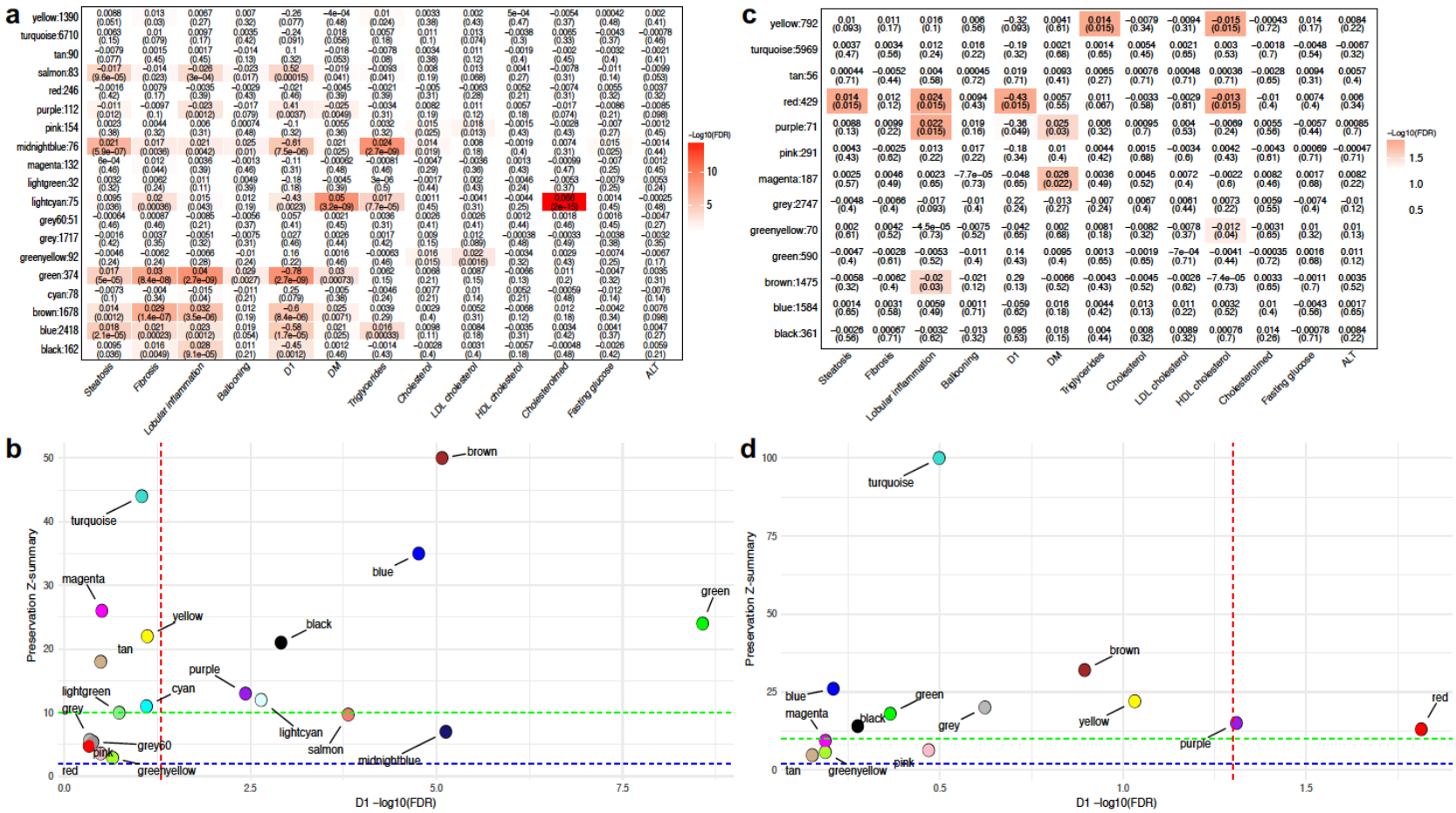


Figure 2. Results of the liver and adipose weighted gene co-expression network analyses (WGCNA). a, The association results between the liver WGCNA modules and histological liver phenotypes and metabolic traits in KOBS. **b**, KOBS WGCNA module preservation in the GTEx liver RNA-seq data (n=96). A Z-score greater than 10 is considered strongly preserved. **c**, The association results between the adipose WGCNA modules and histological liver phenotypes and metabolic traits in KOBS. Numbers in the cells and parenthesis indicate the effect sizes and FDRs, respectively. **d**, Module preservation in an independent Finnish adipose RNA-seq cohort, METSIM (n=335). A Z-score>10 is considered as strongly preserved. In **a** and **c**, the effect sizes and FDRs are listed in each cell and in parenthesis. In **b** and **d**, the blue and green horizontal dotted lines indicate the summary preservation Z-score at 5 and 10 respectively, and the red dotted line indicates the significant threshold for trait association at FDR<0.05.

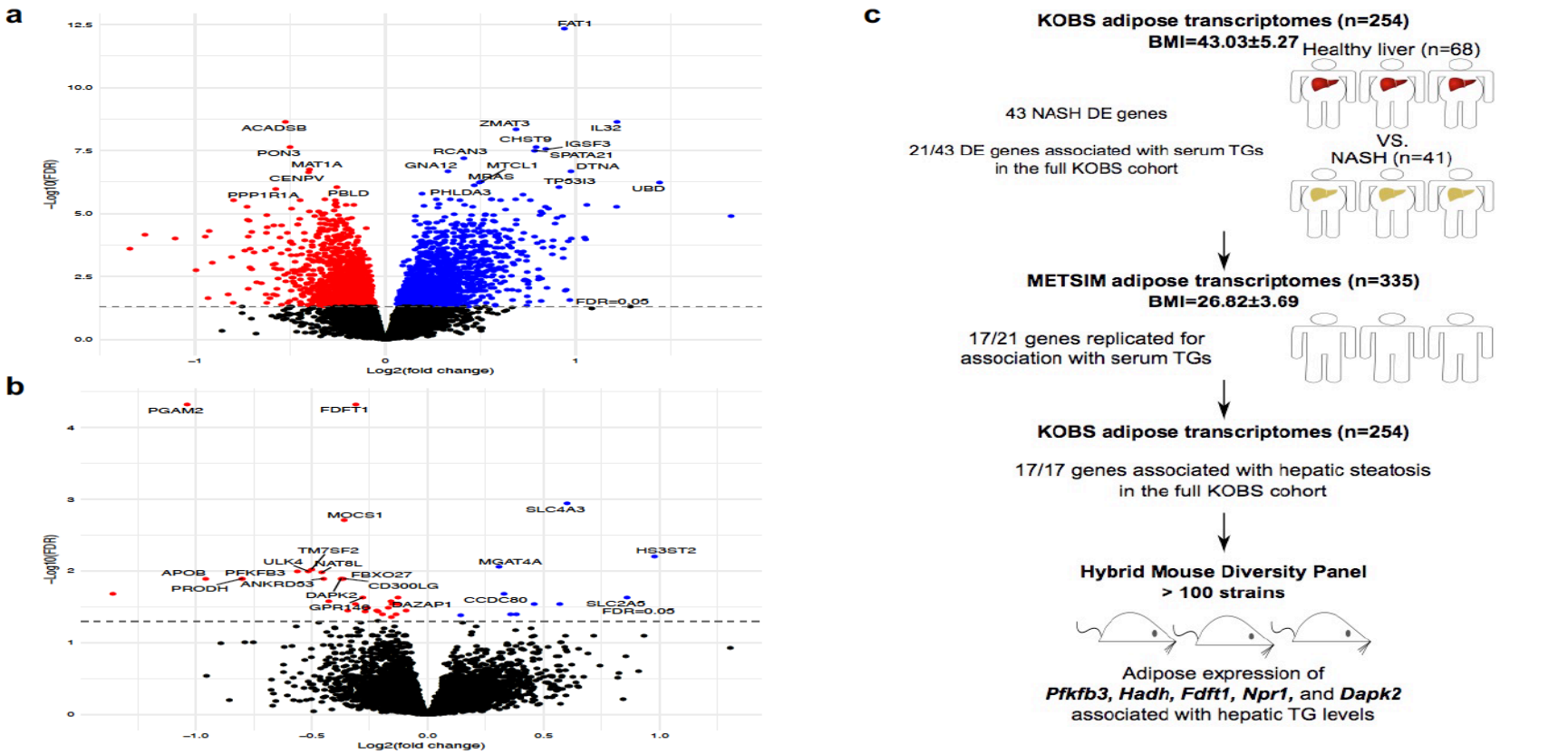


Figure 3. Differential expression (DE) analyses of NASH vs. healthy liver find key genes perturbed by NASH in the liver (a) and baseline adipose tissue (b); as well as a schematic overview of the identification of the five NASH DE adipose genes, consistently associated with hepatic steatosis in both humans and mice (c). a, The volcano plot of the DE results in the liver. A total of 2,823 genes are differentially expressed. **b**, The DE results of the adipose tissue at the baseline, visualized as a volcano plot. There are 43 genes that are differentially expressed. In **a** and **b**, genes with a higher or lower expression in the NASH patients are colored in blue and red, respectively; the genome-wide FDR<0.05 is indicated by the dotted grey line; and the top 20 genes are labeled. **c**, An overview of the identification of the five NASH DE adipose genes, consistently associated with hepatic steatosis in both humans and mice. We tested the association between serum TGs and adipose expression of the 43 NASH DE genes in the KOBS cohort, and replicated the association between serum TGs and adipose expression for 17 of the 21 genes (Bonferroni corrected P<0.05) in the METSIM cohort. The adipose expression of all 17 genes was also associated with hepatic steatosis in KOBS. Finally, we validated the association with the hepatic steatosis for five genes in mice using the Hybrid Diversity Mouse Panel.

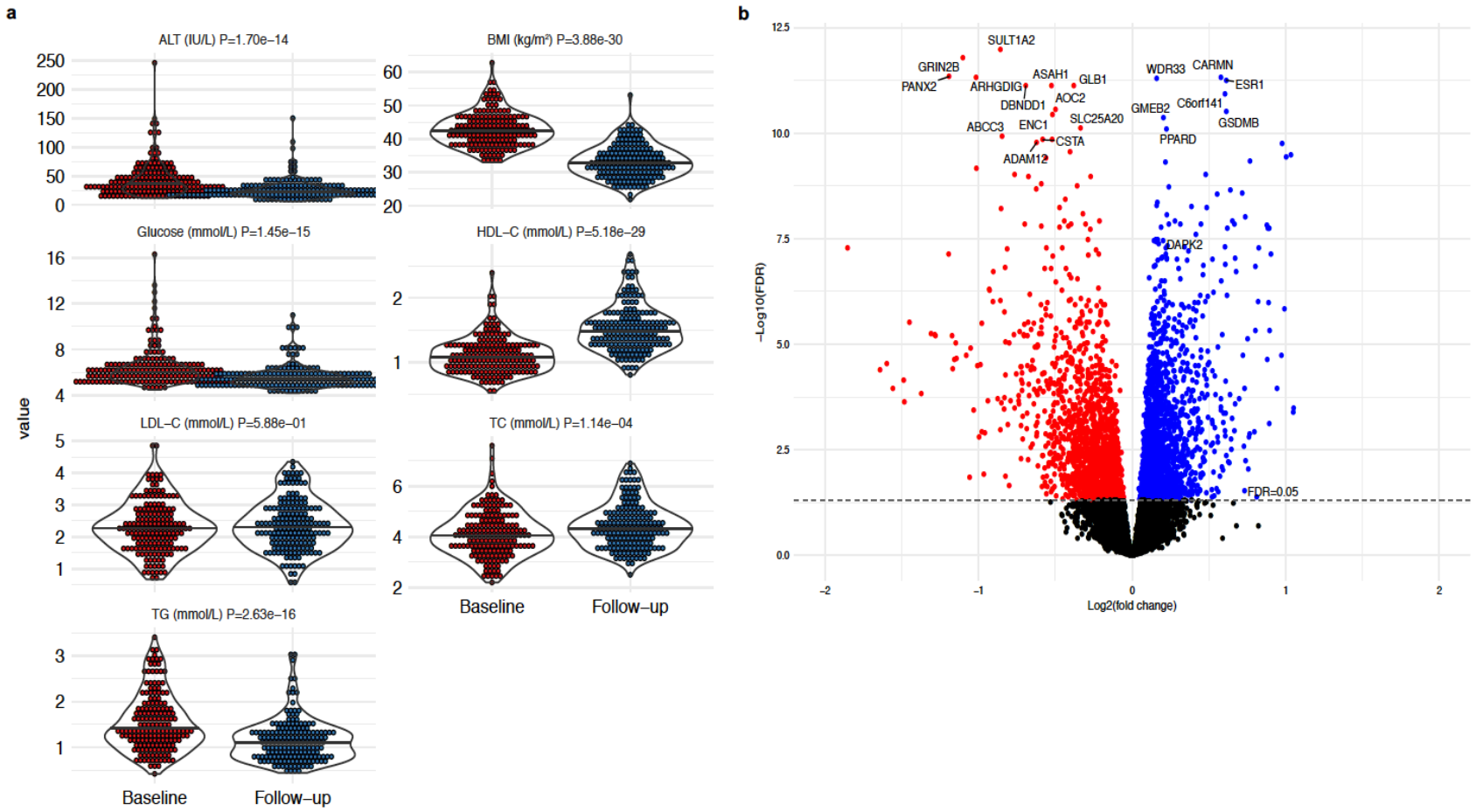


Figure 4. The metabolic profiles and adipose expression change significantly one year post bariatric surgery. a, Metabolic phenotypes clearly improve and BMI drastically decreases at the 1-year post surgery follow-up visit in KOBS when compared to the baseline measurements. The analyses are performed using the paired Wilcoxon signed rank test. **b**, A volcano plot of the adipose DE results between the baseline and one-year follow-up. A total of 3,797 genes are significantly DE (FDR<0.05). The dotted grey line indicates the genome-wide FDR=0.05. Genes that are up-/down-regulated at follow-up are colored as blue and red, respectively. The 20 most significant genes and *DAPK2* are labeled.

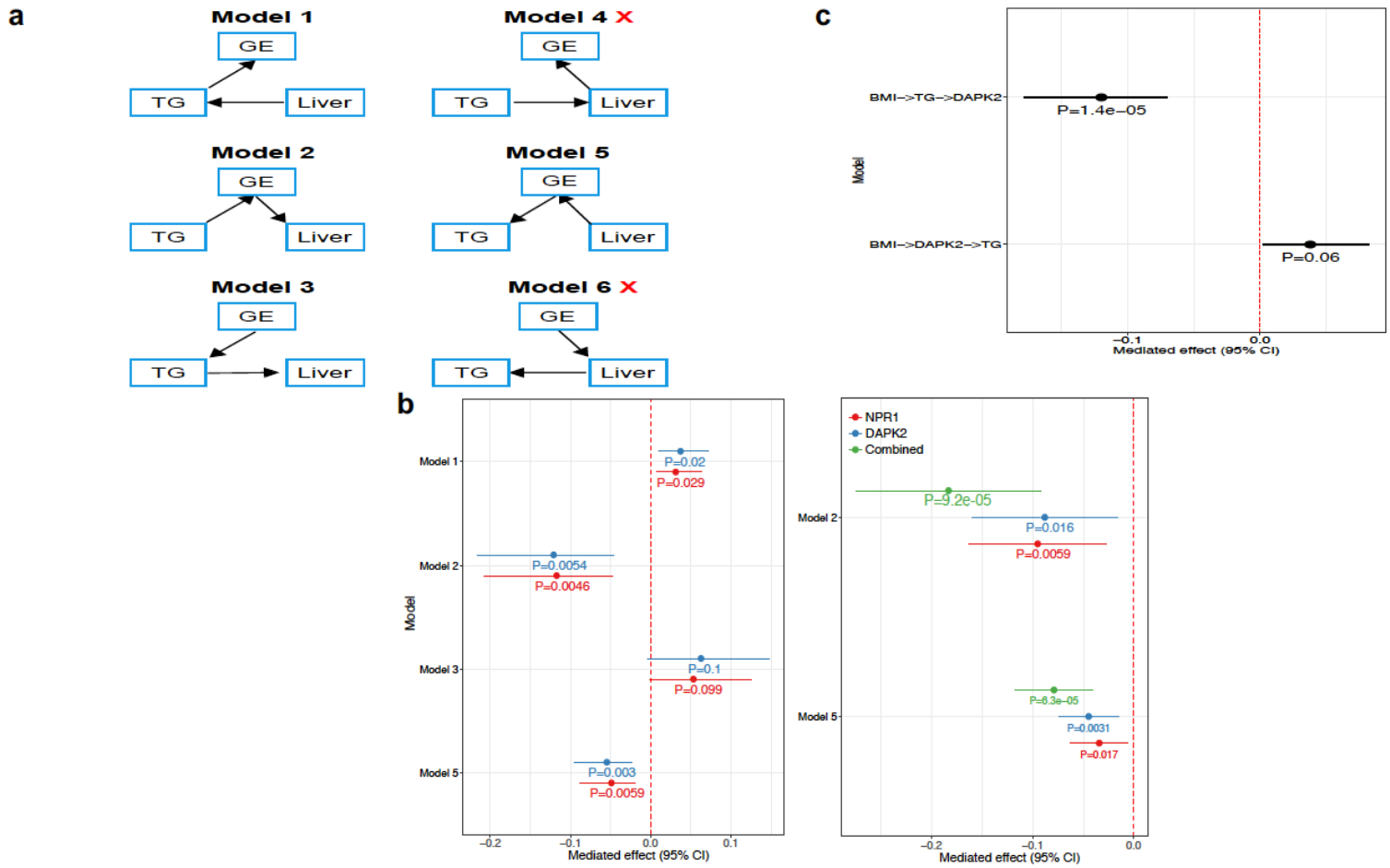


Figure 5. Potential causal pathways in NAFLD, supported by two independent human cohorts. **a**, There are 6 possible causal models given the three variables. Models 4 and 6 were eliminated since the correlations between serum triglycerides (TGs) and gene expression (GE) of *DAPK2/NPR1* in the adipose are consistent between KOBS and METSIM. As the METSIM participants are overall healthier (3.0% T2D rate; mean BMI=26.82, SD=3.69) than the morbidly obese KOBS participants (37.0% T2D rate; mean BMI=43.03, SD=5.27), the risk of NAFLD should be lower in METSIM, NAFLD should not mediate the correlation between serum TGs and GE. **b**, The mediation analysis results with adipose expression of *DAK2/NPR1*, serum TGs, and NAFLD (represented by D1), in the KOBS cohort. On the left panel, the mediated effects and two-sided p-values of each model from (**a**) are shown. Models 2 and 5

displayed significant mediated effects after accounting for the 8 (2x4) tests. On the right panel, models 2 and 5 were extended to simultaneously include the adipose expression of *NRP1* and *DAPK2* as mediators, which accounts for 61% of the correlation between TG and NAFLD (D1). **c**, The mediation analysis with BMI, serum TGs, and *DAPK2* adipose expression in the METSIM cohort. The model with serum TGs as the mediator between BMI and *DAPK2* expression is significant.

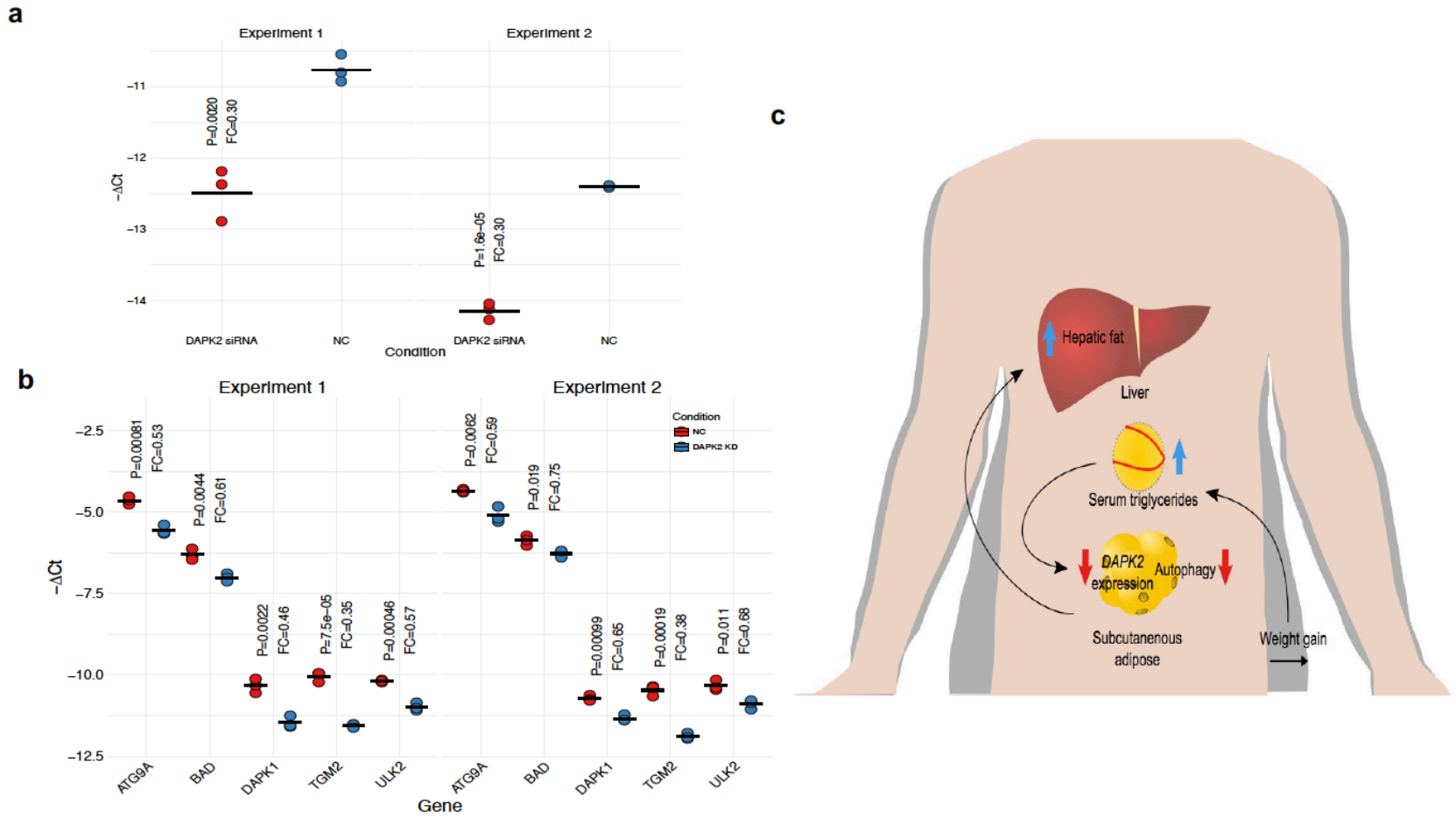
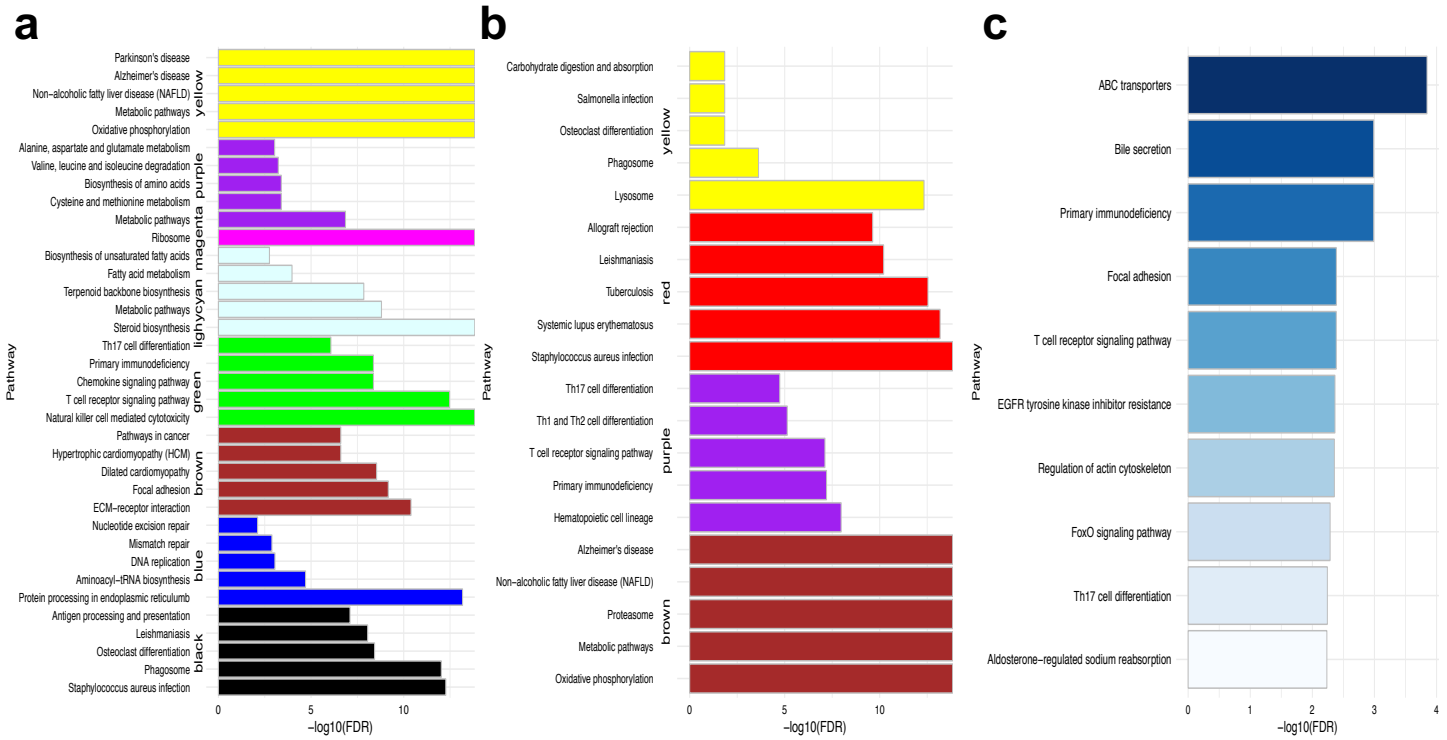
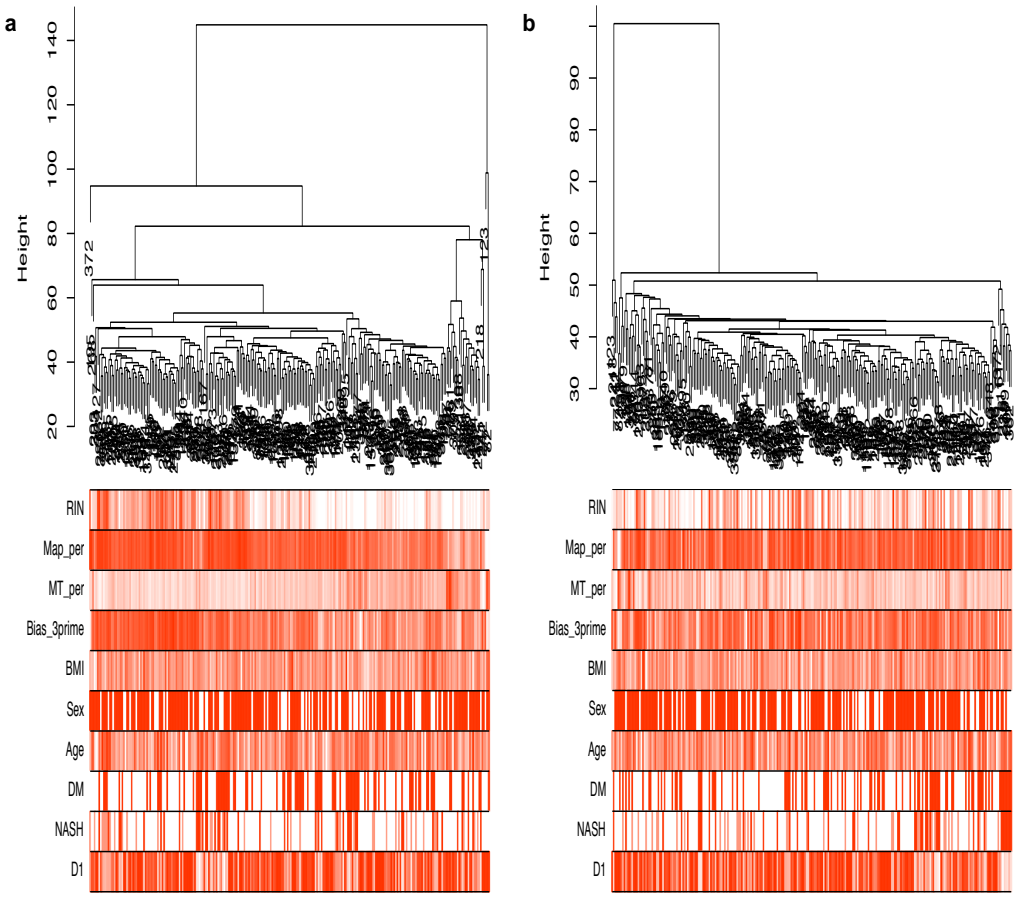


Figure 6. *DAPK2* knockdown by siRNA downregulates expression of key autophagy genes in human primary preadipocytes, providing biological support for the adipose-based role of *DAPK2* in fatty liver. a, siRNA knockdown reduced the *DAPK2* expression in primary human preadipocytes to around 30% of original level. Negative ΔCt values are shown to demonstrate the directionality of the change in expression. The mean negative ΔCt values are labeled by the horizontal black line, and p-values were calculated using two-sample t-test. **b**, The expressions of five autophagy pathway genes are consistently reduced (P<0.05) after *DAPK2* knockdown in two biologically independent human primary preadipocyte cell lines, each in triplicates. Negative ΔCt values are shown, and the horizontal black line indicates the mean. P-values were calculated using two-sample t-test. **c**, An illustration

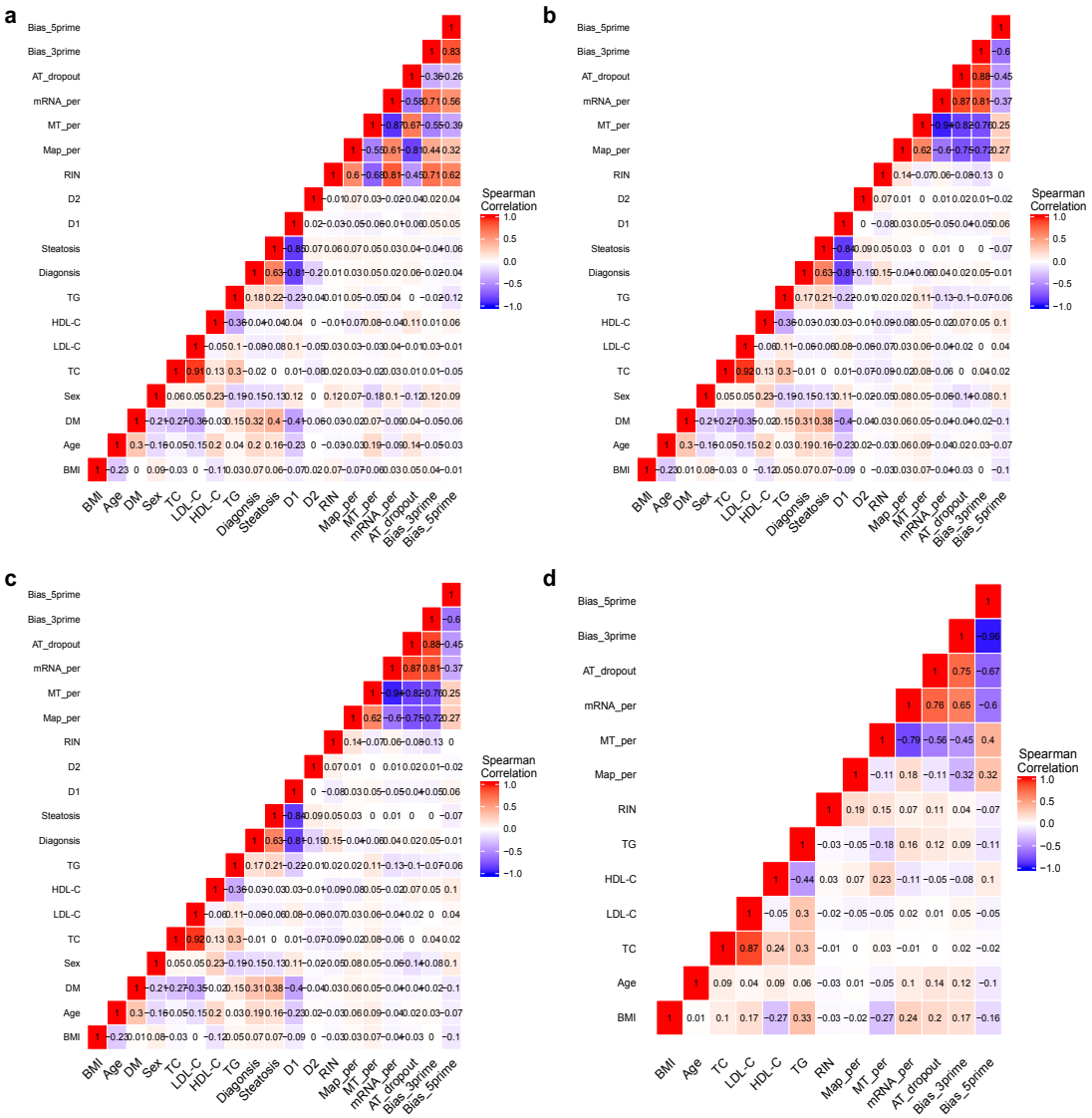
depicting the proposed causal pathway for obesity, adiposity and NAFLD. Weight gain causes the serum TG levels to increase, which in turn reduces the expression of *DAPK2* in the subcutaneous adipose tissue. Lower *DAPK2* adipose expression suppresses expression of autophagy pathway genes, some of which are involved in adipocyte differentiation, which impairs adipogenesis and lipogenesis of the adipose tissue. This cascade promotes the ectopic fat accumulation in the liver.



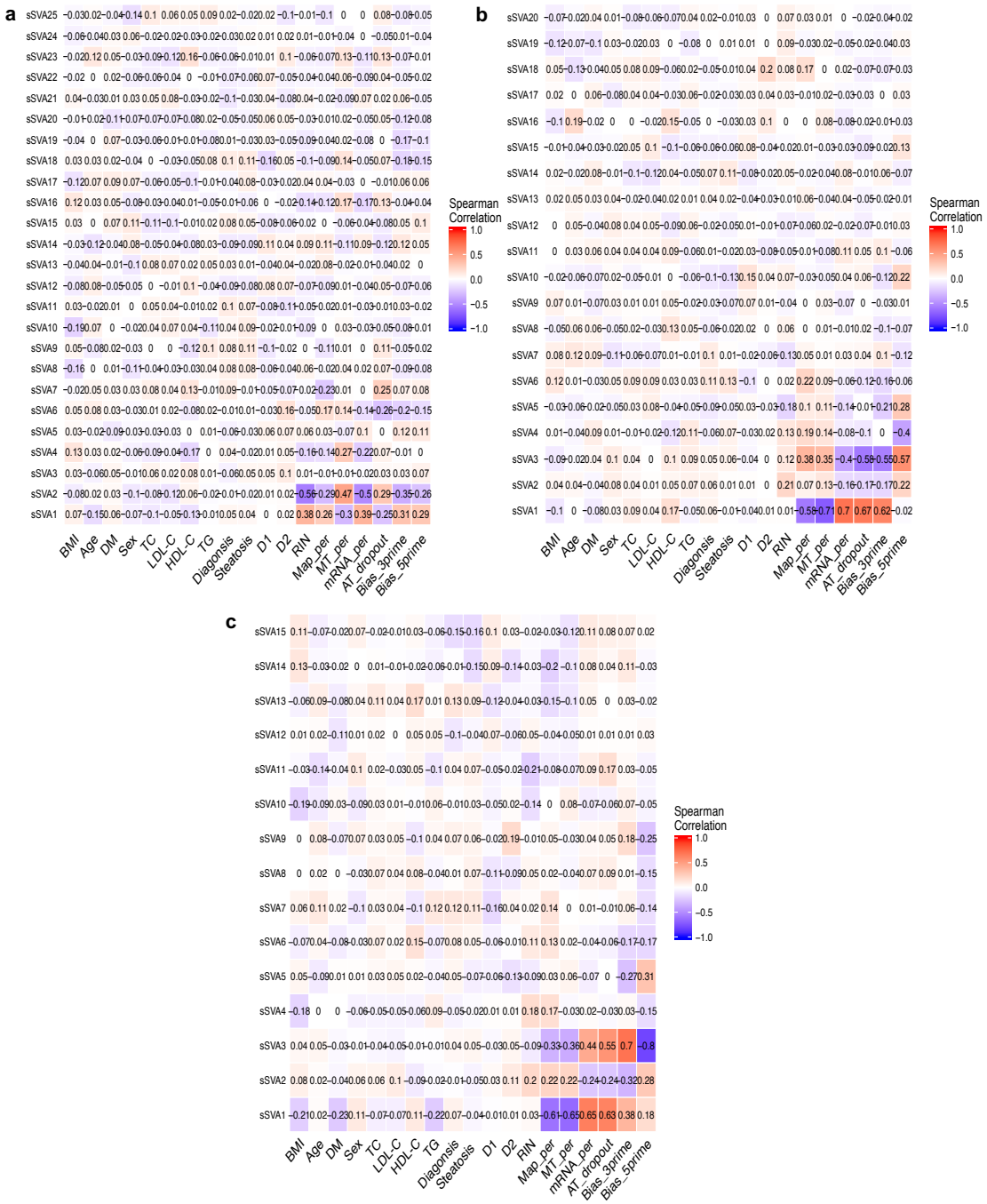
Supplementary Figure 1 | The KEGG pathway enrichment of the preserved WGCNA liver and adipose modules as well as liver NASH DE genes. a, The KEGG pathway enrichment of the preserved WGCNA liver modules. The cyan module is excluded due to no enrichment. Only the top 5 significant pathways (FDR<0.05) of each module are shown. **b**, The KEGG pathway enrichment of the preserved WGCNA adipose modules. Only the top 5 significantly enriched pathways (FDR<0.05) are shown for each module. **c**. The top 10 significantly enriched KEGG pathways (FDR<0.05) based on the liver DE genes are shown.



Supplementary Figure 2. The liver RNA-seq data before and after correcting for confounding technical factors that masked fatty liver effect.
a, The samples are mostly clustered based on RNA-seq technical factors such as RIN value, three prime bias (Bias_3prime), and uniquely aligned read percentage (Map_per). **b**, After adjusting for technical and phenotypic covariates, we observed a cluster of NASH and type 2 diabetes (DM) patients and no trend based on RNA-seq confounders.

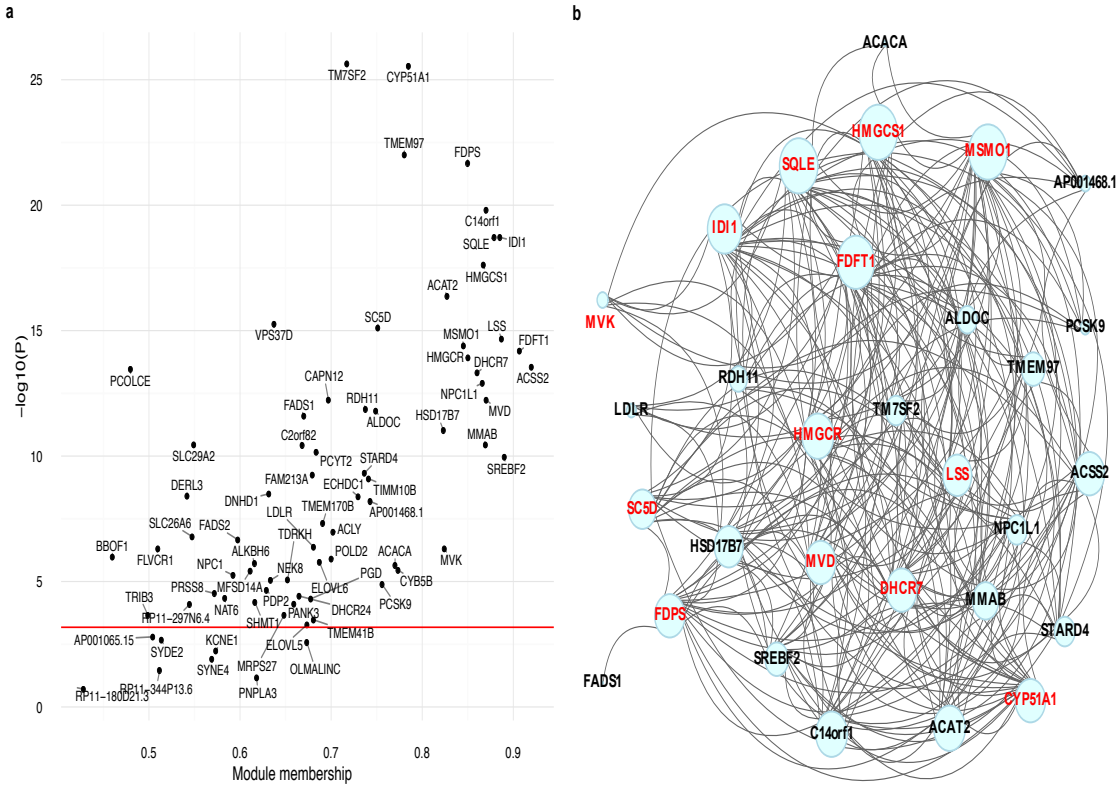


Supplementary Figure 3. RNA-seq technical covariates do not correlate with biological phenotypes. a, KOBS liver RNA-seq data (n=259). **b**, KOBS baseline adipose RNA-seq data (n=254). **c**, KOBS follow-up adipose RNA-seq data. **d**, METSIM adipose RNA-seq data (n=335). DM indicates type 2 diabetes status; TC, total cholesterol; LDL-C, low-density lipoprotein cholesterol; HDL-C, high-density lipoprotein cholesterol; TG, serum triglycerides; Diagnosis, NASH diagnosis; D1 and D2, the first two non-linear principal components of liver histology scores; Map_per, percent of uniquely aligned reads; mRNA_per, percent of reads aligned to mRNA regions; and AT_dropout, AT bias estimated by Picard, respectively.



Supplementary Figure 4. Latent variables from sSVA are highly correlated with RNA-seq technical attributes but not with biological phenotypes.

a, KOBS liver RNA-seq data. **b**, KOBS baseline adipose RNA-seq data. **c**, KOBS follow-up adipose RNA-seq data. DM indicates type 2 diabetes status; TC, total cholesterol; LDL-C, low-density lipoprotein cholesterol; HDL-C, high-density lipoprotein cholesterol; TG, serum triglycerides; Diagnosis, NASH diagnosis; D1 and D2, the first two non-linear principal components of liver histology scores; Map_per; percent of uniquely aligned reads; mRNA_per, percent of reads aligned to mRNA regions; and AT_dropout, AT bias estimated by Picard, respectively.



Supplementary Figure 5. Genes in the light cyan modules are strongly associated with statin medication and are involved in cholesterol synthesis. a, The strength of association with statin medication is highly correlated with the module membership of the light cyan module. The red line indicates the Bonferroni corrected p-value=0.05 threshold. **b**, The network representation of the light cyan module. For visualization purpose, we limited it to gene-gene connections with topological overlap at least 0.05, resulting in the top 30 genes shown here. Red-colored genes are part of the cholesterol biosynthesis pathway according to Wikipathway.

Supplementary note

Statin-associated module reveals potential new statin response genes

Since the liver is the key site of cholesterol synthesis and the target organ for statins, we further investigated the light cyan module that is significantly associated with statin medication and enriched for the steroid biosynthesis pathway (FDR<1.59×10^{-9}). Since T2D is the second most associated phenotype (Fig. 2a), we performed a conditional analysis, in which T2D was included as covariate and found that statin medication remained significantly associated (P=3.53×10^{-10}) with the light cyan module. Therefore, the statin association is not driven by the T2D signal. Conversely, when we conditioned the other significantly associated traits on statin medication, associations of this module with liver fibrosis and T2D were nominally significant (P=0.042 and P=0.012, respectively), and the p-value for serum TGs remained 0.00026. These findings suggest that except for serum TGs, the association signal for other liver traits are likely due to the high rate of statin usage in NAFLD patients.

We further tested which of the 75 genes are significantly associated with statin medication (Bonferroni corrected P<0.05) and found 67 of them to be correlated with statin medication individually (Supplementary Table 10). Hub genes with a high module membership also have the strongest statin associations (Supplementary Fig. 5a), whereas *PNPLA3,* the first known NAFLD gene[9], is not individually associated with statin medication even though it is one of the 75 genes (Supplementary Fig. 5a). However, the module is not significantly associated with total cholesterol (TC) or low-density lipoprotein cholesterol (LDL-C) (P>0.05) after excluding statin users, suggesting that the signal is likely attributed to statin medication; however, the decreased sample size of the non-statin users provides reduced power to detect the module association with LDL-C and TC. We also observed that, *HMGCR*, the therapeutic target

of statin, is among the hub genes, and 19 known cholesterol pathway genes are also part of the module (Supplementary Fig. 5b). The 67 statin-associated genes showed a large overlap with the statin response genes reported previously in the primary human hepatocytes[45]. This robust overlap together with the significant preservation of the statin response module that we observed in GTEx livers indicates that these module genes are responding to statins and that the effect of statin on liver gene expression is not perturbed by morbid obesity or NAFLD.

Given the strong association with statin medication and presence of cholesterol pathway genes, we examined whether the *cis* regulatory regions (within 1Mb of the gene) of these 67 genes contribute significantly to the heritability of LDL-C using LD score regression[46] based on previous LDL-C GWAS summary statistics[47]. All variants in the *cis* regions of the 67 genes account for 5.0% of the LDL-C heritability, representing a 2.5-fold enrichment over the genome-wide background (P=0.0012).

Supplementary Table 1. A summary of metabolic traits of the KOBS cohort.

Qualitative traits	Number	Quantitative traits	Mean (SD)
Sex	78 (30.1%) /181 (69.9%)	Age (years)	48.51 (9.0)
Steatosis (0:1:2:3)	102:93:34:30	BMI (kg/m^2)	43.03 (5.3)
Lobular inflammation (0:1:2:3)	177:65:17	LDL-C (mmol/L)	2.30 (0.8)
Ballooning (0:1:2)	194:63:2	HDL-C (mmol/L)	1.11 (0.3)
Fibrosis (0:1:2:3:4)	142:99:10:6:2	TC (mmol/L)	4.13 (0.9)
NASH (0:1:2)	175:41:43	TG (mmol/L)	1.59 (0.7)
T2D	96 (37.0%)	Glucose (mmol/L)	6.5 (1.9)
NASH+T2D	29 (11.1%)	ALT (IU/L)	42.60 (28.2)

T2D indicates type 2 diabetes; NASH, non-alcoholic steatohepatitis; BMI, body mass index; LDL-C, low-density lipoprotein cholesterol; HDL-C, high-density lipoprotein cholesterol; TC, total cholesterol; TG, serum triglycerides; and ALT, alanine aminotransferase, respectively. The scores of steatosis, lobular inflammation, ballooning, and fibrosis indicate the CNR NALFD activity score and fibrosis staging. The scores of 0, 1, and 2 for NASH stand for no steatohepatitis, possible, and NASH, respectively. SD=standard deviation.

Supplementary Table 7. Mouse adipose gene expression associations with hepatic TG content.

Probe	Gene	Effect size	SE	P
1448130_at*	*Fdft1*	-0.53	0.058	7.55×10^{-17}
1438322_x_at*	*Fdft1*	-0.51	0.059	1.20×10^{-15}
1455972_x_at*	*Hadh*	-0.40	0.061	2.77×10^{-10}
1460184_at*	*Hadh*	-0.38	0.0621	6.93×10^{-9}
1436756_x_at*	*Hadh*	-0.34	0.064	2.13×10^{-7}
1449160_at*	*Npr1*	-0.33	0.066	1.41×10^{-6}
1456676_a_at*	*Pfkfb3*	-0.28	0.069	6.30×10^{-5}
1451453_at*	*Dapk2*	-0.23	0.070	0.0011
1416432_at*	*Pfkfb3*	-0.22	0.071	0.0023
1460684_at	*Tm7sf2*	-0.18	0.066	0.0060
1422042_at	*Gjc3*	-0.18	0.071	0.014
1451060_at	*Gpr146*	-0.13	0.070	0.071
1421944_a_at	*Asgr1*	0.072	0.072	0.32
1425004_s_at	*Mocs1*	-0.062	0.072	0.39
1450717_at	*Ang*	0.057	0.072	0.43
1455593_at	*Apob*	0.026	0.072	0.71
1423632_at	*Gpr146*	-0.0048	0.070	0.94

Table is sorted by p-value. SE indicates standard error estimated by linear regression; P, P-value of the probe in the linear regression and *significant probes using Bonferroni correction for 17 tests, respectively.

Supplementary Table 9. A summary of RNA-seq library types, platform, and coverage.

	KOBS liver	KOBS baseline adipose	KOBS follow-up adipose	METSIM adipose
Library type	Ribo-Zero stranded	TruSeq stranded	TruSeq stranded	TruSeq unstranded
Sequencing platform	HiSeq2500	HiSeq4000	HiSeq4000	HiSeq2000
Read length	50bp	69bp	69bp	50bp
ERCC spike-in	Yes	Yes	Yes	No
Mean read-pair count	39.73M (5.83M)	42.38M (7.39M)	43.57M (7.91M)	47.46M (9.74M)
Unique alignment rate	81.91 (5.0)	89.94 (4.9)	90.17 (2.8)	85.30 (4.1)
# samples passed QC	259	254	169	335

The standard deviations of read-pair counts and alignment rate are shown in parenthesis. M indicates million read pairs.

References

1. Pellicoro, A., Ramachandran, P., Iredale, J.P., and Fallowfield, J.A. (2014). Liver fibrosis and repair: immune regulation of wound healing in a solid organ. Nat. Rev. Immunol. *14*, 181–194.

2. NCD Risk Factor Collaboration, Di Cesare, M., Bentham, J., Stevens, G.A., Zhou, B., Danaei, G., Lu, Y., Bixby, H., Cowan, M.J., Riley, L.M., et al. (2016). Trends in adult body-mass index in 200 countries from 1975 to 2014: a pooled analysis of 1698 population-based measurement studies with 19·2 million participants. Lancet *387*, 1377–1396.

3. GBD 2015 Obesity Collaborators, Afshin, A., Forouzanfar, M.H., Reitsma, M.B., Sur, P., Estep, K., Lee, A., Marczak, L., Mokdad, A.H., Moradi-Lakeh, M., et al. (2017). Health Effects of Overweight and Obesity in 195 Countries over 25 Years. N. Engl. J. Med. *377*, 13–27.

4. Younossi, Z.M., Koenig, A.B., Abdelatif, D., Fazel, Y., Henry, L., and Wymer, M. (2016). Global epidemiology of nonalcoholic fatty liver disease—Meta-analytic assessment of prevalence, incidence, and outcomes. Hepatology *64*, 73–84.

5. Allen, A.M., Terry, T.M., Larson, J.J., Coward, A., Somers, V.K., and Kamath, P.S. (2017). Nonalcoholic Fatty Liver Disease Incidence and Impact on Metabolic Burden and Death: a 20 Year-Community Study. Hepatology.

6. Spengler, E.K., and Loomba, R. (2015). Recommendations for Diagnosis, Referral for Liver Biopsy, and Treatment of Nonalcoholic Fatty Liver Disease and Nonalcoholic Steatohepatitis. Mayo Clin. Proc. *90*, 1233–1246.

7. Williams, C.D., Stengel, J., Asike, M.I., Torres, D.M., Shaw, J., Contreras, M., Landt, C.L., and Harrison, S.A. (2011). Prevalence of Nonalcoholic Fatty Liver Disease and Nonalcoholic Steatohepatitis Among a Largely Middle-Aged Population Utilizing Ultrasound and Liver Biopsy: A Prospective Study. Gastroenterology *140*, 124–131.

8. Oliveira, C.P., Stefano, J.T., and Carrilho, F.J. (2017). Clinical patterns of hepatocellular carcinoma (HCC) in non-alcoholic fatty liver disease (NAFLD): a multicenter prospective study. Hepatobiliary Surg Nutr *6*, 350–352.

9. Romeo, S., Kozlitina, J., Xing, C., Pertsemlidis, A., Cox, D., Pennacchio, L.A., Boerwinkle, E., Cohen, J.C., and Hobbs, H.H. (2008). Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. Nat. Genet. *40*, 1461–1465.

10. Kozlitina, J., Smagris, E., Stender, S., Nordestgaard, B.G., Zhou, H.H., Tybjærg-Hansen, A., Vogt, T.F., Hobbs, H.H., and Cohen, J.C. (2014). Exome-wide association study identifies a TM6SF2 variant that confers susceptibility to nonalcoholic fatty liver disease. Nat. Genet. *46*, 352–356.

11. Adams, L.A., Anstee, Q.M., Tilg, H., and Targher, G. (2017). Non-alcoholic fatty liver disease and its relationship with cardiovascular disease and other extrahepatic diseases. Gut *66*, 1138–1153.

12. Eslam, M., Valenti, L., and Romeo, S. (2017). Genetics and epigenetics of NAFLD and NASH: Clinical impact. J. Hepatol.

13. Hui, S.T., Parks, B.W., Org, E., Norheim, F., Che, N., Pan, C., Castellani, L.W., Charugundla, S., Dirks, D.L., Psychogios, N., et al. (2015). The genetic architecture of NAFLD among inbred strains of mice. Elife *4*.

14. Chalasani, N., Guo, X., Loomba, R., Goodarzi, M.O., Haritunians, T., Kwon, S., Cui, J., Taylor, K.D., Wilson, L., Cummings, O.W., et al. (2010). Genome-Wide Association Study Identifies Variants Associated With Histologic Features of Nonalcoholic Fatty Liver Disease. Gastroenterology *139*, 1567–.

15. Soussi, H., Reggio, S., Alili, R., Prado, C., Mutel, S., Pini, M., Rouault, C., Clément, K., and Dugail, I. (2015). DAPK2 Downregulation Associates With Attenuated Adipocyte Autophagic Clearance in Human Obesity. Diabetes *64*, 3452–3463.

16. Hicks, R., and Tingley, D. (2011). Causal mediation analysis. Stata Journal *11*, 605–619.

17. MacKinnon, D.P., and Valente, M.J. (2014). Mediation from Multilevel to Structural Equation Modeling. Ann. Nutr. Metab. *65*, 198–204.

18. Ber, Y., Shiloh, R., Gilad, Y., Degani, N., Bialik, S., and Kimchi, A. (2015). DAPK2 is a novel regulator of mTORC1 activity and autophagy. Cell Death Differ. *22*, 465–475.

19. Myneni, V.D., Melino, G., and Kaartinen, M.T. (2015). Transglutaminase 2-a novel inhibitor of adipogenesis. Cell Death Dis *6*, –e1868.

20. Ro, S.-H., Jung, C.H., Hahn, W.S., Xu, X., Kim, Y.-M., Yun, Y.S., Park, J.-M., Kim, K.H., Seo, M., Ha, T.-Y., et al. (2013). Distinct functions of Ulk1 and Ulk2 in the regulation of lipid metabolism in adipocytes. Autophagy *9*, 2103–2114.

21. Henegar, C., Tordjman, J., Achard, V., Lacasa, D., Cremer, I., Guerre-Millo, M., Poitou, C., Basdevant, A., Stich, V., Viguerie, N., et al. (2008). Adipose tissue transcriptomic signature highlights the pathological relevance of extracellular matrix in human obesity. Genome Biol. *9*.

22. Baerga, R., Zhang, Y., Chen, P.-H., Goldman, S., and Jin, S. (2009). Targeted deletion of autophogy-related 5 (atg5) impairs adipogenesis in a cellular model and in mice. Autophagy *5*, 1118–1130.

23. Singh, R., Xiang, Y., Wang, Y., Baikati, K., Cuervo, A.M., Luu, Y.K., Tang, Y., Pessin, J.E., Schwartz, G.J., and Czaja, M.J. (2009). Autophagy regulates adipose mass and differentiation in mice. J. Clin. Invest. *119*, 3329–3339.

24. Zhang, Y., Goldman, S., Baerga, R., Zhao, Y., Komatsu, M., and Jin, S. (2009). Adipose-specific deletion of autophagy-related gene 7 (atg7) in mice reveals a role in adipogenesis. Proc. Natl. Acad. Sci. U.S.a. *106*, 19860–19865.

25. Guo, L., Huang, J.-X., Liu, Y., Li, X., Zhou, S.-R., Qian, S.-W., Liu, Y., Zhu, H., Huang, H.-Y., Dang, Y.-J., et al. (2013). Transactivation of Atg4b by C/EBPβ promotes autophagy to facilitate adipogenesis. Mol. Cell. Biol. *33*, 3180–3190.

26. Mannisto, V.T., Simonen, M., Hyysalo, J., Soininen, P., Kangas, A.J., Kaminska, D., Matte, A.K., Venesmaa, S., Kakela, P., Karja, V., et al. (2015). Ketone body production is differentially altered in steatosis and non-alcoholic steatohepatitis in obese humans. Liver Int. *35*, 1853–1861.

27. Stančáková, A., Civelek, M., Saleem, N.K., Soininen, P., Kangas, A.J., Cederberg, H., Paananen, J., Pihlajamäki, J., Bonnycastle, L.L., Morken, M.A., et al. (2012). 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.

28. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet. *45*, 580–585.

29. Kleiner, D.E., Brunt, E.M., Van Natta, M., Behling, C., Contos, M.J., Cummings, O.W., Ferrell, L.D., Liu, Y.C., Torbenson, M.S., Unalp-Arida, A., et al. (2005). Design and validation of a histological scoring system for nonalcoholic fatty liver disease. Hepatology *41*, 1313–1321.

30. Brunt, E.M. (2009). Histopathology of Non-Alcoholic Fatty Liver Disease. Clinics in Liver Disease *13*, 533–.

31. Leeuw, J. de, and Mair, P. (2009). Gifi Methods for Optimal Scaling in R: The Package homals. Journal of Statistical Software *31*, 1–21.

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

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

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

35. Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol *34*, 525–527.

36. Leek, J.T. (2014). svaseq: removing batch effects and other unwanted noise from sequencing data. Nucleic Acids Research *42*, –e161.

37. Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics.

38. Law, C.W., Chen, Y., Shi, W., and Smyth, G.K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. *15*.

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

40. Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics *9*, 559.

41. Langfelder, P., Luo, R., Oldham, M.C., and Horvath, S. (2011). Is my network module preserved and reproducible? PLoS Comput Biol *7*, e1001057.

42. Bennett, B.J., Farber, C.R., Orozco, L., Kang, H.M., Ghazalpour, A., Siemers, N., Neubauer, M., Neuhaus, I., Yordanova, R., Guan, B., et al. (2010). A high-resolution association mapping panel for the dissection of complex traits in mice. Genome Res. *20*, 281–290.

43. Tofighi, D., and MacKinnon, D.P. (2011). RMediation: an R package for mediation analysis confidence intervals. Behav Res Methods *43*, 692–700.

44. Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. Journal of Statistical Software *48*, 1–36.

45. Hafner, M., Juvan, P., Rezen, T., Monostory, K., Pascussi, J.-M., and Rozman, D. (2011). The human primary hepatocyte transcriptome reveals novel insights into atorvastatin and rosuvastatin action. Pharmacogenet. Genomics *21*, 741–750.

46. Bulik-Sullivan, B.K., Loh, P.-R., Finucane, H.K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291–295.

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

Chapter 7

Conclusions and future directions

Since the inception of genome-wide association studies (GWAS)[1], the field of human genetics has made great progress in understanding the genetic architecture of many human traits and diseases. It has, however, become increasingly clear that unlike initially envisioned, GWAS is not a silver bullet that can answer many genetic questions. Nevertheless, GWAS is still a very powerful exploratory approach to guide downstream biological or clinical studies[2,3], especially with the establishment of many large biobanks or electronic health record (EHR) databases that provide great power to detect small effect-size variants. In this dissertation, I have discussed some shortcomings of GWAS and presented our integrative approaches to augment or go beyond GWAS to understand human disease mechanisms. In particular, we designed a new approach, cross-population allele screen (CPAS)[4], to search for population-specific variants associated with traits. We also developed two computational tools that combine GWAS summary statistics or genotype data with transcriptomes to estimate gene expression-trait associations[5] and allele-specific expression (ASE)[6], respectively. In Chapters 5 and 6, we utilized transcriptomic and epigenomic data to identify the regulatory mechanisms of obesity GWAS variants[7] and the obesity-driven pathogenesis of non-alcoholic fatty liver (NAFLD) (Chapter 6). In this final remark, I will address some limitations of our integrative approaches and offer future prospects and directions for our research.

We developed CPAS to identify population-specific variants that might contribute to the difference in quantitative traits or disease susceptibility among ethnic groups. The difference in allele frequency of a variant between populations may not only reflect the demographic history of a population[8,9] but could also have undergone natural selection, driven by environmental factors[10]. One limitation is that our CPAS-GWAS analysis did not interrogate any low-frequency

or rare variants (minor allele frequency <5% and <1%, respectively), which are more likely to differ in frequency among populations[11]. To this end, we are performing targeted and exome sequencing to identify coding variants in loci that harbor common population-specific variants, associated with metabolic traits[4]. In addition, recent genomic initiatives, such as the Trans-omics for Precision Medicine (TOPMed), supported by the National Heart, Lung, and Blood Institute, are performing whole-genome sequencing, genome-wide methylation, RNA-sequencing, as well as collecting phenotypes from diverse populations, which will greatly help cataloging rare variants. During the past decade, we have increased the number of GWAS on non-European people, but similar projects and efforts to TOPMed and All of Us that advocate the inclusion of more diverse populations are needed to decrease this disparity in genomic medicine[12,13].

To better understand the mechanism of cardiovascular disease, researchers are shifting the focus from identifying disease-associated DNA variants to examining molecular phenotypes, such as gene expression and epigenomic modification[14]. In Chapters 3 and 4, I introduced our two newly developed computational tools, Functional Summary-based Imputation (FUSION)[5] and ASElux[6] that integrate genotype and transcriptomic data to infer genes or functional variants underlying a GWAS locus. FUSION enables large-scale transcript-wide association studies (TWAS) without the burden of profiling many transcriptomes. However, FUSION only utilizes *cis* variants (usually within 1 Mb of the gene) to impute the expression-trait associations, even though distal or *trans* variants are estimated to explain more than 50% heritability of gene expression[15,16]. Future development of FUSION could incorporate genome-wide variants to estimate the trait-expression association for each gene, but this will likely require much larger reference panels to obtain robust models between gene expression and genotype. We also note that imputed TWAS via FUSION will only account for the genetic component of gene

regulation, and thus environmental effects on expression will not be captured. However, the advantage of requiring just a small reference panel and the flexibility to utilize additional molecular phenotypes, such as methylation and transcription factor binding data, will allow FUSION to be applicable to many human diseases and traits.

Our new method, ASElux, can estimate allele-specific expression (ASE) much faster than many standard alignment methods while achieving similar accuracy. Nevertheless, we acknowledge that there are still complications when performing an ASE analysis. Particularly, it is unclear how to properly handle multi-aligned reads, as they can lead to false ASE calculation. ASElux tries to align a read to as many loci as possible and excludes the read if it aligns to multiple regions. However, these multi-aligned reads then become a wasted resource or can lead to underestimated ASE. Sequencing mRNA transcripts with much longer reads can partially resolve this issue. Another disadvantage of most ASE methods and ASElux is the requirement of external genotype data, which may be unavailable due to various causes, such as financial constraint. Directly calling single nucleotide polymorphism (SNP) from RNA-sequence (RNA-seq) data will greatly reduce the cost of ASE analysis but requires additional method development. Given the speed and performance of ASElux, we envision ASElux as a computational tool that can extend ASE analysis to many large-scale RNA-seq cohorts. Furthermore, ASElux can also be expanded to analyze other high-throughput functional assays, such as chromatin immunoprecipitation sequencing (ChIP-seq) and assay for transposase-accessible chromatin using sequencing (ATAC-seq), to identify allelic-specific transcription factor (TF) binding or open chromatin sites.

In Chapters 5 and 6, I provided examples of our work of using transcriptomic and epigenomic data to understand the molecular mechanism of obesity and NAFLD. To determine

the functional variants and their underlying mechanisms at the obesity GWAS loci, we integrated gene expression and expression quantitative trait loci (eQTLs) data from human subcutaneous adipose tissue with chromosomal interaction data from human primary adipocytes. We utilized the more cost-effective promoter capture Hi-C method to specifically assay promotor-enhancer interactions. A follow-up experiment using deeply sequenced data from whole-genome Hi-C[17] to catalog a more comprehensive chromatin interaction map of adipocytes might help us identify additional functional variants that do not act through canonical promoters.

In our obesity study, we demonstrated that the alternative allele of the SNP, rs4776984, exhibits a preferential protein binding at the mitogen-activated protein kinase 5 (*MAP2K5*) locus. However, it remains unclear what protein or protein complex targets this regulatory variant. Performing ChIP-seq targeting the predicted TFs that bind at this site in adipocytes can potentially elucidate the true regulatory protein. We can also determine the allelic effect of the SNP by applying Hi-C and ChIP-seq on a heterozygous carrier and estimate allele-specific TF-binding or chromatin-looping to further elucidate the functional role of the SNP. Additional similar analyses and experiments can also further elucidate the molecular mechanisms of the other 41 new obesity genes that we identified to be under *cis* genetic regulation via chromosomal interactions[7].

Our genomic and experimental data from human and mouse support the role of death associated protein kinase 2 (*DAPK2*) in the pathogenesis of NAFLD. We leveraged the longitudinal information from the bariatric surgery cohort to help orient the direction of our causal analyses. We showed that reducing *DAPK2* expression in human primary preadipocytes leads to lower expression of five key autophagy pathway genes. In particular, two of the autophagy genes are also involved in adipocyte differentiation. We thus proposed that impaired

autophagy and adipocyte differentiation in the adipose tissue due to obesity-induced *DAPK2* down-regulation could lead to fat accumulation in the liver. To fully support this hypothesis and the causal effect of *DAPK2* in the pathogenesis of NAFLD, we are currently performing an adipose-specific *Dapk2* knockdown experiment in mice that are fed with a high fat diet to first induce obesity and then measure their hepatic TG content. Future studies are warranted to further explore the potential of using adipose *DAPK2* expression as a biomarker or a therapeutic target for NAFLD in humans.

In this dissertation, I illustrated methods and examples utilizing integrative genomic approaches to shed light into the molecular mechanisms of genetic variants in human disease. However, generating and analyzing functional genomic data still present technical and methodological challenges. In particular, when sampling from human tissues, cell type heterogeneity might lead to noisy data. Computational tools have been developed to decompose bulk RNA-seq[18-20] or methylation data[21,22] in order to estimate the cell-type composition or transcript/methylation levels within each cell type of a heterogenous tissue. With the recent emergence of more efficient single-cell RNA-seq technology, we can now assay transcriptome[23], methylation[24] , and open chromatin[25] at the single cell level. These data can provide a high resolution to examine cell-type specific transcriptomic and epigenomic phenotypes.

Another difficulty for human functional genomic studies is the accessibility of tissue. For example, neurodevelopmental and cardiometabolic disorders involve tissues, such as the brain or heart, that can only be sampled during a surgery or an autopsy. With the introduction of human induced pluripotent stem cells (iPSCs), we now have the alternative to collect more accessible cells, such as fibroblasts[26,27], and reprogram them into iPSCs that can be further differentiated into the cell-type of interest. In this exciting post-GWAS era, new and more efficient functional

genomic assays are rapidly being developed. As we continue to generate more functional data, new computational tools must also be developed to enable integrative data analysis and improve the interpretation. Our work presented here provides a framework and tools to integrate genomic data and understand the molecular mechanisms of disease. In the future, large genomic initiatives, such as the UK Biobank[28], TOPMed, and All of Us, will not only help provide a more comprehensive genetic catalog and better power for association studies in diverse populations, but also generate molecular phenotypes that will enable integrative approaches to delineate the molecular mechanism of variants and genes. In conjunction with emerging new molecular and cellular technologies including single-cell -omics and iPSCs, we can now circumvent some of the previous limitations of functional genomics and further translate findings from GWAS into clinical knowledge.

References

1. Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGiovanni, J.P., Mane, S.M., Mayne, S.T., et al. (2005). Complement factor H polymorphism in age-related macular degeneration. Science *308*, 385–389.

2. Visscher, P.M., Brown, M.A., McCarthy, M.I., and Yang, J. (2012). Five Years of GWAS Discovery. Am. J. Hum. Genet. *90*, 7–24.

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

4. Ko, A., Cantor, R.M., Weissglas-Volkov, D., Nikkola, E., Reddy, P.M.V.L., Sinsheimer, J.S., Pasaniuc, B., Brown, R., Alvarez, M., Rodriguez, A., et al. (2014). Amerindian-specific regions under positive selection harbour new lipid variants in Latinos. Nat Commun *5*.

5. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W.J.H., Jansen, R., de Geus, E.J.C., Boomsma, D.I., Wright, F.A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. Nat. Genet. *48*, 245–252.

6. Miao, Z., Alvarez, M., Pajukanta, P., and Ko, A. (2018). ASElux: an ultra-fast and accurate allelic reads counter. Bioinformatics *34*, 1313–1320.

7. Pan, D.Z., Garske, K.M., Alvarez, M., Bhagat, Y.V., Boocock, J., Nikkola, E., Miao, Z., Raulerson, C.K., Cantor, R.M., Civelek, M., et al. (2018). Integration of human adipocyte chromosomal interactions with adipose gene expression prioritizes obesity-related genes from GWAS. Nat Commun *9*, 1512.

8. Barreiro, L.B., Laval, G., Quach, H., Patin, E., and Quintana-Murci, L. (2008). Natural selection has driven population differentiation in modern humans. Nat. Genet. *40*, 340–345.

9. Novembre, J., and Ramachandran, S. (2011). Perspectives on Human Population Structure at the Cusp of the Sequencing Era. Annu Rev Genomics Hum Genet *12*, 245–274.

10. Guo, J., Wu, Y., Zhu, Z., Zheng, Z., Trzaskowski, M., Zeng, J., Robinson, M.R., Visscher, P.M., and Yang, J. (2018). Global genetic differentiation of complex traits shaped by natural selection in humans. Nat Commun *9*, 1865.

11. Bentley, D.R., Donnelly, P., Lehrach, H., Nickerson, D.A., Schmidt, J.P., Dinh, H., Lee, S., Muzny, D., Wang, M., Fang, X., et al. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

12. Popejoy, A.B., and Fullerton, S.M. (2016). Genomics is failing on diversity. Nature *538*, 161–164.

13. Need, A.C., and Goldstein, D.B. (2009). Next generation disparities in human genomics: concerns and remedies. Trends Genet. *25*, 489–494.

14. Musunuru, K., Ingelsson, E., Fornage, M., Liu, P., Murphy, A.M., Newby, L.K., Newton-Cheh, C., Perez, M.V., Voora, D., Woo, D., et al. (2017). The Expressed Genome in Cardiovascular Diseases and Stroke: Refinement, Diagnosis, and Prediction A Scientific Statement From the American Heart Association. Circ Cardiovasc Genet *10*, E1–E24.

15. Price, A.L., Helgason, A., Thorleifsson, G., McCarroll, S.A., Kong, A., and Stefansson, K. (2011). Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. PLoS Genet. *7*, e1001317.

16. Albert, F.W., Bloom, J.S., Siegel, J., Day, L., and Kruglyak, L. (2017). Genetics of trans-regulatory variation in gene expression. bioRxiv 208447.

17. Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293.

18. Li, Y., and Xie, X. (2013). A mixture model for expression deconvolution from RNA-seq in heterogeneous tissues. BMC Bioinformatics *14*.

19. Gong, T., and Szustakowski, J.D. (2013). DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. Bioinformatics *29*, 1083–1085.

20. Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., and Alizadeh, A.A. (2015). Robust enumeration of cell subsets from tissue expression profiles. Nat Meth *12*, 453–.

21. Houseman, E.A., Accomando, W.P., Koestler, D.C., Christensen, B.C., Marsit, C.J., Nelson, H.H., Wiencke, J.K., and Kelsey, K.T. (2012). DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics *13*.

22. Houseman, E.A., Kile, M.L., Christiani, D.C., Ince, T.A., Kelsey, K.T., and Marsit, C.J. (2016). Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. BMC Bioinformatics *17*, 259.

23. Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., et al. (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science *343*, 776–779.

24. Zhu, P., Guo, H., Ren, Y., Hou, Y., Dong, J., Li, R., Lian, Y., Fan, X., Hu, B., Gao, Y., et al. (2018). Single-cell DNA methylome sequencing of human preimplantation embryos. Nat. Genet. *50*, 12–19.

25. Buenostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y., and Greenleaf, W.J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. Nature *523*, 486–490.

26. Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell *131*, 861–872.

27. Park, I.-H., Zhao, R., West, J.A., Yabuuchi, A., Huo, H., Ince, T.A., Lerou, P.H., Lensch, M.W., and Daley, G.Q. (2008). Reprogramming of human somatic cells to pluripotency with defined factors. Nature *451*, 141–U141.

28. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). 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.