UNIVERSITY OF CALIFORNIA

Los Angeles

Integrating molecular phenotypes and gene expression to characterize DNA variants for cardiometabolic traits

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Human Genetics

by

Alejandra Rodriguez

2018

ABSTRACT OF THE DISSERTATION

Integrating molecular phenotypes and gene expression to characterize DNA variants for cardiometabolic traits

by

Alejandra Rodriguez

Doctor of Philosophy in Human Genetics

University of California, Los Angeles, 2018

Professor Päivi Elisabeth Pajukanta, Chair

In-depth understanding of cardiovascular disease etiology requires characterization of its genetic, environmental, and molecular architecture. Genetic architecture can be defined as the characteristics of genetic variation responsible for broad-sense phenotypic heritability. Massively parallel sequencing has generated thousands of genomic datasets in diverse human tissues. Integration of such datasets using data mining methods has been used to extract biological meaning and has significantly advanced our understanding of the genome-wide nucleotide sequence, its regulatory elements, and overall chromatin architecture. This dissertation presents integration of "omics" data sets to understand the genetic architecture and molecular mechanisms of cardiovascular lipid disorders (further reviewed in Chapter 1).

In 2013, Daphna Weissglas-Volkov and coworkers[1] published an association between the chromosome 18q11.2 genomic region and hypertriglyceridemia in a genome-wide association

study (GWAS) of Mexican hypertriglyceridemia cases and controls. In chapter 2, we present the fine-mapping and functional characterization of the molecular mechanisms underlying this triglyceride (TG) association signal on chromosome 18q11.2[2]. Specifically, we found nine additional variants in linkage disequilibrium (LD) with the lead single nucleotide polymorphism (SNP). Using luciferase transcriptional reporter assays, electrophoretic mobility shift assays, and HNF4 ChIP-qPCR (chromatin immunoprecipitation coupled with quantitative polymerase chain reaction), we found that the minor G allele of rs17259126 disrupts an HNF4A binding site. Furthermore, using *cis* expression quantitative trait locus (eQTL) analysis, we found that the G allele of rs17259126 is associated with decreased expression of the regional transmembrane protein 241 (TMEM241) gene[2]. Our results suggest that reduced transcript levels of TMEM241 likely contribute to the increased serum TG levels in Mexicans.

GWAS variants typically have small effect sizes, and about 40% of them are located in intergenic regions and 40% in intronic regions. Since a large number of GWAS variants reside in non-coding regions, these SNPs are thought to affect gene regulation via disruption of functional elements, such as transcription factor binding sites (TFBS). Mapping genome-wide TFBS using chromatin immunoprecipitation followed by sequencing (ChIP-Seq) can identify such binding sites for specific transcription factors (TFs). It can also help identify unknown TF targets, complex interaction networks, and hub genes that can ultimately lead to the discovery of pharmaceutical targets. In Chapter 3, we present our results of the investigation of genome-wide targets of the RAR Related Orphan Receptor A (RORA), a high-density lipoprotein cholesterol (HDL-C) GWAS gene in Mexicans[3] and a known regulator of the apolipoproteins, APOA5, APOA1, and APOC3.

Despite the several hundred lipid loci identified by GWAS, it has become increasingly clear that variation at these known loci explains only a small fraction of the trait heritability. In addition to rare variants, contributions to variation in lipid traits that can be attributed to complex genetic models, such as gene-environment and epistatic interactions, have been hypothesized to be additional sources of this "missing heritability." In chapter 4, we present our findings of the investigation of genes that exhibit context-dependent expression variance and their underlying variance expression quantitative trait loci (ve-QTLs). Our cohort consisted of Mexicans exhibiting extreme TG values with subcutaneous adipose tissue expression microarrays available for study. We found that individuals with low serum TGs displayed a greater ATP citrate lyase (ACLY) expression variance than the individuals with high TGs. We replicated this observation in the Finnish METabolic Syndrome In Men[4] (METSIM) adipose RNA-Sequence cohort (p-value=$1.8x10^{-3}$). ACLY encodes the primary enzyme responsible for the synthesis of cytosolic acetyl-CoA in many tissues, which is vital for the biosynthesis of fatty acids, a precursor of TGs. One hypothesis is that reduced ACLY expression variance under increased TG context leads to an increased degree of constraint in lipid biosynthesis pathways, followed by decreased robustness in its response to environmental stimuli and buffering ability against cryptic genetic variation. We used a correlation least squared (CLS) test and found that the reference allele of variant rs34272903 (T/C) is associated with an increased ACLY expression variance (FWER p-value=$1.0x10^{-4}$). Our results suggest that the reference T allele of rs34272903 interacts with an unknown factor under the low TG context, increasing ACLY expression variance. This interaction may contribute to efficient responses in the lipid pathway activation to endo-exogenous stimuli via unknown mechanisms.

The dissertation of Alejandra Rodriguez is approved.

Jeff S. Abramson

Rita M. Cantor

Janet S. Sinsheimer

Päivi Elisabeth Pajukanta, Committee Chair

University of California, Los Angeles

2018

DEDICATION

I dedicate this thesis to my mother Elisa, for everything you have done for me.

To my grandmother Roselia who thought me never to give up.

To my husband and friend Saul for your love, mentorship, and support.

To Samuel and Yaren, I pray one day you will super-exceed this work beyond my imagination.

To JMS for never giving up on me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to thank:

My father Jose Luis for inspiring me to love nature.

My brothers, and sisters for their unconditional love, support, and encouragement.

All of the Camarena family members their love and support.

My church Fe Esperanza Y Amor for your unconditional support.

My mentor Paivi Pajukanta, an inspiring and young-spirited visionary.

My Committee members for their guidance and encouragement.

John Roche for your unconditional support.

Mayumi Prins for your support during the graduate application process.

Laurent Vergnes for the many great science conversations and the knowledge you shared.

Marcus Alvarez for always willing to help with R.

Zong Miao for all your help with Bash.

All the fantastic people who have shared, knowledge, words of encouragement and guidance.

These studies were funded by the Institutes of Health (NIH) grants HL-095056, HL-28481, and DK093757. A. Rodríguez was supported by the National Science Foundation Graduate Research Fellowship Program NSF grant number DGE-1144087 and A. Ko by NIH grants F31HL127921 and T32HG002536. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the article. Genotyping services for the METSIM cohort were supportedby 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.

VITA

EDUCATION

B.S., Biotechnology	2005, California State University, Northridge Northridge, California
Ph.D., Candidate	2018, University of California, Los Angeles Department of Human Genetics

RESEARCH PAPERS (PEER-REVIEWED)

1. **Alejandra Rodríguez**, Luis Gonzalez, Arthur Ko, Marcus Alvarez, Zong Miao, Yash Bhagat, Elina Nikkola, Ivette Cruz-Bautista, Olimpia Arellano-Campos, Linda L. Muñoz-Hernández, Maria-Luisa Ordóñez-Sánchez, Rosario Rodriguez-Guillen, Karen L. Mohlke, Markku Laakso, Teresa Tusie-Luna, Carlos A. Aguilar-Salinas, and Päivi Pajukanta. Molecular characterization of the lipid GWAS signal on chromosome 18q11.2 implicates HNF4A-mediated regulation of the TMEM241 gene. Arterioscler Thromb Vasc Biol. 2016;36:1350-1355.
2. Arthur Ko, Rita M. Cantor, Daphna Weissglas-Volkov, Elina Nikkola, Prasad M. V. Linga Reddy, Janet S. Sinsheimer, Bogdan Pasaniuc, Robert Brown, Marcus Alvarez, **Alejandra Rodriguez**, Rosario Rodriguez-Guillen, Ivette C. Bautista, Olimpia Arellano-Campos, Linda L. Muñoz-Hernández, Veikko Salomaa, Jaakko Kaprio, Antti Jula, Matti Jauhiainen, Markku Heliövaara, Olli Raitakari, Terho Lehtimäki, Johan G. Eriksson, Markus Perola, Kirk E. Lohmueller, Niina Matikainen, Marja-Riitta Taskinen, Maribel Rodriguez-Torres, Laura Riba, Teresa Tusie-Luna, Carlos A. Aguilar-Salinas & Päivi Pajukanta. Amerindian-specific regions under positive selection harbour new lipid variants in Latinos. Nature Communications. 2014;5:3983.
3. Elina Nikkola, Arthur Ko, Marcus Alvarez, Rita M. Cantor, Kristina Garske, Elliot Kim, Stephanie Gee, **Alejandra Rodriguez**, Reinhard Muxel, Niina Matikainen, Sanni Soderlund, Mahdi M. Motazacker, Jan Boren, Claudia Lamina, Florian Kronenberg, Wolfgang J. Schneider, Aarno Palotie, Markku Laakso, Marja-Riitta Taskinen, Paivi Pajukanta. Family-specific aggregation of lipid GWAS variants confers the susceptibility to familial hypercholesterolemia in a large Austrian family. Atherosclerosis. 2017;264:58-66.

ABSTRACTS

1. **Alejandra Rodríguez** and Päivi Pajukanta. Genome-wide profiling of context-specific RORA binding sites provides a catalog of transcriptional targets associated with lipid metabolism in the human liver. Department of Human Genetics Academic Retreat, Los Angeles, CA 2016.
2. **Alejandra Rodríguez**, Luis Gonzalez, Arthur Ko, Marcus Alvarez, Zong Miao, Yash Bhagat, Elina Nikkola, Ivette Cruz-Bautista, Olimpia Arellano-Campos, Linda L. Muñoz-Hernández, Maria-Luisa Ordóñez-Sánchez, Rosario Rodriguez-Guillen, Karen L. Mohlke, Markku Laakso, Teresa Tusie-Luna, Carlos A. Aguilar-Salinas, and Päivi Pajukanta. Identification of TMEM241 as the underlying gene in the chromosome 18q11.2 triglyceride region in Mexicans. Presented at the 64th Annual Meeting of the American Society of Human Genetics, San Diego, CA 2014.
3. **Alejandra Rodriguez** and Mayumi Prins. The age dependent effects of ketones on oxidative damage following controlled cortical impact. Presented at the 29th Annual National Neurotrauma Symposium, Hollywood Beach, FL 2011.

PROFESSIONAL EXPERIENCE

2010 – 2012	Research Associate II University of California- Los Angeles Los Angeles, California Department of Neurosurgery
2009 – 2009	Associate Scientist Iris International Diagnostics, Chatsworth, CA Department of Hematology

TEACHING EXPERIENCE

University of California-Los Angeles, Los Angeles, CA

Winter 2017	Teaching Associate, Biotechnology and Society 71B
Fall 2016	Teaching Associate, Biotechnology and Society 71A
Winter 2015	Teaching Assistant, Human Genetics and Genomics CM165
Winter 2014	Teaching Assistant, Undergraduate life science LS 23 L

AWARDS

2016	Cold Spring Harbor, NY Helmsley Scholarship
2015	Wellcome Trust Sanger Institute, London WTAC Award
2015	University of California, Los Angeles Department of Human Genetics and Biostatistics Diversity Fellowship Award, NIH Grant 1R25GM103774
2014	University of California, Los Angeles Department of Human Genetics Leena Peltonen Award Runner-up
2014 – 2017	National Science Foundation NSF Graduate Research Fellowship
2012 – 2013 2017 – 2018	University of California, Los Angeles Graduate Division Eugene V. Cota-Robles Fellowship

CHAPTER 1

INTRODUCTION

Short review of adipose tissue biology

In mammals, excess energy is stored as fat in adipose tissue. In humans, adipose tissue mainly consists of white adipose tissue (WAT) and the amount of brown adipose tissue (BAT) is small. Here we will focus on the metabolic functions of WAT, which primarily stores excess energy in the form of TGs to be used during periods of food deprivation[5]. WAT is distributed as the subcutaneous adipose tissue, located beneath the skin, and as intra-abdominal adipose depot, which in humans surrounds the gastrointestinal organs. Intra-abdominal fat accumulation is strongly associated with the development of obesity and related diseases, including type 2 diabetes, while the accumulation of subcutaneous fat exhibits weaker correlations[5]. White adipocytes acquire the expression of specific enzymes for TG synthesis (fatty acid synthase (FAS)) and lipolysis (hormone sensitive lipase (LIPE)) during their differentiation, which enable both the accumulation and mobilization of fat[5]. Given that they have a central role in the molecular mechanisms regulating lipid biosynthesis and metabolism, these enzymes have been targeted for drug development[6]. Most lipids stored in WAT come from circulating fatty acids and triglycerides from the liver and small intestine. The liver produces most of the lipids in de novo lipogenesis and TG synthesis. These lipids are insoluble, and thus they are efficiently packed into very low-density lipoproteins (VLDL) particles, secreted into the circulation, and delivered for storage in WAT as energy supply to be used in other peripheral tissues[5]. Fat from the small intestine is incorporated into large chylomicron particles. These also enter the circulation and are delivered to WAT for storage. Fatty acid uptake by adipocytes is believed to occur both by passive diffusion and active transport mediated by membrane enzymes including fatty acid transport protein (FATP), fatty acid binding protein plasma membrane (FABPpm), caveolin, and fatty acid translocase (CD36/FAT)[5] The mechanisms regulating intracellular levels of lipids are essential for maintaining homeostasis.

Lipid overflow is toxic to cells, and excessive fat intake in modern human populations has been hypothesized to induce decanalization[7] (i.e. the loss of a stable equilibrium in the underlying molecular pathways), ultimately leading to cardiometabolic disease[8]. In chapter 4, we target genes in the lipid metabolism pathways for an expression variance quantitative trait locus (QTL) analysis, which may signal decanalization of the molecular mechanisms in lipid metabolism[7–10].

Genetic architecture of hypertriglyceridemia and hypercholesterolemia

Elevated serum TG concentrations contribute to an increased risk of cardiovascular disease[25]. Hypertriglyceridemia is clinically defined as an elevated concentration of serum TGs (>150 mg/dL) and hypercholesterolemia as an elevated concentration of serum total cholesterol (TC) (>200 mg/dL), respectively[25]. Serum cholesterol and TG concentrations are heritable (56-77%), and based on current research, they are influenced by genetic variants with a broad spectrum of effect sizes and allele frequencies[11].

Genetic architecture: pharmaceutical application

The ability to predict a trait from the genetic sequence means that one can predict the risk of heart attack, stroke, early detection through improved screening, and drug response, which can ultimately help tailor personalized treatments. An individual's genetic susceptibility to disease is the sum of the effects of genetic risk variants and their interactions. Genetic architecture contains the genetic factors that together contribute to the broad-sense phenotypic heritability[12]. It defines the number and type of genes and alleles affecting the trait. To date, most of the genetic variants that contribute to disease susceptibility have been uncovered using GWAS. However, the total contribution of GWAS variants to phenotypic variation of complex traits only accounts for a small fraction of their total estimated heritability. The gap between the estimated heritability and the total genetic contribution from all GWAS variants is known as the "missing heritability." Epistatic

interactions may explain in part this missing heritability, which are well documented in model organisms[13–15]. Although studies in humans have reported epistatic interactions[9,10], they remain underinvestigated in quantitative genetics. Epistatic interactions could occur between multiple genetic factors and therefore, be population-specific, representing a challenge for their replication. In the near future, the anticipated lower cost of sequencing will lead to larger study samples and an increase in the number of the identified variants. In-depth characterization of the genetic architecture of lipid disorders will also help identify therapeutic targets. For example, to better target expensive and time consuming clinical trials, scientists can now use low and common frequency variants to predict drug responses using a drug-response curve[16–18] For serum TG levels, common variants with small effect sizes and rare variants with larger effects have been found in and near apolipoprotein C3 (APOC3), the gene encoding apolipoprotein CIII. This has allowed scientists to predict how pharmaceuticals targeting APOC3 will affect TGs[19]. A major promise of the characterization of the genetic architecture of lipid disorders is to improve the current low efficacy of drug targets, which are failing the expensive and time consuming clinical trials.

Waddington's original definition of epigenetics and decanalization

Recent conceptual and empirical developments[8,10,20] have expanded the definition of genetic architecture to include not only gene and allele number and the distribution of allelic and mutational effects, but also patterns of pleiotropy[21] and epistasis[10,20]. In 1942, C.H. Waddington coined the term "canalization[7]" which refers to the buffering of the genotype against variations in the environment and genetic composition. That is to say, organismal trails and physiological pathways, such as glucose and lipid metabolism, have reached a stable equilibrium and are robust against minor variations in environmental conditions. According to Waddington, the wild type of a trait is always less variable following a determined path; whereas the mutant phenotype is much

more variable. In 2009, Greg Gibson extended Waddington's observations, suggesting that decanalization events could explain the rising incidence of complex genetic diseases[8]. Decanalization of physiological processes, such as lipid metabolism, may be mediated by gene-environment interactions. Specifically, environmental metabolites can act as signaling molecules to chromatin, inducing chromatin remodeling and gene expression changes. Gibson proposed that under millions of years of stabilizing selection, biological systems evolved to a stable equilibrium and that under significant environmental pressures (dietary shifts, tobacco smoking, air pollution, altered pathogen exposure, and psychological stress), this equilibrium has been perturbed, which helps uncover cryptic genetic variation. According to Gibson, this hypothesis explains the increase in major common disease susceptibility in modern societies[8]. Decanalization will result in the loss of highly evolved reduced genetic variation, an increased "sensitivity" to environmental and genetic changes (which is mediated by complex gene-gene, gene-environment interactions), and increased variance for a particular phenotype.

Since Waddington's original theory of decanalization was published, others contributed to this theoretical framework[8–10,22]. Several studies have reported epistatic interactions in the human genome[9,10]. In model organisms, epistasis is common. In fact, it has been reported that epistasis dominates the genetic architecture of Drosophila[23]. Epistatic interactions can be investigated through gene expression variance-genotype association analysis (ve-QTLs) even though currently there is a lack of replicated epistatic observations in humans. Epistatic interactions may in part underlie the so-called "missing heritability" and together with undiscovered low-frequency variants and the known common GWAS variants may advance our understanding of the human genome-phenome map necessary for precision medicine, early detection, accurate prognosis, increased pharmaceutical efficacy, and other similar advances in molecular medicine.

DEFINITIONS

Canalization	The evolution of reduced genetic variability. Evolution of reduced gene effects through epistatic interactions with an evolving genetic background.
Epistasis	The phenomenon whereby one polymorphism's effect on a trait depends on other polymorphisms present in the genome.
Genetic architecture	The characteristics of genetic variation responsible for heritable phenotypic variability.
Complex traits	Traits that do not follow Mendelian inheritance patterns and are derived from any combination of multiple genetic factors, environmental factors, and their interactions.
GWAS	Studies that test the association of all measured genetic variation across the genome with a trait or disease.
Heritable	A characteristic or trait that has a portion of variability that is accounted for by genetic factors.
Pleiotropy	The phenomenon of one genetic locus influencing several traits.

REFERENCES

1. 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).
2. Rodríguez, A. *et al.* Molecular characterization of the lipid GWAS signal on chromosome 18q11.2 implicates HNF4A-mediated regulation of the TMEM241 gene. *Arterioscler. Thromb. Vasc. Biol.* **36,** 1350–1355 (2016).
3. Ko, A. *et al.* Amerindian-specific regions under positive selection harbour new lipid variants in Latinos. *Nat. Commun.* **5,** 3983 (2014).
4. Laakso, M. *et al.* METabolic Syndrome In Men (METSIM) Study: a resource for studies of metabolic and cardiovascular diseases. *J. Lipid Res.* jlr.O072629 (2017). doi:10.1194/jlr.O072629
5. Gesta, S. & Kahn, C. White Adipose Tissue. in *Adipose Tissue Biology: Second Edition* 149–199 (2017). doi:10.1007/978-3-319-52031-5_5
6. Shi, Y. & Burn, P. Lipid metabolic enzymes: emerging drug targets for the treatment of obesity. *Nat. Rev. Drug Discov.* **3,** 695–710 (2004).
7. Waddington, C. H. Canalization of Development and the Inheritance of Acquired Characters. *Nature* **150,** 563–565 (1942).
8. Gibson, G. Decanalization and the origin of complex disease. *Nat. Rev. Genet.* **10,** 134–140 (2009).
9. Burrows, E. L. & Hannan, A. J. Decanalization mediating gene-environment interactions in schizophrenia and other psychiatric disorders with neurodevelopmental etiology. *Front. Behav. Neurosci.* **7,** 157 (2013).

10. Wang, G. *et al.* Epistasis and decanalization shape gene expression variability in humans via distinct modes of action. *bioRxiv* 026393 (2015). doi:10.1101/026393
11. Fuchsberger, C. *et al.* The genetic architecture of type 2 diabetes. *Nature* **536,** 41–47 (2016).
12. Timpson, N. J., Greenwood, C. M. T., Soranzo, N., Lawson, D. J. & Richards, J. B. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nat. Rev. Genet.* **19,** 110–124 (2018).
13. Yang, Z. *et al.* Analysis of Epistasis among QTLs on Heading Date based on Single Segment Substitution Lines in Rice. *Sci. Rep.* **8,** 3059 (2018).
14. Dasmeh, P., Girard, É. & Serohijos, A. W. R. Highly expressed genes evolve under strong epistasis from a proteome-wide scan in E. coli. *Sci. Rep.* **7,** 15844 (2017).
15. Tong, A. H. Y. *et al.* Global mapping of the yeast genetic interaction network. *Science* **303,** 808–813 (2004).
16. Plenge, R. M., Scolnick, E. M. & Altshuler, D. Validating therapeutic targets through human genetics. *Nat. Rev. Drug Discov.* **12,** 581–594 (2013).
17. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47,** 856–860 (2015).
18. Ference, B. A., Majeed, F., Penumetcha, R., Flack, J. M. & Brook, R. D. Effect of naturally random allocation to lower low-density lipoprotein cholesterol on the risk of coronary heart disease mediated by polymorphisms in NPC1L1, HMGCR, or both: a 2 × 2 factorial Mendelian randomization study. *J. Am. Coll. Cardiol.* **65,** 1552–1561 (2015).
19. Gaudet, D. *et al.* Antisense Inhibition of Apolipoprotein C-III in Patients with Hypertriglyceridemia. *N. Engl. J. Med.* **373,** 438–447 (2015).

20. Hemani, G. *et al.* Detection and replication of epistasis influencing transcription in humans. *Nature* **508,** 249–253 (2014).
21. Hackinger, S. & Zeggini, E. Statistical methods to detect pleiotropy in human complex traits. *Open Biol.* **7,** (2017).
22. Tronick, E. & Hunter, R. G. Waddington, Dynamic Systems, and Epigenetics. *Front. Behav. Neurosci.* **10,** 107 (2016).
23. Huang, W. *et al.* Epistasis dominates the genetic architecture of Drosophila quantitative traits. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 15553–15559 (2012).
24. Plaisier, C. L. *et al.* Galanin Preproprotein Is Associated With Elevated Plasma Triglycerides. *Arterioscler. Thromb. Vasc. Biol.* **29,** 147–152 (2009).
25. Toth, P. P. Triglyceride-rich lipoproteins as a causal factor for cardiovascular disease. *Vasc. Health Risk Manag.* **12,** 171–183 (2016).

CHAPTER 2

MOLECULAR CHARACTERIZATION OF THE LIPID GENOME-WIDE ASSOCIATION STUDY SIGNAL ON CHROMOSOME 18Q11.2 IMPLICATES HNF4A-MEDIATED REGULATION OF THE *TMEM241* GENE

This work has been published and is freely accessible beginning May 19, 2016 in PubMed Central. PMCID: PMC5154300

Molecular Characterization of the Lipid Genome-Wide Association Study Signal on Chromosome 18q11.2 Implicates HNF4A-Mediated Regulation of the *TMEM241* Gene

Alejandra Rodríguez, Luis Gonzalez, Arthur Ko, Marcus Alvarez, Zong Miao, Yash Bhagat, Elina Nikkola, Ivette Cruz-Bautista, Olimpia Arellano-Campos, Linda L. Muñoz-Hernández, Maria-Luisa Ordóñez-Sánchez, Rosario Rodriguez-Guillen, Karen L. Mohlke, Markku Laakso, Teresa Tusie-Luna, Carlos A. Aguilar-Salinas, Päivi Pajukanta

Objective—We recently identified a locus on chromosome 18q11.2 for high serum triglycerides in Mexicans. We hypothesize that the lead genome-wide association study single-nucleotide polymorphism rs9949617, or its linkage disequilibrium proxies, regulates 1 of the 5 genes in the triglyceride-associated region.

Approach and Results—We performed a linkage disequilibrium analysis and found 9 additional variants in linkage disequilibrium (r^2>0.7) with the lead single-nucleotide polymorphism. To select the variants for functional analyses, we annotated the 10 variants using DNase I hypersensitive sites, transcription factor and chromatin states and identified rs17259126 as the lead candidate variant for functional in vitro validation. Using luciferase transcriptional reporter assay in liver HepG2 cells, we found that the G allele exhibits a significantly lower effect on transcription (P<0.05). The electrophoretic mobility shift and ChIPqPCR (chromatin immunoprecipitation coupled with quantitative polymerase chain reaction) assays confirmed that the minor G allele of rs17259126 disrupts an hepatocyte nuclear factor 4 α–binding site. To find the regional candidate gene, we performed a local expression quantitative trait locus analysis and found that rs17259126 and its linkage disequilibrium proxies alter expression of the regional transmembrane protein 241 (*TMEM241*) gene in 795 adipose RNAs from the Metabolic Syndrome In Men (METSIM) cohort (P=6.11×10^{-07}–5.80×10^{-04}). These results were replicated in expression profiles of *TMEM241* from the Multiple Tissue Human Expression Resource (MuTHER; n=856).

Conclusions—The Mexican genome-wide association study signal for high serum triglycerides on chromosome 18q11.2 harbors a regulatory single-nucleotide polymorphism, rs17259126, which disrupts normal hepatocyte nuclear factor 4 α binding and decreases the expression of the regional *TMEM241* gene. Our data suggest that decreased transcript levels of *TMEM241* contribute to increased triglyceride levels in Mexicans.

Key Words: chromatin ■ dyslipidemias ■ functional genomics ■ gene expression and regulation ■ genome-wide association study ■ mechanisms ■ triglycerides

Serum triglyceride levels are heritable and environmentally modifiable risk factor for cardiovascular disease.[1] Several groups have successfully used genome-wide association studies (GWAS) to identify signals for triglycerides and other lipid traits, including high-density lipoprotein cholesterol, low-density lipoprotein cholesterol, and total cholesterol.[2] However, the lead GWAS signals may not themselves be functional rather in linkage disequilibrium (LD) with the actual underlying susceptibility variant. This limitation in GWAS derives from the fact that the human genome is only relatively superficially screened in GWAS using common tag single-nucleotide polymorphisms (SNPs). Furthermore, the functional variant often acts through a regional gene. Therefore, GWASs are only a starting point and require subsequent fine mapping and functional validation studies to identify the actual susceptibility variants and genes.

According to a recent survey, both the US Hispanic men and women have higher levels of serum triglycerides than non-Hispanic whites or blacks,[3] a result consistently reported for the past 2 decades.[4] Recent studies using Latino cohorts have successfully narrowed European lipid loci.[5] Moreover, because of the higher incidence of metabolic

Received on: November 20, 2014; final version accepted on: May 9, 2016.

From the Department of Human Genetics, David Geffen School of Medicine (A.R., L.G., A.K., M.A., Z.M., Y.B., E.N., P.P.), Molecular Biology Institute (A.K., P.P.), and Bioinformatics Interdepartmental Program (P.P.), University of California, Los Angeles; Instituto Nacional de Ciencias Médicas y Nutrición, Salvador Zubiran, Mexico City, Mexico (I.C.-B., O.A.-C., L.L.M.-H., M.-L. O.-S., R.R.-G., T.T.-L., C.A.A.-S.); Department of Genetics, University of North Carolina, Chapel Hill (K.L.M.); Department of Medicine, University of Eastern Finland and Kuopio University Hospital (M.L.); and Instituto de Investigaciones Biomédicas de la UNAM, Mexico City, Mexico (T.T.-L.).

Nonstandard Abbreviations and Acronyms	
***cis*-eQTL**	*cis*-expression quantitative trait locus
ENCODE	encyclopedia of DNA elements
GWAS	genome-wide association study
HNF4A	hepatocyte nuclear factor 4 α
LD	linkage disequilibrium
SNP	single-nucleotide polymorphism
TFBS	transcription factor–binding sites
TMEM241	transmembrane protein 241

disease in the Amerindian origin populations, the investigation of their admixed genomes provides an opportunity to identify Amerindian-specific susceptibility variants for complex cardiovascular traits.[6] Despite their high predisposition to dyslipidemias, Hispanics remain underinvestigated as the discovery study stage in genomic cardiovascular studies. Previously, we identified a locus on chromosome 18q11.2 associated with high serum triglycerides in Mexicans using GWAS.[5] However, similar to other GWAS, the functional variants and the underlying gene(s) through which these variants exert their effects in the triglyceride phenotype remain to be elucidated. To find the actual functional risk variant(s), we systematically annotated the SNPs in the triglyceride-associated LD block with chromatin state marks and transcription factor–binding events which nominated rs17259126 as the top candidate functional variant. Its genomic landscape harbors regulatory sites and is predicted to disrupt an hepatocyte nuclear factor 4 α (HNF4A)–binding site. We show that the G allele of rs7259126 reduces expression of the luciferase reporter gene in a human liver cell line. Consistent with this result, the mobility shift and ChIPqPCR (chromatin immunoprecipitation coupled with quantitative polymerase chain reaction) assays confirmed that the same allele disrupts an HNF4A-binding site. Replicated *cis*-expression quantitative trait locus (*cis*-eQTL) analyses also implicate the minor G allele of rs17259126 for reduced expression of transmembrane protein 241 (*TMEM241*), suggesting *TMEM241* as the regional candidate gene. Taken together, we found that the triglyceride locus on chromosome 18q11.2 harbors at least one functional variant, rs17259126, associated with a decreased expression of the regional *TMEM241* gene, a novel gene for triglycerides in the rapidly growing Hispanic population with a high predisposition to dyslipidemias.

Materials and Methods

Materials and Methods are available in the online-only Data Supplement.

Results

Pairwise LD Analysis to Identify LD Proxies

In our original GWAS,[5] conditional association analyses at the top 12 genotyped loci did not reveal additional independent SNPs with $P \leq 2.5 \times 10^{-3}$. To identify the full set of variants in LD with the lead GWAS SNP rs9949617, we first performed a regional LD analysis in the triglyceride-associated LD block. The LD block was determined in our previous study as the region spanning SNPs in LD of $r^2 \geq 0.5$ with the lead SNP rs9949617.[5] For the LD analysis, we used our genotyped and imputed GWAS data,[5] and we also verified using the 1000 Genomes Project data that no additional SNP(s) inside or outside this LD block (±500 kb from the block borders) have emerged to be in LD with the lead SNP rs9949617 since our previous study.[5] We found 3 genotyped and 6 imputed SNPs in LD ($r^2 \geq 0.7$) with the lead SNP rs9949617 (Table). Two of these 10 SNPs (rs9949617 and rs4800467) were genotyped in stages 1 and 2 of our original GWAS scan,[5] both resulting in P values $<5 \times 10^{-8}$. Because any of these 10 SNPs in LD can be the functional variant underlying the triglyceride association on chromosome 18q11.2, we first performed functional annotation followed by hypothesis-driven functional assays to uncover the functional variant in the triglyceride-associated LD block. We also tested the candidate variant and its LD proxies for regional effects on gene expression among the 5 genes in the triglyceride-associated LD block using a *cis*-eQTL analysis to investigate whether the variant changes expression of a particular regional gene.

Functional Genomics Analysis Using Encyclopedia of DNA Elements Data

Cis-eQTL variants often reside in regulatory elements such as transcription factor–binding sites (TFBS) and interrupt transcription factor occupancy, leading to transcriptional changes. However, functional variants may also act through multiple other mechanisms making functional validation studies challenging. To facilitate the identification of suitable functional assays, we used the encyclopedia of DNA elements (ENCODE) data sets to give biological interpretation to the variants, and based on their predicted functionality, we conducted hypothesis-driven functional assays. TFBS often coincide with regions of open chromatin; hence, we annotated the chromatin state using ENCODE DNase I hypersensitive sites and histone marks in disease-relevant cell lines and control cell lines. In addition to the ENCODE biochemical annotations, we looked for transcription factor motif disruptions using HaploReg. We hypothesized that variants with the greatest amount of regulatory evidence from experimental data sets and bioinformatic predictions are more likely to be functional. Using this approach, we screened all 10 SNPs (the lead SNP and its 9 LD proxies) and selected rs17259126 as a top candidate for functional validation because it resides in a TFBS and a likely regulatory element defined by the co-occurrence of H3K27ac and H3K4me1. The G allele of rs17259126 is also predicted to disrupt a HNF4A regulatory motif (Figure I in the online-only Data Supplement). HNF4A is a known regulator of several metabolic genes.[7] On the basis of these annotations, we hypothesized that rs17259126 resides in a TFBS and regulates expression of one of the regional genes on chromosome 18q11.

Functional Validation of Candidate Variants

We sought to validate our predicted functional variant rs17259126. We performed luciferase reporter assays using engineered vectors containing a 600-bp sequence around the SNP. At 48 hours post transfection of HepG2 cells, we found that the minor allele G displays a 1.5-fold decreased reporter expression ($P<0.05$) compared with the major A allele in 3 biological replicates (Figure 1). These results are consistent with

Table 1. The Lead SNP rs9949617 and its LD Proxies Are *cis*-eQTLs for the *TMEM241* Gene

SNP	Minor Allele	r^2	MAF (AMR)	MAF (EUR)	MAF (FIN)	*cis*-eQTL *P* value* (β)
rs9949617†	T	1.00	0.34	0.17	0.14	5.96×10^{-06} (–0.129)
rs9962573†	T	0.97	0.32	0.17	0.14	5.96×10^{-06} (–0.129)
rs4800467†	G	0.92	0.35	0.2	0.18	6.11×10^{-07} (–0.131)
rs1276322†	G	0.82	0.33	0.17	0.14	1.7×10^{-04} (–0.109)
rs17259126	G	0.77	0.25	0.06	0.08	1.1×10^{-04} (–0.149)
rs9954334	G	1.00	0.34	0.17	0.14	5.96×10^{-06} (–0.128)
rs67124903	G	0.97	0.34	0.17	0.14	5.96×10^{-06} (–0.128)
rs71360517	A	0.88	0.29	0.14	0.10	5.80×10^{-04} (–0.112)
rs77127070	A	0.74	0.22	0.03	0.04	N/A
rs4800154	A	0.74	0.23	0.03	0.04	N/A

The regional LD analysis uncovered 3 genotyped and 6 imputed variants in LD (r^2≥0.7) with the lead SNP, rs9949617 in Mexicans. *cis*-eQTL indicates *cis*-expression quantitative trait locus; LD, linkage disequilibrium; MAF, minor allele frequency in the 1000 Genomes Project on the admixed American (AMR) individuals, European ancestry (EUR) individuals, and Finns (FIN); NA, not available; SNP, single-nucleotide polymorphism; and *TMEM241*, transmembrane protein 241.

*The cis-eQTL *P* values obtained in the analysis of the Finnish Metabolic Syndrome In Men (METSIM) RNA-seq data (n=795) pass the Bonferroni correction for 50 tests (10 SNPs and 5 regional genes in the triglyceride-associated LD block; $P<0.001$). The β is shown for the minor allele.

†Genotyped SNPs.

the observed direction of the *cis*-eQTL effect (β=–0.149; Table). Similar assays for the lead SNP rs9949617 and rs4800467 did not reveal significant expression changes in the luciferase assay.

To further investigate whether the variant disrupts an HNF4A motif, we performed electrophoretic mobility shift assays (EMSA) using isolated HNF4A protein (Figure 2A) or HepG2 cell nuclear extracts (Figure 2B) and found evidence that HNF4A preferentially binds the major A allele of rs17259126 in 4 biological replicates. We also performed EMSA assays for the 9 other LD proxies variants. No allele-specific shifts were observed (Figure II in the online-only Data Supplement). Together, the luciferase (Figure 1) and EMSA (Figure 2) assays suggest that HNF4A may regulate expression of a target gene by directly binding to the rs17259126 regulatory site.

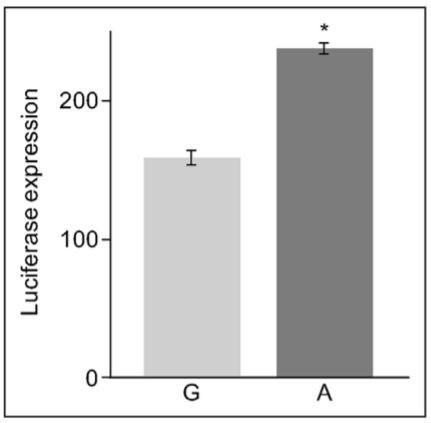


Figure 1. Luciferase expression of rs17259126 alleles in HepG2 cells. The reference allele A exhibits a 1.5-fold increased reporter expression ($P=2.97\times10^{-4}$) when compared with the minor allele G. Luciferase assays were read at 48 hours post transfection. Statistical analysis was performed using the *t* test function in R. The Y axis represents the percent change in reporter expression relative to the empty pGL4.23[luc2/minP] vector. Error bars indicate the SE for the means of 3 independent biological replicates each done in triplicate.

To confirm that HNF4A interacts with the variant site in HepG2 cells, we performed chromatin immunoprecipitation followed by qPCR targeting a 71-bp (site 1) or 151-bp (site 2) sequence surrounding rs17259126 (Figure 3). We found an average enrichment of 4.23 and 2.29 for the sequences, respectively, when compared with an unbound control site. Our functional studies provide converging evidence that the sequence underlying rs17259126 is an HNF4A-binding site and that the G minor allele significantly inhibits this interaction in vitro.

cis-eQTL Analysis

GWAS variants residing in regulatory elements such as TFBS can lead to gene expression changes and contribute to disease susceptibility. We investigated whether the lead GWAS SNP may affect expression of the regional genes in the ≈300-kb region defining the triglyceride-associated window on chromosome 18q11.2 (LD $r^2>0.5$ with the lead SNP).[5] We performed a *cis*-eQTL analysis for the 5 genes within this triglyceride-associated LD block using adipose RNA-seq samples (n=795) from the Metabolic Syndrome In Men (METSIM) cohort and discovered that the lead SNP rs9949617 (ie, the SNP with the strongest triglyceride association signal[4]) and its LD proxies are a *cis*-eQTL, regulating the expression of one regional gene, the *TMEM241* ($P=6.11\times10^{-07}$–5.80×10^{-04}; Table). These results pass the Bonferroni correction for the 50 performed tests (10 SNPs tested for 5 regional genes; $P<0.001$; Table), and the 10 triglyceride-associated SNPs did not regulate expression of any of the 4 other genes within the LD block (Bonferroni corrected $P>0.05$).

To validate and replicate these regional *cis*-eQTL analysis results, we used expression data from 856 publicly available human adipose, skin, and lymphocyte RNA

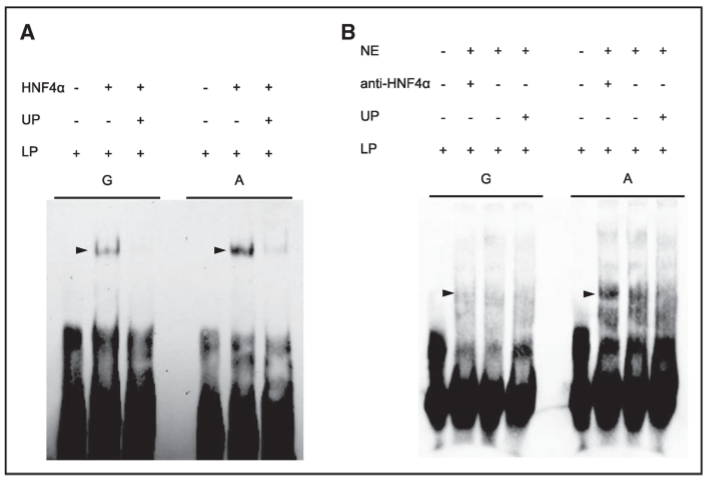


Figure 2. As predicted by motif analysis (Figure I in the online-only Data Supplement), EMSA (electrophoretic mobility shift assays) revealed that hepatocyte nuclear factor 4 α (HNF4A) recombinant protein directly interacts and displays a higher affinity for the reference A allele of the transmembrane protein 241 expression quantitative trait locus SNP rs17259126 (**A**), although no super shift was obtained when using HepG2 nuclear extract and anti-HNF4A (**B**). **A**, HNF4A isolate. **B**, HepG2 hepatoma nuclear extract. HNF4A (HNF4A isolated protein), unlabeled probe (UP), labeled probe (LP), nuclear extract (NE), and anti-HNF4A (anti-HNF4A). SNP indicates single-nucleotide polymorphism.

microarray samples from the Multiple Tissue Human Expression Resource (MuTHER)[8] and similarly discovered that the lead SNP rs9949617 is a *cis*-eQTL (Figure III in the online-only Data Supplement), regulating the expression of *TMEM241* ($P<1\times10^{-5}$ across all 3 tissues; β=−0.107 for adipose). These replication data are consistent, including the direction of the effect, with our *cis*-eQTL signal in Finns and our luciferase assays in which the minor G allele results in a decreased expression (Table and Figure 1). We also found comparable *cis*-eQTL results for the lead SNP rs9949617 in the HapMap3 data sets for the CEU (Utah residents with Northern and Western European ancestry from the CEPH collection; $P=0.0010$), CHB (Han Chinese in Beijing, China; $P=0.0019$), and JPT (Japanese in Tokyo, Japan; $P=6.0\times10^{-4}$) samples in lymphoblastoid cells. Although there was a trend toward significance, this relationship did not hold for the MEX HapMap sample ($P=0.20$), perhaps because of the low number of Mexican-American samples (n=45) included in the HapMap project. These results implicate *TMEM241* as a likely regional gene underlying the GWAS association because the lead SNP and its LD proxies robustly regulate *TMEM241* expression through multiple cohorts. Taken together, these data suggest that rs17259126 is at least one of the functional SNPs underlying the original triglyceride GWAS signal[5] on chromosome 18q11.2 in Amerindian origin populations.

Discussion

We recently identified a locus on chromosome 18q11.2 associated with high serum triglycerides in Mexicans using GWAS.[5] However, GWAS typically do not conclusively identify a functional regulatory variant and candidate gene, rather they require statistical and biochemical follow-up studies.[9,10] We used statistical fine mapping to first identify variants in the triglyceride-associated LD block. Because all variants represent 3′UTR (untranslated region) or noncoding variants, we annotated their biological function using available regulatory data sets and bioinformatic tools and subsequently validated our recorded annotations using appropriate molecular assays.

Our LD analyses uncovered 9 variants in LD with the lead GWAS SNP rs9949617. Functional annotations using HaploReg[11] found that rs17259126 is predicted to disrupt an HNF4A-binding site, the minor G allele exhibiting a lower

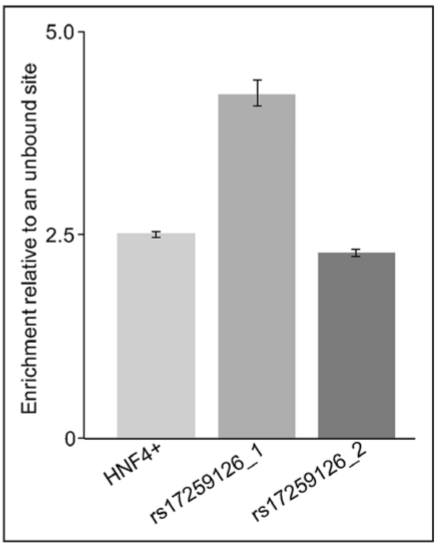


Figure 3. ChIP DNA analyzed by real-time polymerase chain reaction showed that hepatocyte nuclear factor 4 α (HNF4A) binds the transmembrane protein 241 expression quantitative trait locus SNP rs17259126 site in hepatoma cells. The y axis represents the enrichment of HNF4A at the rs17259126 site relative to an unbound control site. Error bars represent the SE for the means of 3 experimental replicates each done in triplicate. The region surrounding rs17259126 was targeted using 2 different pairs of oligos, indicated as rs17259126_1 and rs17259126_2. The sequences of the oligos are given in Table I in the online-only Data Supplement. SNP indicates single-nucleotide polymorphism.

enrichment score. Furthermore, the ENCODE TF ChIP-seq data in HepG2 showed evidence of HNF4A enrichment around rs17259126. These findings prompted us to nominate rs17259126 as the lead candidate for molecular validation. We performed HNF4A ChIPqPCR targeting the SNP region and confirmed that HNF4A indeed binds the SNP site. HNF4A is a well-known, central regulator of hepatocyte development, differentiation, and gene expression[7,12] associated with type 2 diabetes mellitus, consistent with the triglyceride association. In line with our bioinformatics prediction, we also show that the G allele of rs17259126 reduces transcription of the luciferase reporter and significantly inhibits HNF4A binding in mobility shift assays. It is worth noting that Amerindian origin populations have >3-fold higher frequency of the minor allele G of rs17259126 when compared with Europeans (minor allele frequencies for admixed American=0.22, European=0.06, African=0.08, and Asian=0.20, respectively).

To identify the regional gene, we performed cis-eQTL analyses using expression data from multiple cohorts, tissues, and platforms. We provide replicated evidence that the minor G allele of rs17259126 and its LD proxies are a robust *cis*-eQTL decreasing expression of the regional *TMEM241* gene across many cohorts. Our results suggest that HNF4A binds the A allele of rs17259126 site and increases expression of the *TMEM241* gene, 1 of the 5 regional genes in the LD block. We hypothesize that individuals with the G allele have decreased *TMEM241* expression which affects the normal triglyceride synthesis or secretory pathways through an unknown mechanism.

The *TMEM241* gene is a yeast VRG4 homolog, a Golgi-localized GDP (guanosine diphosphate mannose)-mannose transporter. Yeast VRG4 is pleiotropically required for a range of Golgi functions, including N-linked glycosylation, secretion, protein sorting, and the maintenance of a normal endomembrane system.[13,14] In the mammalian Golgi, carbohydrate processing is a highly diverse process. Carbohydrate chains may contain galactose, sialic acid, fucose, xylose, *N*-acetylglucosamine, and *N*-acetylgalactosamine unlike in the yeast *Saccharomyces cerevisiae*, where glycosylation is restricted to mannosylation. Thus, human TMEM241 may function in the transport of other nucleotide sugars required in mammalian systems. In addition to glycoproteins, sphingolipids are also modified in the Golgi and have been implicated in metabolic disease.[15] TMEM241 is believed to function as a nucleotide sugar transporter and, when defective, may lead to underglycosylation of glycoproteins and sphingolipids, potentially resulting in dysregulation of triglyceride synthesis.

Together, our results provide converging evidence suggesting rs17259126 as one of the functional variants underlying the GWAS association signal on 18q11.2,[5] and *TMEM241* as the underlying gene for triglycerides in Amerindian origin populations. However, because not all individuals of Mexican ancestry share the same composition of Amerindian DNA, additional cohorts may or may not replicate this particular association.

Future studies focusing on characterizing the role of *TMEM241* in triglyceride metabolism could include CRISPR/Cas9,[16] an emerging technology for targeted genomic modification. This technology allows a site-specific genetic engineering in disease-relevant cell lines to interrogate the function of specific genes and single-nucleotide variants in their native chromatin state. Elucidation of the role of *TMEM241* in triglyceride metabolism may help guide future research and development of new therapies for effective triglyceride management and prevention of heart disease in the rapidly growing Hispanic populations, currently underinvestigated in genomic cardiovascular studies despite their high predisposition to dyslipidemias.

Acknowledgments

We thank the Mexican and Finnish individuals who participated in this study. We also thank Saúl Cano-Colín for laboratory technical assistance. Michael Boehnke and Francis Collins are thanked for providing the METSIM genotype data.

Sources of Funding

This study was funded by the Institutes of Health (NIH) grants HL-095056, HL-28481, and DK093757. A. Rodríguez was supported by the National Science Foundation Graduate Research Fellowship Program NSF grant number DGE-1144087 and A. Ko by NIH grants F31HL127921 and T32HG002536. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the article. Genotyping services 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.

Disclosures

None.

References

1. Willer CJ, Schmidt EM, Sengupta S, et al. Discovery and refinement of loci associated with lipid levels. *Nat Genet*. 2013;45:1274–1283. doi: 10.1038/ng.2797.
2. Go AS, Mozaffarian D, Roger VL, et al; American Heart Association Statistics Committee and Stroke Statistics Subcommittee. Executive summary: heart disease and stroke statistics–2014 update: a report from the American Heart Association. *Circulation*. 2014;129:399–410. doi: 10.1161/01.cir.0000442015.53336.12.
3. Aguilar-Salinas CA, Canizales-Quinteros S, Rojas-Martínez R, Mehta R, Villarreal-Molina MT, Arellano-Campos O, Riba L, Gómez-Pérez FJ, Tusié-Luna MT. Hypoalphalipoproteinemia in populations of Native American ancestry: an opportunity to assess the interaction of genes and the environment. *Curr Opin Lipidol*. 2009;20:92–97. doi: 10.1097/MOL.0b013e3283295e96.
4. Aguilar-Salinas CA, Tusie-Luna T, Pajukanta P. Genetic and environmental determinants of the susceptibility of Amerindian derived populations for having hypertriglyceridemia. *Metabolism*. 2014;63:887–894. doi: 10.1016/j.metabol.2014.03.012.
5. Weissglas-Volkov D, Aguilar-Salinas CA, Nikkola E, et al. Genomic study in Mexicans identifies a new locus for triglycerides and refines European lipid loci. *J Med Genet*. 2013;50:298–308. doi: 10.1136/jmedgenet-2012-101461.
6. Ko A, Cantor RM, Weissglas-Volkov D, et al. Amerindian-specific regions under positive selection harbour new lipid variants in Latinos. *Nat Commun*. 2014;5:3983. doi: 10.1038/ncomms4983.
7. Hayhurst GP, Lee YH, Lambert G, Ward JM, Gonzalez FJ. Hepatocyte nuclear factor 4alpha (nuclear receptor 2A1) is essential for maintenance of hepatic gene expression and lipid homeostasis. *Mol Cell Biol*. 2001;21:1393–1403. doi: 10.1128/MCB.21.4.1393-1403.2001.
8. Grundberg E, Small KS, Hedman ÅK, et al; Multiple Tissue Human Expression Resource (MuTHER) Consortium. Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat Genet*. 2012;44:1084–1089. doi: 10.1038/ng.2394.
9. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. *Genome Res*. 2012;22:1748–1759. doi: 10.1101/gr.136127.111.

10. Lo KS, Vadlamudi S, Fogarty MP, Mohlke KL, Lettre G. Strategies to fine-map genetic associations with lipid levels by combining epigenomic annotations and liver-specific transcription profiles. *Genomics*. 2014;104:105–112. doi: 10.1016/j.ygeno.2014.04.006.
11. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res*. 2012;40(Database issue):D930–D934. doi: 10.1093/nar/gkr917.
12. Parviz F, Matullo C, Garrison WD, Savatski L, Adamson JW, Ning G, Kaestner KH, Rossi JM, Zaret KS, Duncan SA. Hepatocyte nuclear factor 4alpha controls the development of a hepatic epithelium and liver morphogenesis. *Nat Genet*. 2003;34:292–296. doi: 10.1038/ng1175.
13. Hansen HG, Schmidt JD, Søltoft CL, Ramming T, Geertz-Hansen HM, Christensen B, Sørensen ES, Juncker AS, Appenzeller-Herzog C, Ellgaard L. Hyperactivity of the Ero1α oxidase elicits endoplasmic reticulum stress but no broad antioxidant response. *J Biol Chem*. 2012;287:39513–39523. doi: 10.1074/jbc.M112.405050.
14. Dean N, Zhang YB, Poster JB. The VRG4 gene is required for GDP-mannose transport into the lumen of the Golgi in the yeast, Saccharomyces cerevisiae. *J Biol Chem*. 1997;272:31908–31914. doi: 10.1074/jbc.272.50.31908.
15. Brice SE, Cowart LA. Sphingolipid metabolism and analysis in metabolic disease. *Adv Exp Med Biol*. 2011;721:1–17. doi: 10.1007/978-1-4614-0650-1_1.
16. Hsu PD, Lander ES, Zhang F. Development and applications of CRISPR-Cas9 for genome engineering. *Cell*. 2014;157:1262–1278. doi: 10.1016/j.cell.2014.05.010.

Highlights

- The triglyceride locus on chromosome 18q11.2 harbors at least one functional variant, rs17259126, associated with a decreased expression of the regional *TMEM241* gene, a novel gene for triglycerides in the Hispanic population.
- HNF4A may regulate the expression of the *TMEM241* gene by directly binding the rs17259126 regulatory site.
- Our findings suggest that decreased transcript levels of *TMEM241* contribute to increased triglyceride levels in Mexicans.

Arteriosclerosis, Thrombosis, and Vascular Biology

JOURNAL OF THE AMERICAN HEART ASSOCIATION

Molecular Characterization of the Lipid Genome-Wide Association Study Signal on Chromosome 18q11.2 Implicates HNF4A-Mediated Regulation of the *TMEM241* Gene

Alejandra Rodríguez, Luis Gonzalez, Arthur Ko, Marcus Alvarez, Zong Miao, Yash Bhagat, Elina Nikkola, Ivette Cruz-Bautista, Olimpia Arellano-Campos, Linda L. Muñoz-Hernández, Maria-Luisa Ordóñez-Sánchez, Rosario Rodriguez-Guillen, Karen L. Mohlke, Markku Laakso, Teresa Tusie-Luna, Carlos A. Aguilar-Salinas and Päivi Pajukanta

Detailed Materials and Methods
Materials and Methods are available in the online-only Data Supplement

Study cohorts
All Mexican participants were recruited at the Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán (INCMNSZ), Mexico City, as described in detail previously[1]. The study design was approved by the local ethics committee and all participants gave an informed consent. Measurements of fasting lipid levels were performed with commercially available standardized methods[1].

A total of 795 participants from the Finnish METabolic Syndrome in Men (METSIM, total n=10,197)[2] were included in the regional *cis*-eQTL analysis of this study. The METSIM participants who underwent a subcutaneous abdominal adipose biopsy for RNA-seq were recruited at the University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland. All 10,197 METSIM participants are male with a median age at examination 57 years (range: 45-74 years)[2]. The study design was approved by the local ethics committee and all participants gave an informed consent.

Linkage disequilibrium analysis
To search for variants in linkage disequilibrium (LD) ($r^2 \geq 0.7$) with the lead SNP rs9949617, we performed an LD analysis using PLINK[3] in the ~300-kb region on chr18q11.2 associated with TGs[1]. For the LD analysis, we used our genotyped and imputed GWAS data[1], and we also verified utilizing the 1000 Genomes Project data that no additional SNP(s) inside or outside this LD block (+/-500 kb from the block borders) has emerged to be in LD with the lead SNP rs9949617 since our previous study[1]. To maintain the strength of the association evidence while searching for a maximum number of variants in relatively tight LD with the lead SNP, we arbitrarily set the LD cut-point to $r^2=0.7$.

Genotyping rs17259126 in 200 Mexican GWAS controls
We genotyped the variant rs17259126 in 200 Mexican GWAS controls to validate the imputation results. This analysis confirmed that rs17259126 is in LD ($r^2=0.77$) with the lead GWAS SNP rs9949617. The genotype concordance between the genotyped and imputed genotypes of rs17259126 was 98.8%.

Prioritizing functional variation through epigenomic annotation
To prioritize the variants for experimental analysis, we filtered the variants through systematic data mining, including enrichment of signatures for regulatory elements at the variation site in disease relevant cell lines, including adipose and liver when available and utilizing ENCODE and Roadmap Epigenomics Project data[4-5]. The functional potential was evaluated based on whether the SNP resided in a region containing DNase I hypersensitive site, transcription factor binding site and by evaluating its chromatin state using and histone mark ChIP-seq data.

Plasmid DNA constructs and luciferase assays

To experimentally validate the lead candidate variants, rs9949617, rs4800467 and rs17259126, for enhancer activity, a sequence of 500-bp surrounding the SNPs was amplified from genomic DNA and cloned upstream of a minimal promoter in the pGL4.23[luc2/minP] vector (Promega). To obtain constructs with the alternative allele, we utilized the GeneArt® Site-Directed Mutagenesis PLUS System (Invitrogen A14604) following the manufacturer's protocol. All plasmid constructs were verified by sequencing using the RVprimer3 (Promega) and other appropriate custom designed sequencing primers.

Reporter constructs were transiently transfected using Lipofectamine 2000 (invitrogen 11668-027) into human liver hepatocellular carcinoma cells (HepG2) (ATCC HB-8065). Luciferase assays were read at 48 hrs post transfection in triplicates and the results were reproduced in three independent biological replicates each with three technical replicates. To normalize the activity of the pGL4.23[luc2/minP] engineered vector, we used the pGL4.74[hRluc/TK] internal control, which minimizes the experimental variability caused by differences in cell viability or transfection efficiency. All experiments were performed following the manufacturer's recommendations with minor modifications. Statistical analysis was performed using the R t test function. Expression values were normalized and expressed as a percent change from the un-engineered control pGL4.23[luc2/minP] vector.

Mobility shift assays

To detect DNA-protein interactions, we performed electrophoretic mobility shift assays (EMSA) (Active Motif, 37341) for the 10 variants in the LD TG block. HepG2 cell extracts (Active motif, 36011) or purified HNF4A factor (OriGene, TP316588) were incubated with a biotin labeled DNA probe (IDT) containing the SNP of interest. Samples were then resolved by electrophoresis on a 6% DNA retardation gels (ThermoFisher, EC6365BOX) and transferred to a nylon membrane. The biotin end-labeled DNA probe was detected using streptavidin conjugated to horseradish peroxidase (HRP) and a chemiluminescent substrate (Active Motif, 37341).

ChIPqPCR

HepG2 cells were grown to confluence and cross-linked using 1% formaldehyde. Chromatin was sheared by sonication to an average size of 500 bp and incubated with anti-HNF4A (PPMX, PP-H1415-00) antibodies overnight at 4C. Immunoprecipitated complexes were captured using magnetic protein A beads (ThermoFisher, 10001D). DNA was eluted from the beads and incubated over night at 65^{o}C to reverse the cross-links. After purification (Active Motif, 58002), DNA was analyzed by qPCR with primers corresponding to the regions around the rs17259126 site. Sites corresponding to peaks for transcriptional repression mark, H3K27me3 or HNF4A peaks in HepG2 ChIP-seq from ENCODE/Broad promoter sites were included as negative and positive controls, respectively (suppl table I). For each primer pair, the average cycle threshold values of the triplicates (variability <0.2) were calculated. The relative DNA amount was calculated using the $\Delta\Delta$CT method for relative quantification.

METSIM RNA-seq

We isolated total RNA from abdominal subcutaneous adipose needle biopsy using Qiagen miRNeasy kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. Polyadenylated mRNA was prepared using Illumina TruSeq RNA Sample Preparation Kit v2. Paired end, 50-bp reads were generated using the Illumina HiSeq 2000 platform. We used STAR[6] to align the reads to the hg19 reference genome, allowing up to 2 mismatches per read-pair. To quantify gene expression, we first cleaned the Gencode v19 annotation data using the following steps mimicking closely the GTEx RNA-seq analysis pipeline[7]: 1) use only genes annotated as "protein_coding" or "lincRNA"; 2) remove transcripts annotated as "retained_intron"; 3) collapse overlapping exons from all transcripts of a gene; and 4) remove exonic intervals belonging to more than 1 gene. This generated non-redundant exons from all transcripts for each gene. For transcript assembly, we employed Cufflinks[8] v2.2.1 using the -G option and --overhang-tolerance set to 0 in order to use only reads that fully overlap exons. To normalize gene expression, we analyzed genes with expression FPKM > 0 in more than 90% of individuals. FPKM values were rank transformed to a standard normal distribution for each gene. We corrected for technical and environmental confounders using PEER[9] with 40 factors. As an additional quality control step, we tested the RNA-seq data for genotype concordance between the DNA and RNA samples using the VerifyBamID[10], MixupMapper[11], and GATK best practice guidelines to call variants from RNA-seq data, and matched the data accordingly.

METSIM genotypes and Imputation

METSIM genotypes were generated from the Illumina OmniExpress and Illumina ExomeChip arrays. All imputation analyses in the study were performed first pre-phasing SNPs with SHAPEIT[12] v2.r727 and then performing imputations using IMPUTE2[13-15] v2.3.0 with default parameters. The 1000 Genomes phase 1 version 3 (March 2012 release) was utilized for both pre-phasing and imputation. We removed variants with genotype posterior probabilities less than 0.9, minor allele frequency less than 0.01, and Hardy-Weinberg Equilibrium p-value less than $1x10^{-5}$.

METSIM *cis*-eQTL analysis

We used linear regression employing an additive model implemented in Matrix-eQTL[16] to map *cis*-eQTLs from genotype/imputation data and normalized gene expression in the chr18q11.2 region. The ten TG-associated SNPs in LD ($r^2 \geq 0.7$) were tested for *cis*-eQTL effects for the 5 genes within the TG-associated ~300-kb LD block. P-values passing the Bonferroni correction for 50 tests (10 SNPs tested for 5 genes; $p<0.001$) were considered as statistically significant.

MuTHER *cis*-eQTL analysis

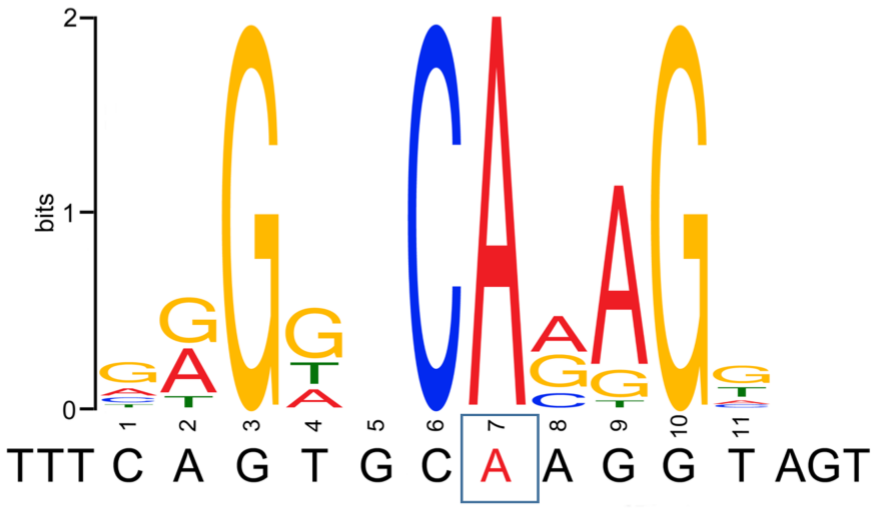
To replicate the *cis*-eQTL found in METSIM, we performed a *cis*-eQTL analysis utilizing the Multiple Tissue Human Expression Resource (MuTHER) study, a publicly available microarray data set of 856 samples from the TwinsUK registry[17]. We analyzed differential gene expression in the 18q11.2 GWAS locus across three different tissues: human adipose, skin, and transformed lymphocytes.

Supplemental References

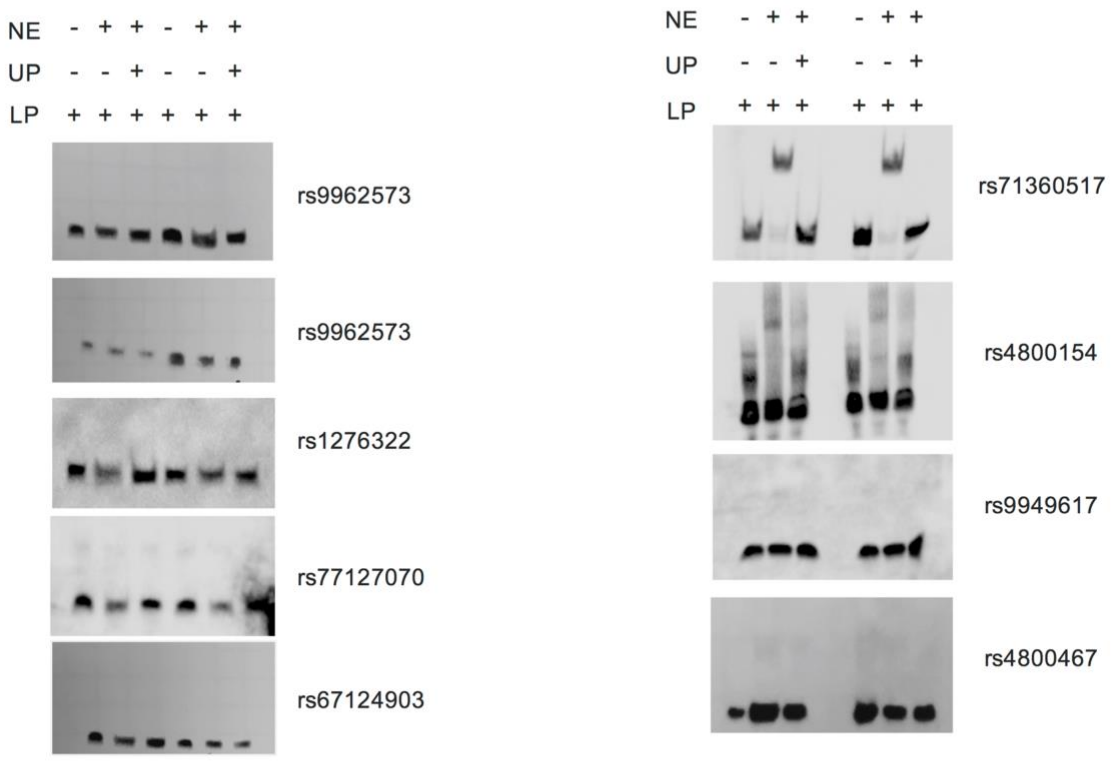
1. Weissglas-Volkov D, Aguilar-Salinas CA, Deere K, et al. Genomic study in Mexicans identifies a new locus for triglycerides and refines European lipid loci. J Med Genet. 2013;50:298-308.
2. Stancáková A, Civelek M, Saleem NK, et al. Hyperglycemia and a common variant of GCKR are associated with the levels of eight amino acids in 9,369 Finnish men. Diabetes. 2012;61:1895-1902.
3. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 2007;81:559-575.
4. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.
5. Bernstein BE, Stamatoyannopouos JA, Costello JF, et al. The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol*. 2010;28:1045-1048.
6. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15-21.
7. Melé M, Ferreira PG, Reverter F, et al. Human genomics. The human transcriptome across tissues and individuals. Science. 2015;348:660-665.
8. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28:511-115.
9. Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat Protoc. 2012;7:500-507.
10. Jun G1, Flickinger M, Hetrick KN, Romm JM, Doheny KF, Abecasis GR, Boehnke M, Kang HM. Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data. Am J Hum Genet. 2012;91:839-848.
11. Westra HJ1, Jansen RC, Fehrmann RS, te Meerman GJ, van Heel D, Wijmenga C, Franke L. MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. Bioinformatics. 2011;27:2104-2111.
12. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. Nat Methods. 2011;9:179-181.
13. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009;5:e1000529.
14. Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. G3. 2011;1:457-470.

15. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet. 2012;44:955-959.
16. Shabalin AA. Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. Bioinformatics. 2012;28:1353-1358.
17. Grundberg E, Small KS, Hedman ÅK, et al. Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat Genet. 2012;44:1084-1089.

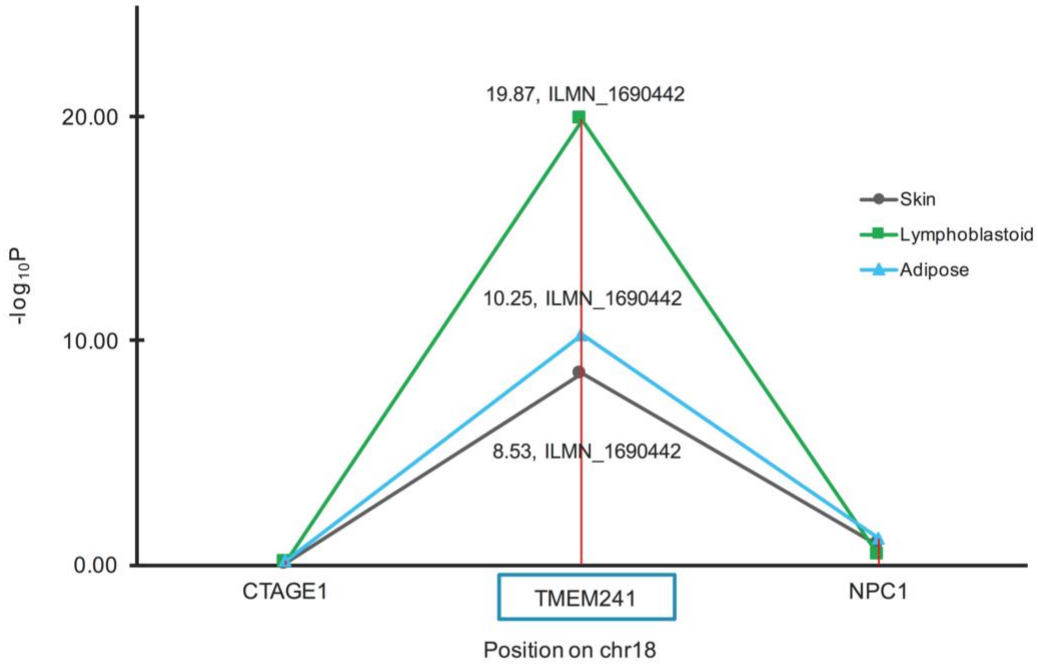
SUPPLEMENTAL MATERIAL



Supplemental Figure I. Motif discovery analyses using publicly available HNF4A ChIP-seq data sets (Kheradpour and Kellis 2013) show that the sequence underlying rs17259126 is an HNF4A binding motif (HNF4_disc3) and that the A major allele has a higher enrichment score (14 vs. 2) when compared to the G allele.



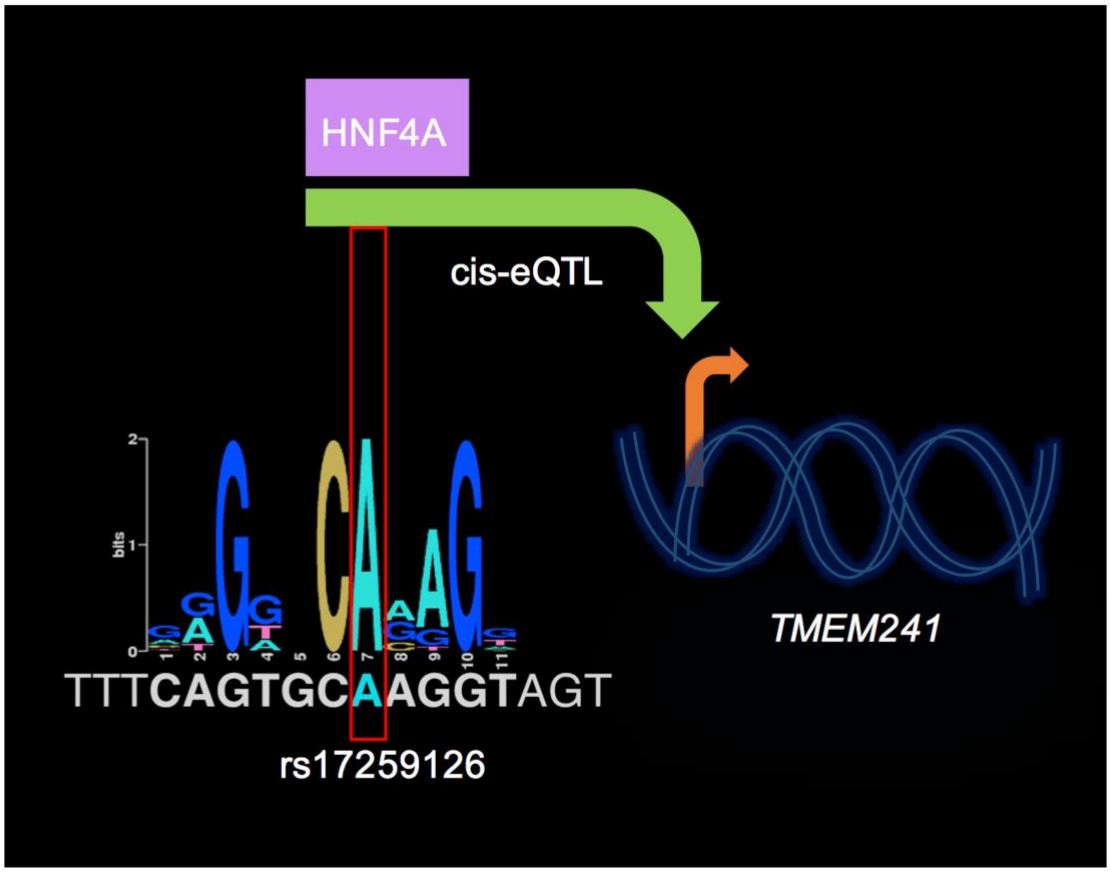
Supplemental Figure II. Representative images for EMSAs for the 9 SNPs in the TG-associated LD block that did not result in allele-specific shifts. Variants rs71360517 and rs4800154 display a TFB shift but no allele-specific differences were observed. No shift was observed for the remaining 7 SNPs (rs9949617, rs67124903, rs9962573, rs4800467, rs1276322, rs9954334, and rs77127070). Three different experimental conditions were tested for each allele: 1) LP (labeled probe); 2) LP and HepG2 nuclear extract (NE); and 3) LP, NE, and unlabeled probe (UP) to act as a competitor.



Supplemental Figure III. The lead TG-associated GWAS SNP rs9949617 is a cis-eQTL for one of the 5 regional genes, transmembrane protein 241 (TMEM241), on chromosome 18q11.2 ($p<1x10^{-5}$ for all 3 tissues). Expression analyses were performed for the regional genes in the ~300 kb region defining the TG-associated window (LD r^2 >0.5 with the lead SNP rs9949617)[5] using experimental data from 856 publicly available human adipose, skin, and lymphocyte RNA microarray samples from the MuTHER resource[7].

Supplemental table I. Primers used for ChIPqPCR.

site	sequence
rs17259126_1	CCCCTTAGGCCCACTACC
	CTTCACACCTCTGGATGCCA
rs17259126_2	TTAGGCCCACTACCTTGCAC
	ACCTCTCTCAACCGTGCTGT
H3k27me3-unbound site	GGACAAGAGAAGAGAGCACAG
	AGGACAGGTAGAAGCCCTAATA
	CTTCTACCTGTCCTGGAAACAG
	GAACCGTGGATGCTTAACAAAG
HNF4 bound site	GGCAAGACTCCTCTGAAAGGA
	TAGGACAGTAGTGGGGCATCA


HNF4A
cis-eQTL
TMEM241
bits
TTTCAGTGCAAGGTAGT
rs17259126

CHAPTER 3

GENOME-WIDE PROFILING OF CONTEXT-SPECIFIC RORA BINDING SITES PROVIDES A CATALOG OF TRANSCRIPTIONAL TARGETS ASSOCIATED WITH LIPID METABOLISM IN THE HUMAN LIVER

Genome-wide profiling of context-specific RORA binding sites provides a catalog of transcriptional targets associated with lipid metabolism in the human liver

Alejandra Rodríguez[1] and Päivi Pajukanta[1,2,3]

[1]Department of Human Genetics, David Geffen School of Medicine at University of California, Los Angeles. [2]Molecular Biology Institute, and [3]Bioinformatics Interdepartmental Program, University of California, Los Angeles

Correspondence: Paivi Pajukanta, Dept. of Human Genetics, David Geffen School of Medicine at UCLA, 695 Charles E. Young Drive South, Los Angeles, CA 90095.

Email:PPajukanta@mednet.ucla.edu; Phone (310) 267-2011; Fax (310) 794-5446

ABSTRACT

Although genome-wide association studies (GWAS) have uncovered a myriad of lipid loci, the molecular mechanisms underlying these associations remain largely uncharacterized. Here we profile genome-wide targets of RAR Related Orphan Receptor A (RORA), a high-density lipoprotein cholesterol (HDL-C) gene identified in our previous Mexican lipid GWAS[1]. Previous studies in mice have shown that when compared to wild type mice, RORA staggerer mutant mice display lower total cholesterol, TGs and hypoalphalipoproteinemia, a condition in which concentrations of HDL-C and its major apolipoprotein component, ApoAI, are reduced. These findings in mice suggest that RORA has a key role in HDL-C metabolism and emphasize the need to profile transcriptional targets of RORA in the human liver. We hypothesized that RORA may display context-specific transcriptional regulation of key lipid genes. To test our hypothesis, we used ChIP-Seq to profile genome-wide context-specific transcriptional targets of RORA in the human liver cancer cell line, HepG2. We treated human HepG2 cells with 200 uM palmitic acid (PA) or BSA (baseline) for 24 hrs. We crossed-linked cells using 1% formaldehyde, fragmented nuclear extracts to 100-300bp, and immunoprecipitated using anti-RORA antibody. We used the Illumina HiSeq 2000 platform and obtained 50-80M reads (75bp single-end reads) per biological replicate (2 BSA and 3 PA) and input controls. Our preliminary results using first ChIP-qPCR show that there is a higher enrichment of RORA in the cytochrome P450 promoter region in the palmitic acid treated HepG2 cells compared to base line treatment. We also saw a small enrichment of RORA bound to APOA3 in palmitic acid treated compared to base line treatment Figure 3-2. Our ChIP-Seq peak annotations using HOMER[9] revealed 62 genes with cardiometabolic functions are located near our peaks. We also found 66 metabolic genes. Gene ontology Analysis[9] also revealed genes with functions in chemical dependency, neurological, and psychiatric pathways.

Our results are consistent with RORA's known function as a regulatory TF in neurological pathways.

INTRODUCTION

Systematic integration of genomics data sets can reveal functional mechanisms underlying GWAS variants and identify new genes with roles in lipid metabolism and networks and key targets for pharmaceutical development. One important target currently investigated for its pharmaceutical potential in multiple sclerosis[2] is the Retinoic acid receptor-related Orphan Receptor A (RORA). RORA is widely expressed nuclear receptor that binds the ROR response element (RORE) sequence within the promoter regions of target genes and modulates their expression across several tissues. In the vascular system, RORA is involved in differentiation of adipocytes[3], control of the vascular tone of small arteries, ischemia-induced angiogenesis, lipid metabolism, and inflammation[4]. Recently is was associated with cellular stress response[5] and found to control hepatic lipid homeostasis[6]. Studies in mice have showed that RORA binds to a RORE in the promoter of APOA1 that encodes a protein component of HDL, and to APOC3 that encodes a protein component of both triglyceride-rich lipoproteins and HDL. RORA is also known to regulate genes in the lipid metabolism pathways such as APOA1, APOA5, APOC3 and PPARG[4,6]. These findings suggest that RORA has a key role in TG and lipoprotein metabolism. We propose to conduct ChIP-Seq to identify targets of RORA in the liver, a key metabolic tissue.

RESULTS

Our ChIP-qPCR results show that there is a higher enrichment of RORA in the cytochrome P450 promoter region in the palmitic acid treated HepG2 cells when compared to the baseline treatment. We also saw a small enrichment of RORA bound to APOA3 in palmitic acid treated cells when compared to the baseline treatment (Figure 3-2). These results prompted us to find

genome-wide sites bound by the TF RORA using ChIP-Seq. For our ChIP-Seq peaks, first we annotated the regional landscape around the peaks by performing functional analysis using HOMER[9]. We annotated genes located next to our peaks and then performed Gene Ontology Analysis to annotate their function. Our results revealed that a total of 62 cardiovascular (Table 3-3) and 66 metabolic genes (Table 3-4) were located near our ChIP-Seq peaks. Our Gene Ontology Analysis revealed that these genes function in chemical dependency, neurological, and psychiatric pathways (Table 3-2). This result is consistent with RORA's known function as a regulatory TF in neurological pathways. Disease annotations were also consistent with RORA's roles in the central nervous system and in metabolic pathways (Tables 3-2 and Table 3-5). Our results represent an initial step in mapping genome-wide targets of RORA in a human liver cancer cell line.

DISCUSSION

The lack of replication between biological replicates for our observed peaks makes our results inconclusive. Several improvements could advance our results. We sequenced our RORA-chromatin IPs using the Illumina HiSeq 2000 platform to identify genome-wide RORA binding sites. With this approach, we were hoping to identify previously unknown RORA binding sites outside promoter regions. However, as all currently available anti-RORA antibodies have low affinities and poor chromatin yield, future experiments could employ a targeted approach using the Affymetrix Human Promoter 1.0R array, which contains over 4.6×106 probes tiled over 25,500 promoter regions of annotated genes. This approach would enrich for promoter regions that are more likely to be bound by RORA[7]. To the best of our knowledge, only one anti-RORA antibody exists with primary characterizations. If secondary characterization of this antibody becomes available, that will improve the chromatin yields. Despite the limitations of our current study, our

results seem consistent with the known biological role of RORA. These results should be considered preliminary due to the lack of replication and the low enrichment of our peaks.

EXPERIMENTAL PROCEDURES

Cell culture

HepG2 cells were grown in ATCC-formulated Eagle's Minimum Essential Medium, supplemented with 10% FBS and 1% penicillin/streptomycin, incubated at 37 °C in a humidified 5% CO2 incubator, and split 1:2 every 2 or 3 days when the cells reached ~80% confluence.

ChIP assays

After cells reached ~80% confluence, the medium was aspirated, and the cells were fixed with 37% formaldehyde for exactly 10 minutes. Glycine was added to a final concentration of 0.125 M to stop the cross-linking, and the cells were rinsed with cold PBS, supplemented with a protease inhibitor cocktail. The cells were scraped and then transferred to a pre-chilled centrifuge tube. Crosslinked cells were pelleted, and nuclear extraction was performed, as previously described with some modifications[8]. Nuclear pellets were re-suspended in SDS Lysis Buffer and fragmented using the BioRuptor sonicator to achieve an average chromatin length of 100-500 bp. Fragmented chromatin was divided into several aliquots and immunoprecipitated one μg of goat anti-RORA1 or control IgG antibody at 4C overnight. On the following day, chromatin was reverse-crosslinked by adding 5 M NaCl to the final concentration of 0.2 M and incubated at 65°C overnight and purified using the Chromatin IP DNA Purification Kit and then submitted for sequencing at the UCLA sequencing core.

Identification of ChIP-seq peaks and tag density profiles

ChIP-Seq analysis was done using HOMER[9] using first the option makeTagDirectory to create a "tag directory" from our high-throughput sequencing alignment files. We also performed

the 4 basic quality control tests produced by the default options. ChIP-Seq enriched regions were identified using HOMER, as has been previously published[9]. Briefly, all ChIP-Seq experiments used the input sequencing as a control, and peaks were called using the findPeaks option in "factor" mode to select a fixed peak size based on estimates from the autocorrelation analysis. This maximizes sensitivity for identifying locations where RORA makes a single contact with the DNA. To increase the overall quality of peaks, we used 3 separate filtering steps employing the input sequencing as a control, filtering based on local signal, and filtering based on clonal signal.

Annotation of genomic regions

To annotate the regional landscape around the peaks, we used the perl script from HOMER[9] annotatePeaks.pl. We performed two main types of annotations. First, we associated peaks with the nearby genes. Second, we performed the Gene Ontology Analysis. Third, we performed genomic feature association analysis.

Basic annotation

We mined the data using all options in the annotatePeaks.pl script from HOMER[9]. In summary, the output generated using the basic gene annotation options included the following: Nearest TSS, Nearest TSS: Entrez Gene ID, Nearest TSS: Unigene ID, Nearest TSS: RefSeq ID, and Nearest TSS: Ensembl ID.

Peak annotation enrichment: Gene ontology analysis of associated genes

ChIP-Seq peaks can preferentially be found near genes with specific biological functions. We used the annotatePeaks.pl from HOMER[9], using the gene ontology classifications, to perform an annotation enrichment analysis. In this analysis, HOMER uses the list of genes associated with our regions and searches for enriched functional categories.

FIGURE LEGENDS

Table 3-1 **RORA ChIP-qPCR primers for known targets of RORA and control repressed chromatin sites**

Table 3-2 **Functional enrichment analysis for RORA ChIP-Seq peaks**

Table 3-3 **Functional enrichment analysis for RORA ChIP-Seq peaks: Cardiovascular genes**

Table 3-4 **Functional enrichment analysis for RORA ChIP-Seq peaks: Metabolic genes**

Table 3-5 **Functional enrichment analysis for RORA ChIP-Seq peaks: Disease categories**

Figure 3-1 **HepG2 chromatin fragments used in the RORA ChIP-Seq assays**

Nuclear pellets were fragmented using the BioRuptor sonicator to an average chromatin length of 100-500 bp.

Figure 3-2 **RORA ChIP-qPCR assays show enrichment of known RORA target sites relative to control sites**

ChIP-qPCR results for RORA in HepG2 cells shows a higher enrichment of the RORA TF in the Cytochrome P450 promoter region with palmitic acid treatment when compared to the baseline treatment. We also saw a small enrichment of RORA bound to APOA3 with the palmitic acid treatment when compared to baseline treatment.

Figure 3-1 **HepG2 cell chromatin fragments used in the RORA ChIP-Seq assays**

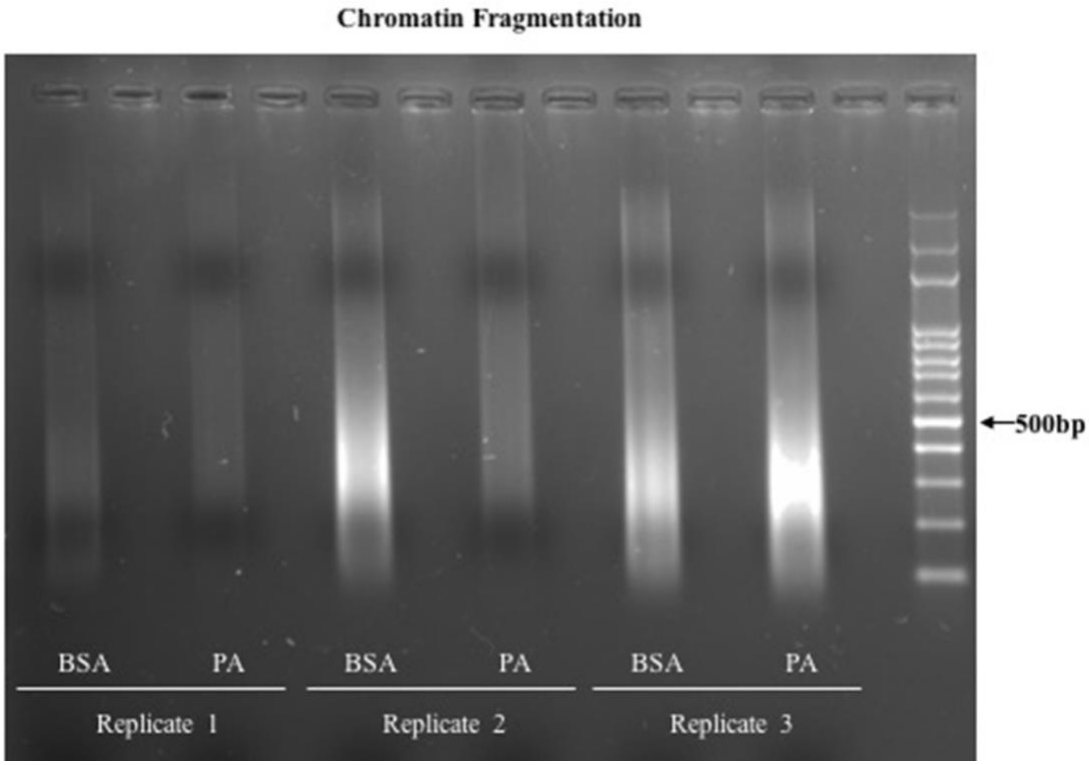


Table 3-1 **RORA ChIP-qPCR primers for known targets of RORA and control repressed chromatin sites**

Target	Sequence
APOA1-PROMOTER FWD Set 1	TGG AGG CGG ACA ATA TCT TTA C
APOA1-PROMOTER FWD Set 4	TTA CCA GTT TGG GAG GCT TG
APOA1-PROMOTER FWD Set 5	CTT TGC CCA GAG GTC TTC TC
APOA1 FWD Set 1	TGG AGG CGG ACA ATA TCT TTA C
APOA1 FWD Set 4	TTA CCA GTT TGG GAG GCT TG
APOA1 FWD Set 5	CTT TGC CCA GAG GTC TTC TC
APOA5 FWD Set 1	GCC TCT TGC CAT CTC ATC TT
APOA5 FWD Set 2	CAG GTC AGT GGG AAG GTT AAA G
APOA5 FWD Set 4	GAG GGA TGT GGT TGG TCT TT
APOA3-PROMOTER FWD Set 1	GGA TTG AAA CCC AGA GAT GGA
APOA3-PROMOTER FWD Set 4	AAG CCA CCC ACT TGT TCT C
APOA3-PROMOTER FWD Set 5	GGC CTA TGT CCA AGC CAT TT
CYP FWD Set 1	CTA AGA AGT GAG GAA CCC AAG G
CYP FWD Set 2	CCT GTC TCA CTC TCT TCC TGT A
CYP FWD Set 3	TAC CAC GCT GTT CTG CAA TC
CYP FWD Set 4	CAT CTT GAG GGA CAA GCA GAG
SULT2A1 FWD Set 1	GAG GTA TAA TGT GAC CCA TAC TCAA
SULT2A1 FWD Set 2	CGA ATA ACA AAC ACG AGG ACA AA
SULT2A1 FWD Set 3	GAA GAT GTT GAG CAA TCA TGA ACT
SULT2A1 FWD Set 4	ATA ATC CTG CAA TCG TGC ATT T

Figure 3-2 **RORA ChIP-qPCR assays show an enrichment of known RORA target sites when compared to the control sites**

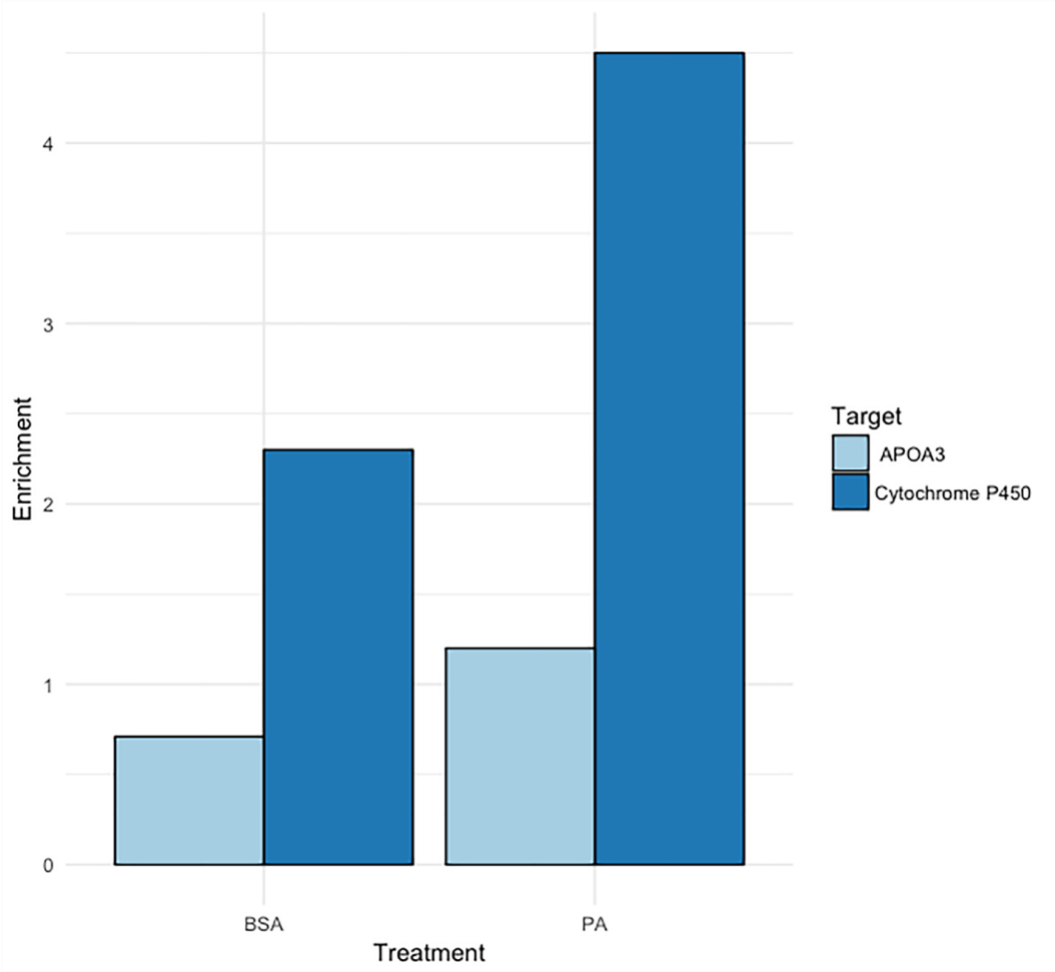


Table 3-2 **Functional enrichment analysis for RORA ChIP-Seq peaks.**

Term	p-value	Adjusted p-value	Number of genes
Cardiovascular	5.5×10^{-7}	9.9×10^{-6}	62
Metabolic	4.3×10^{-5}	3.9×10^{-4}	66
Chemical dependency	2.0×10^{-3}	1.2×10^{-2}	46
Hematological	1.5×10^{-2}	6.7×10^{-2}	22
Neurological	1.9×10^{-2}	6.8×10^{-2}	35
Psychiatric	3.7×10^{-2}	1.1×10^{-1}	25

Table 3-1 **Functional enrichment analysis for RORA ChIP-Seq peaks: Cardiovascular genes.**

ID	Name
NM_004274	A-kinase anchoring protein 6(AKAP6)
NM_152522	ADP ribosylation factor like GTPase 6 interacting protein 6(ARL6IP6)
NM_175060	C-type lectin domain family 14 member A(CLEC14A)
NM_001805	CCAAT/enhancer binding protein epsilon(CEBPE)
NM_001250	CD40 molecule(CD40)
NM_017774	CDK5 regulatory subunit associated protein 1 like 1(CDKAL1)
NM_020943	CWC22 spliceosome associated protein homolog(CWC22)
NM_203301	F-box protein 33(FBXO33)
NM_001002911	G protein-coupled receptor 139(GPR139)
NM_148963	G protein-coupled receptor class C group 6 member A(GPRC6A)
NM_017769	G2/M-phase specific E3 ubiquitin protein ligase(G2E3)
NM_018557	LDL receptor related protein 1B(LRP1B)
NM_080676	MACRO domain containing 2(MACROD2)
NM_002515	NOVA alternative splicing regulator 1(NOVA1)
NM_153355	Na+/K+ transporting ATPase interacting 2(NKAIN2)
NM_032784	R-spondin 3(RSPO3)
NM_018723	RNA binding protein, fox-1 homolog 1(RBFOX1)
NM_018460	Rho GTPase activating protein 15(ARHGAP15)
NM_005168	Rho family GTPase 3(RND3)
NM_030623	SPHK1 interactor, AKAP domain containing(SPHKAP)
NM_015910	WD repeat containing planar cell polarity effector(WDPCP)
NM_003917	adaptor related protein complex 1 gamma 2 subunit(AP1G2)
NM_001195	beaded filament structural protein 1(BFSP1)
NM_001200	bone morphogenetic protein 2(BMP2)
NM_005795	calcitonin receptor like receptor(CALCRL)
NM_201548	ceramide kinase like(CERKL)
NM_004824	chromodomain Y-like(CDYL)
NM_032221	chromodomain helicase DNA binding protein 6(CHD6)
NM_015585	cilia and flagella associated protein 61(CFAP61)
NM_018431	docking protein 5(DOK5)
NM_022073	egl-9 family hypoxia inducible factor 3(EGLN3)
NM_005235	erb-b2 receptor tyrosine kinase 4(ERBB4)

NM_000145	follicle stimulating hormone receptor(FSHR)
NM_004131	granzyme B(GZMB)
NR_004855	hepatocellular carcinoma up-regulated long non-coding RNA(HULC)
NM_138574	hepatoma derived growth factor-like 1(HDGFL1)
NM_003855	interleukin 18 receptor 1(IL18R1)
NM_024336	iroquois homeobox 3(IRX3)
NM_018214	leucine rich repeat containing 1(LRRC1)
NM_015702	methylmalonic aciduria and homocystinuria, cblD type(MMADHC)
NM_004801	neurexin 1(NRXN1)
NM_002511	neuromedin B receptor(NMBR)
NM_002500	neuronal differentiation 1(NEUROD1)
NM_138285	nucleoporin 35(NUP35)
NM_000933	phospholipase C beta 4(PLCB4)
NM_002742	protein kinase D1(PRKD1)
NM_001080545	protein phosphatase 1 regulatory inhibitor subunit 1C(PPP1R1C)
NM_002844	protein tyrosine phosphatase, receptor type K(PTPRK)
NM_001145204	shisa family member 9(SHISA9)
NM_014178	syntaxin binding protein 6(STXBP6)
NM_173485	teashirt zinc finger homeobox 2(TSHZ2)
NM_052913	transmembrane protein 200A(TMEM200A)
NM_006296	vaccinia related kinase 2(VRK2)

Table 3-2 **Functional enrichment analysis for RORA ChIP-Seq peaks: Metabolic genes.**

ID	Name
NM_004274	A-kinase anchoring protein 6(AKAP6)
NM_175060	C-type lectin domain family 14 member A(CLEC14A)
NM_001250	CD40 molecule(CD40)
NM_017774	CDK5 regulatory subunit associated protein 1 like 1(CDKAL1)
NM_020943	CWC22 spliceosome associated protein homolog(CWC22)
NM_203301	F-box protein 33(FBXO33)
NM_001002911	G protein-coupled receptor 139(GPR139)
NM_018557	LDL receptor related protein 1B(LRP1B)
NM_080676	MACRO domain containing 2(MACROD2)
NM_014048	MKL1/myocardin like 2(MKL2)
NM_002515	NOVA alternative splicing regulator 1(NOVA1)
NM_153355	Na+/K+ transporting ATPase interacting 2(NKAIN2)
NM_016436	PHD finger protein 20(PHF20)
NM_032784	R-spondin 3(RSPO3)
NM_018723	RNA binding protein, fox-1 homolog 1(RBFOX1)
NM_018460	Rho GTPase activating protein 15(ARHGAP15)
NM_005168	Rho family GTPase 3(RND3)
NM_030623	SPHK1 interactor, AKAP domain containing(SPHKAP)
NM_001200	bone morphogenetic protein 2(BMP2)
NM_005795	calcitonin receptor like receptor(CALCRL)
NM_080617	cerebellin 4 precursor(CBLN4)
NM_004824	chromodomain Y-like(CDYL)
NM_001898	cystatin SN(CST1)
NM_018431	docking protein 5(DOK5)
NM_022073	egl-9 family hypoxia inducible factor 3(EGLN3)
NM_005235	erb-b2 receptor tyrosine kinase 4(ERBB4)
NM_000145	follicle stimulating hormone receptor(FSHR)
NM_004752	glial cells missing homolog 2(GCM2)
NR_004855	hepatocellular carcinoma up-regulated long non-coding RNA(HULC)
NM_138574	hepatoma derived growth factor-like 1(HDGFL1)
NM_017545	hydroxyacid oxidase 1(HAO1)
NM_003855	interleukin 18 receptor 1(IL18R1)
NM_024336	iroquois homeobox 3(IRX3)
NM_024704	kinesin family member 16B(KIF16B)
NM_152447	leucine rich repeat and fibronectin type III domain containing 5(LRFN5)

NM_178839	leucine rich repeat transmembrane neuronal 1(LRRTM1)
NM_019888	melanocortin 3 receptor(MC3R)
NM_199290	nascent polypeptide associated complex alpha subunit 2(NACA2)
NM_004801	neurexin 1(NRXN1)
NM_002500	neuronal differentiation 1(NEUROD1)
NM_005048	parathyroid hormone 2 receptor(PTH2R)
NM_207499	patched domain containing 4(PTCHD4)
NM_021213	phosphatidylcholine transfer protein(PCTP)
NM_000933	phospholipase C beta 4(PLCB4)
NM_144773	prokineticin receptor 2(PROKR2)
NM_002742	protein kinase D1(PRKD1)
NM_002844	protein tyrosine phosphatase, receptor type K(PTPRK)
NM_001145204	shisa family member 9(SHISA9)
NM_001049	somatostatin receptor 1(SSTR1)
NM_001052	somatostatin receptor 4(SSTR4)
NM_014178	syntaxin binding protein 6(STXBP6)
NM_173485	teashirt zinc finger homeobox 2(TSHZ2)
NM_021156	thioredoxin related transmembrane protein 4(TMX4)
NM_018286	transmembrane protein 100(TMEM100)
NM_052913	transmembrane protein 200A(TMEM200A)
NM_006296	vaccinia related kinase 2(VRK2)

Table 3-3 **Functional enrichment analysis for RORA ChIP-Seq peaks: Disease categories.**

Disease	Number of genes	p-value	Adjusted p-value
Mental abnormalities,	2	$6.40x^{-5}$	$1.60x^{-2}$
Neuroblastoma	8	$7.40x^{-5}$	$4.50x^{-4}$
Waist-Hip Ratio	8	$1.00x^{-4}$	$5.90x^{-4}$
Urinalysis	4	$1.30x^{-4}$	$2.20x^{-3}$
lymphoblastic leukemia	2	$1.90x^{-4}$	$2.40x^{-2}$
Apolipoproteins B	4	$2.90x^{-4}$	$3.90x^{-3}$
Stroke	13	$3.80x^{-4}$	$1.20x^{-3}$
Platelet Count	6	$4.00x^{-4}$	$2.70x^{-3}$
Hematocrit	5	$2.10x^{-3}$	$1.20x^{-2}$
Obesity	7	$4.00x^{-3}$	$1.50x^{-2}$
Body Mass Index	10	$5.00x^{-3}$	$1.40x^{-2}$
Erythrocytes	4	$5.90x^{-3}$	$3.50x^{-2}$
Cholesterol, HDL	9	$8.00x^{-3}$	$2.20x^{-2}$
Albuminuria	3	$8.30x^{-3}$	$6.40x^{-2}$
Bone Density	6	$9.10x^{-3}$	$3.20x^{-2}$
Tobacco Use Disorder	35	$1.10x^{-2}$	$1.60x^{-2}$
Heart Failure	8	$1.20x^{-2}$	$3.40x^{-2}$
Basophils	3	$1.30x^{-2}$	$8.50x^{-2}$
Breast |prostate cancer	3	$1.30x^{-2}$	$8.70x^{-2}$
Diabetes Mellitus, Type 2	5	$1.40x^{-2}$	$5.30x^{-2}$
Body Height	10	$1.60x^{-2}$	$3.80x^{-2}$
Monocytes	3	$1.70x^{-2}$	$9.90x^{-2}$
Hemoglobins	6	$1.80x^{-2}$	$5.60x^{-2}$
Erythrocyte Count	5	$1.90x^{-2}$	$6.70x^{-2}$
Blood Pressure	8	$3.20x^{-2}$	$7.50x^{-2}$

REFERENCES

1. Ko, A. *et al.* Amerindian-specific regions under positive selection harbour new lipid variants in Latinos. *Nat. Commun.* **5,** 3983 (2014).
2. Eftekharian, M. M. *et al.* RAR-related orphan receptor A (RORA): A new susceptibility gene for multiple sclerosis. *J. Neurol. Sci.* **369,** 259–262 (2016).
3. Gaudet, D. *et al.* Antisense Inhibition of Apolipoprotein C-III in Patients with Hypertriglyceridemia. *N. Engl. J. Med.* **373,** 438–447 (2015).
4. Liu, Y. *et al.* Retinoic acid receptor-related orphan receptor α stimulates adipose tissue inflammation by modulating endoplasmic reticulum stress. *J. Biol. Chem.* **292,** 13959–13969 (2017).
5. Zhu, Y., McAvoy, S., Kuhn, R. & Smith, D. I. RORA, a large common fragile site gene, is involved in cellular stress response. *Oncogene* **25,** 2901–2908 (2006).
6. Kim, K. *et al.* RORα controls hepatic lipid homeostasis via negative regulation of PPARγ transcriptional network. *Nat. Commun.* **8,** 162 (2017).
7. Sarachana, T. & Hu, V. W. Genome-wide identification of transcriptional targets of RORA reveals direct regulation of multiple genes associated with autism spectrum disorder. *Mol. Autism* **4,** 14 (2013).
8. Blattler, A. *et al.* ZBTB33 binds unmethylated regions of the genome associated with actively expressed genes. *Epigenetics Chromatin* **6,** 13 (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).

CHAPTER 4

ACLY MAY MEDIATE DECANALIZATION OF LIPID METABOLISM PATHWAYS VIA HISTONE ACETYLATION

ACLY may mediate decanalization of lipid metabolism pathways via histone acetylation

Alejandra Rodríguez[1] and Päivi Pajukanta[1,2,3]

[1]Department of Human Genetics, David Geffen School of Medicine at University of California, Los Angeles. [2]Molecular Biology Institute, and [3]Bioinformatics Interdepartmental Program, University of California, Los Angeles

Correspondence: Paivi Pajukanta, Dept. of Human Genetics, David Geffen School of Medicine at UCLA, 695 Charles E. Young Drive South, Los Angeles, CA 90095.

Email:PPajukanta@mednet.ucla.edu; Phone (310) 267-2011; Fax (310) 794-5446

ABSTRACT

Genome-wide association studies (GWAS) have discovered hundreds of lipid loci, yet it has become increasingly clear that variation at the known loci explains only a small fraction of the heritability. Contributions to variation in lipid traits that can be attributed to complex genetic models, such as gene-environment and epistatic interactions, may help uncover additional sources of this "missing heritability." The presence of an interaction is expected to increase the variance of a trait in subjects having the interacting genetic variant. To this end, we performed a two-step approach to identify genes with increased expression variance and their underlying variance expression (ve)-QTLs. We investigated Mexican[1] individuals exhibiting extreme serum TGs from an adipose expression microarray cohort (n=64). We found that individuals with low TGs displayed a greater ATP citrate lyase (ACLY) expression variance than the individuals with high TGs (p-value=$2.0x10^{-2}$), and this result was replicated in the Finnish METabolic Syndrome In Men[2] (METSIM) adipose RNA-Seq cohort (n=335) (p-value=$1.8x10^{-03}$).

ACLY encodes the primary enzyme responsible for the synthesis of nuclear and cytosolic acetyl-CoA, which is important for the biosynthesis of fatty acids, a precursor of TGs. One hypothesis is that a reduced ACLY expression variance under increased TG-context corresponds to an increased degree of constraint in lipid biosynthesis pathways and decreased robustness in its response to environmental stimuli and buffering ability against cryptic genetic variation. We used a correlation least squared (CLS) test to uncover SNPs associated with ACLY expression variance and found that the reference allele of variant rs34272903 (T/C) is associated with increased ACLY expression variance (FWER p-value=$1.0x10^{-4}$). Our results suggest that the reference T allele of rs34272903 interacts with an unknown factor under the low TG-context, increasing the variance

of ACLY expression. Accordingly, this interaction may contribute to efficient responses in lipid pathway activation to endo-exogenous stimuli via unknown mechanisms.

INTRODUCTION

Lipid GWASs have identified hundreds of new loci associated with lipid traits, yet it has become increasingly clear that variation at the known loci explains only a small part of the trait heritability. GWASs focus on associations between a single tag locus and a trait and do not capture gene-gene, or gene-environment interactions. Contributions to variation in lipid-related traits that can be attributed to complex genetic models, such as gene-environment and epistatic interactions, may help uncover additional sources of this "missing heritability." It is known that about 85% of the common genetic variation is associated with the expression of protein coding genes[3]. Fine-mapping eQTL studies typically search for differences in the mean of gene expression between the two alleles of a GWAS hit, whereas differences invariance of expression is often regarded as noise. Recent studies have reported that variance in gene expression is also genetically determined and should be regarded as a quantitative trait that can be investigated using ve-QTL mapping[4,5]. Two mechanisms have been put forward to explain the formation of ve-QTLs. First, ve-QTLs could underlie epistatic interactions and secondly, ve-QTL may be created via decanalization of biochemical pathways[6–8]. Environmentally derived metabolites can act as signaling molecules to chromatin, which may lead to chromatin remodeling and regulation of entire biochemical pathways[9]. This "openness" or interplay between biological systems and their environment can also lead to toxicity and disease[6,7,10] (further reviewed in Chapter 5). Using variance association mapping, we and others identified genetic loci associated with gene expression variance[4,8,11,12]. The ve-QTLs may be part of the genetic architecture of dyslipidemias. Detailed characterization of genes and variants underlying cardiometabolic disease will lead to previously undiscovered

insights of disease etiology as well as to improved screening, diagnosis, prognosis, and drug development for these common disorders.

RESULTS

Subcutaneous adipose tissue is a metabolically active tissue. During the terminal stage of adipocyte differentiation, the transcriptions factors C/EBPα and PPARγ turn on the expression of adipogenic enzymes, which together form networks with highly specialized functions, such as lipid storage and synthesis**.** We hypothesized that a high TG context might interact with one or more genetic variants leading to an increased variance in gene expression of lipogenic and lipolytic enzymes. Despite its relevance as a vital metabolic tissue, there is a remarkable lack of large enough data sets of human adipose tissue. Here, we obtained microarray expression data of 17 genes in the lipid metabolism pathway for 64 Mexican subcutaneous adipose samples[1]. We corrected the expression values for age, gender, and ancestry using PC1 and PC2 from a genetic principal component analysis (PCA). To remove unknown confounders, we also removed the first three principal components of the expression values. To search for genes exhibiting context-dependent expression variance, we divided the samples by TG group (high versus low), and as a measure of variance, we computed the observed standard deviation for each gene. To derive the null distribution, we combined low and high TG samples and permuted their group categories and randomly assigned their TG groups and again computed a permuted standard deviation for each gene. Of the 17 genes, we found a total of 13 probes exhibiting significant context-dependent expression variance p-value ≤ 0.05, and after removing redundant probes, we had nine significant genes (Table 4-1, Figure 4-1). Of these, only ACSBG1 passed the Bonferroni correction for multiple testing. ACSBG1, very long-chain acyl-CoA synthetase, is capable of activating very long-chain fatty acids. To replicate our nominally significant genes, we obtained RNA-Seq data

from subcutaneous adipose tissue for 64 samples with similar TG characteristics from the METSIM cohort. Three (ACACA, p-value=$2.3x10^{-2}$; ACLY, p-value =$1.8x10^{-3}$; and ACSBG1, p-value <0.0001) of the 9 genes passed the significance threshold for multiple testing correction (Table 4-2, Figure 4-2).

Variance expression quantitative trait loci (ve-QTLs) or genetic variants associated with expression variance differences between two alleles have been observed in genetic studies[4,11,12] and may underlie our observed context-dependent expression variance for the ACACA, ACLY, and ACSBG1 genes. To investigate this possibility, we obtained the genotypes and RNA-Seq expression data from 335 subcutaneous adipose samples from the Finnish METSIM cohort[2]. We took variants with MAF>5% located within +/- 1M window from the transcription start site (TSS) of each of the three genes and performed SNP-gene expression variance associations using ve-QTL mapper[4]. We found that variant rs34272903 is a ve-QTL for the ATP citrate lyase (ACLY) gene in the METSIM cohort. ACLY is the primary enzyme responsible for the synthesis of cytosolic and nuclear acetyl-CoA. We also found rs6607284 (p-value=$3.8x10^{-3}$) to be associated with variance in expression of ACACA and rs6495382 (p-value=$8.0x10^{-4}$) for ACSBG1; however, these associations did not pass multiple testing correction.

DISCUSSION

In-depth understanding the disease etiology of lipid disorders requires characterization of their genetic architecture. Most current methods in quantitative genetics focus on the effect of a genetic variant on the gene expression mean (eQTLs); however, variants have also been shown to affect the gene expression variance[4,5,11,12] (ve-QTLs). Genetic variants that control the gene expression variance may underlie non-additive epistatic and gene environment interactions. The complexity of these genetic interactions also means that they may be population-dependent, and

therefore difficult to replicate; hence, there is a lack of replicated human ve-QTL studies in the literature. Here we used ve-QTL-mapper[4] to perform SNP-gene expression variance analyses. To alleviate multiple testing and focus on genes with a function in lipid metabolism, we selected genes that exhibited context-dependent gene expression variance. We hypothesized that ve-QTLs are formed as a result of a decanalization event in individuals with high TG-levels and that this decanalization is mediated via interactions between environment (TG-context) and multiple genetic factors. For our discovery cohort, we used Mexican microarray data. RNA-Seq expression data from the METSIM cohort were used for replication. We found 3 genes with significant context-dependent expression variance in both populations. Next, we looked for variants within +/- 1M of the gene TSS and performed ve-QTL mapping. We found that rs34272903 (T/C) is associated with increased ACLY expression variance. ACLY is the primary enzyme responsible for the synthesis of cytosolic and nuclear acetyl-CoA in many tissues. Acetyl-CoA, is a precursor of malonyl CoA a substrate for FASN in the synthesis of fatty acids, while nuclear acetyl-CoA regulates histone acetylation levels. A limitation of this study is that we only tested for the SNP-SNP interaction and did not take the TG context into account. Replication of our ve-QTL variant in the GTEx subcutaneous adipose RNA-Seq did not pass the significance threshold. However, the GTEx dataset is small and comes from several different populations. Furthermore, the biochemical properties of post-mortem tissue constitute a significant source of heterogeneity. These factors may contribute to the lack of replication.

We focused on genes known to function in the lipid metabolism pathway. Future studies could also include pathways in the insulin pathway, the adrenergic pathway, and the atrial natriuretic hormone pathway, all of which control lipid storage and secretion in adipose tissue[13]. Particular emphasis could be placed on two enzymes, lipoprotein lipase (LPL) and hormone-

sensitive lipase (HSL), both of which are part of the adrenergic pathway, and they are thus important regulators of lipid storage and mobilization.

EXPERIMENTAL PROCEDURES

Mexican gene expression microarrays

Sample collection and processing have been described in detail previously[1]. Briefly, 70 Mexican familial combined hyperlipidemia (FCHL) case/control fat biopsies were collected from umbilical subcutaneous adipose tissue under local anesthesia. Detailed procedures for RNA extraction and microarray hybridization have been described previously[1]. The microarray data can be accessed in MIAME compliant format from the NCBI Gene Expression Omnibus (GEO) database (GSE17170). Each participant provided a written informed consent. The study design was approved by the ethics committees of the INCMNSZ and UCLA.

Mexican microarray data: Pre-processing and quality control

CEL files were imported into R version 3.3.3 as an AffyBatch using the ReadAffy function from the Bioconductor Affy package[14]. The affyBatch was then converted into an ExpressionSet class using the gcrma package. The gcrma function adjusts for background intensities, including optical noise and non-specific binding. In addition, background adjusted probe intensities were converted to expression measures using the same normalization and summarization methods as rma (Robust Multiarray Average).

Adjusting for batch effects

The study sample consisted of Mexican individuals with extreme TG values and subcutaneous adipose tissue expression microarrays available for study (n=64). Samples were divided into extremely low (n=32) and high (n=32) TG groups and linear mixed model was applied

with gene expression as dependent variable and using age, sex, and batch as covariates. Principal component analyses (PCA) were performed using the R function prcomp.

Testing for significance in the Mexican discovery cohort

The residuals from the linear model were used to compute the standard deviation for each gene of the 17 lipid pathway genes. A total of 13 genes with a p-value ≤0.05 in the Mexican discovery cohort were used in the replication METSIM cohort. The residuals from the linear model were used to compute the group standard deviation for each of the nine genes. For each gene, we defined our test statistic as the absolute value of the difference between the standard deviation in the low TG and that of the high TG group: gene standard deviation difference=$|(sd_{low}-sd_{high})|$. For our test statistic, the null distribution was computed by permuting the TG status labels and the p-value was computed by dividing the number of times the shuffled test statistic exceeded the observed test statistic divided by the number of permutations. Genes with a p-value ≤ 0.05 were considered as promising and subsequently used in the replication data set.

METSIM replication cohort

Subcutaneous adipose biopsies and RNA extraction

Subcutaneous adipose tissue samples were taken by needle biopsy under local anesthesia (lidocaine 10 mg/mL without adrenaline)[15]. Total RNA was isolated using Qiagen miRNeasy kit (Qiagen, Hilden, Germany) according to manufacturer's instructions. Polyadenylated mRNA was prepared for sequencing using the Illumina TruSeq RNA Sample Preparation Kit v2[15]. The qualities of the total RNA were evaluated using the Agilent 2100 Bioanalyzer with the RNA 6000 Nano kit (Agilent Technologies). Only RNA with RIN value greater than 6.8 and 28S/18S rRNA ratios greater than 1.5 was used for downstream RNA sequencing.

RNA sequencing and read mapping

RNA was sequenced on the Illumina HiSeq 2000 platform (Ilumina, San Diego, CA, USA)[15], generating 50-base-pair, paired-end reads at an average depth of 43 million reads, according to manufacturer's protocol. Paired-end reads were aligned using STAR version 2.4.1d[16] using annotations-based mapping method with Gencode v19 transcriptome definition[17]. Samtools version 0.1.18 was used to process the SAM and BAM alignment files[18]. To identify any mix-ups between the DNA and RNA samples, we used the VerifyBamID program[19].

Gene expression estimation

To estimate gene expression, we aligned reads to the human genome and counted reads for each gene. Specifically, we first used FastQC[20] to check the quality of these generated reads. Then, we aligned reads using the 2-pass version of STAR 2.4.1d[16] with the GENCODE v19 annotations[17]. To generate read counts for each gene, we first generated annotations that merge exons from all transcript isoforms of a gene. We only counted reads for genes annotated as "lincRNA" or "protein-coding". Since the reads were non-stranded, we removed genomic intervals that contain 2 overlapping genes on opposite strands to remove the ambiguity from which strand the read came from. We used custom software to count reads that fully overlapped with these exon regions without aligning to introns. To adjust for library size and gene lengths, we generated fragments per kb mapped per million mapped reads (FPKM) estimates. We removed lowly expressed genes by selecting genes that had FPKM > 0 in at least 90% of individuals. FPKM values for filtered genes were log transformed. To identify technical variation and estimate quality control metrics, we applied Picard Tools[21] to the alignments and principal components analysis (PCA) to the gene expression data. We regressed out the following technical covariates: percent coding bases, percent UTR bases, percent intronic bases, percent intergenic bases, percent mRNA

bases, median 3' bias, median 5' to 3' bias, and median CV coverage, RIN, total mapped reads, percent of reads uniquely mapped, percent reads from mitochondria, batch, and average allelic expression imbalance.

Adjusting for batch effects

Data modeling was done following the same design used in the discovery Mexican cohort. For extreme low (n=34) and high (n=34) TG groups, a linear mixed model was applied with gene expression as a dependent variable and using a total of 17 technical factors as covariates. Principal component analyses (PCA) were performed using the R function prcomp.

Testing for significance in the METSIM replication cohort

A total of 13 genes with a nominal p-value ≤ 0.05 in the Mexican discovery cohort were tested in the METSIM[2] replication cohort. The residuals from the linear model were used to compute the group standard deviation for each of the nine genes. For each gene, we defined our test statistic as the absolute value of the difference between the standard deviation in the low TG and that of the high TG group: gene standard deviation difference $=|(sd_{low}-sd_{high})|$. For our test statistic, the null distribution was computed by permuting the TG status labels and the p-value was computed by dividing the number of times the shuffled test statistic exceeded the observed test statistic divided by the number of permutations. All p-values were adjusted for multiple testing using Bonferroni correction.

FIGURE LEGENDS

Table 4-1 **Variance of gene expression in the Mexican discovery cohort**

Table 4-2 **Variance of gene expression in the METSIM replication cohort**

Figure 4-1 **Genes in the lipid metabolism pathway that display context-dependent expression variance in the Mexican cohort**
A total of 13 probes representing 9 genes displayed a significant context-dependent expression variance. Six genes had a higher expression variance in the low TG group (orange, n=32), while the remaining three genes, ACSBG1, PRKAA2, and PRKAR1A displayed a higher expression variance in the high TG (blue, n=32) group. The p-values were obtained from the two-sided test with 1,000 permutations. Genes with a p-value ≤ 0.05 were subsequently tested in the replication data set.

Figure 4-2 **Genes in the lipid metabolism pathway that display context-dependent expression variance in the METSIM cohort**
We replicated the observed context-dependent expression variance of 3 of the 9 genes (ACACA, ACLY, and ACSBG1) originally found in the discovery cohort. The expression variance for ACACA and ACLY was higher in the low TG group (n=34, orange), while ACSBG1 displayed a higher expression variance in the high TG group (n=34, blue), consistent with our observations in the Mexican discovery cohort.

Figure 4-3 **Variant rs34272903 is a ve-QTL for the ACLY gene in the METSIM cohort**
The first two columns indicate the gene ID and SNP ID. Then we report the Spearman correlation between the distance measure and genotype (Cor), p-value, permuted p-value (not controlled for multiple testing), and permuted p-value controlled for multiple testing across the SNPs (FWER).

Table 4-1 **Variance of gene expression in the Mexican discovery cohort**

Probe	Gene	p-value	Adjusted p-value
206465_at	ACSBG1	<0.001	<0.001
225278_at	PRKAB2	$6.0x10^{-3}$	0.40
212186_at	ACACA	$7.0x10^{-3}$	0.40
200604_s_at	PRKAR1A	$1.9x10^{-3}$	1.0
210337_s_at	ACLY	$2.0x10^{-2}$	1.0
212609_s_at	AKT3	$2.5x10^{-2}$	1.0
201128_s_at	ACLY	$2.7x10^{-2}$	1.0
212607_at	AKT3	$2.7x10^{-2}$	1.0
201127_s_at	ACLY	$3.3x10^{-2}$	1.0
207163_s_at	AKT1	$3.8x10^{-2}$	1.0
227892_at	PRKAA2	$3.8x10^{-2}$	1.0
226156_at	AKT2	$5.0x10^{-2}$	1.0
225471_s_at	AKT2	$5.5x10^{-2}$	1.0

Table 4-1 **Summary of the results for the gene expression variance in the Mexican discovery cohort.** In the discovery data set of Mexicans (n=64), 13 probes representing 9 genes reached a nominal significance and of these ACSBG1 passed the multiple testing correction (p-value <0.001). Column "Adjusted p-value" shows the Bonferroni corrected p-value for the two-sided test after 1,000 permutations. Genes with a p-value ≤ 0.05 were considered as promising candidates and they were subsequently tested in the replication data set.

Table 4-2 **Variance of gene expression in the METSIM replication cohort**

Gene	p-value	Adjusted p-value
ACACA	$2.5x10^{-3}$	**$2.0x10^{-2}$**
ACLY	$2.0x10^{-4}$	**$2.0x10^{-3}$**
ACSBG1	<0.0001	**<0.0001**
AKT1	1.0	1.0
AKT2	0.20	1.0
AKT3	$1.5x10^{-2}$	0.10
PRKAA2	$7.8x10^{-3}$	$7.0x10^{-2}$
PRKAB2	0.20	1.0
PRKAR1A	0.30	1.0

Table 4-2 **Summary of the results for the gene expression variance in the METSIM Replication Cohort.** Three of the nine nominally significant genes from the discovery cohort were replicated in the METSIM cohort (n=68). Column "Adjusted p-value" shows the Bonferroni corrected p-value for the two-sided test after 10,000 permutations.

Figure 4-1 **Genes in the lipid metabolism pathway that display context-dependent expression variance in the Mexican cohort**

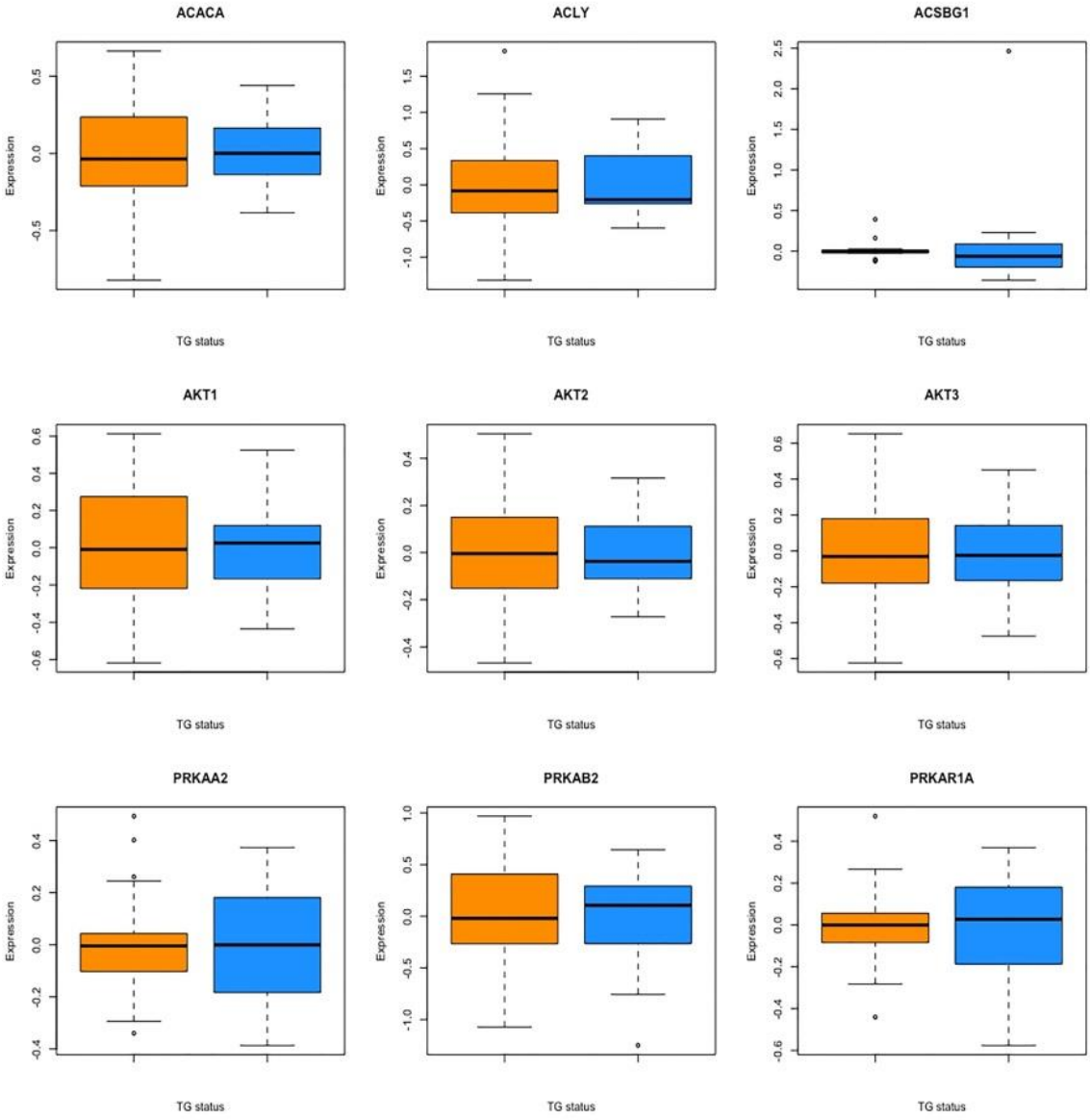


Figure 4-2 **Genes in the lipid metabolism pathway that display context-dependent expression variance in the METSIM cohort**

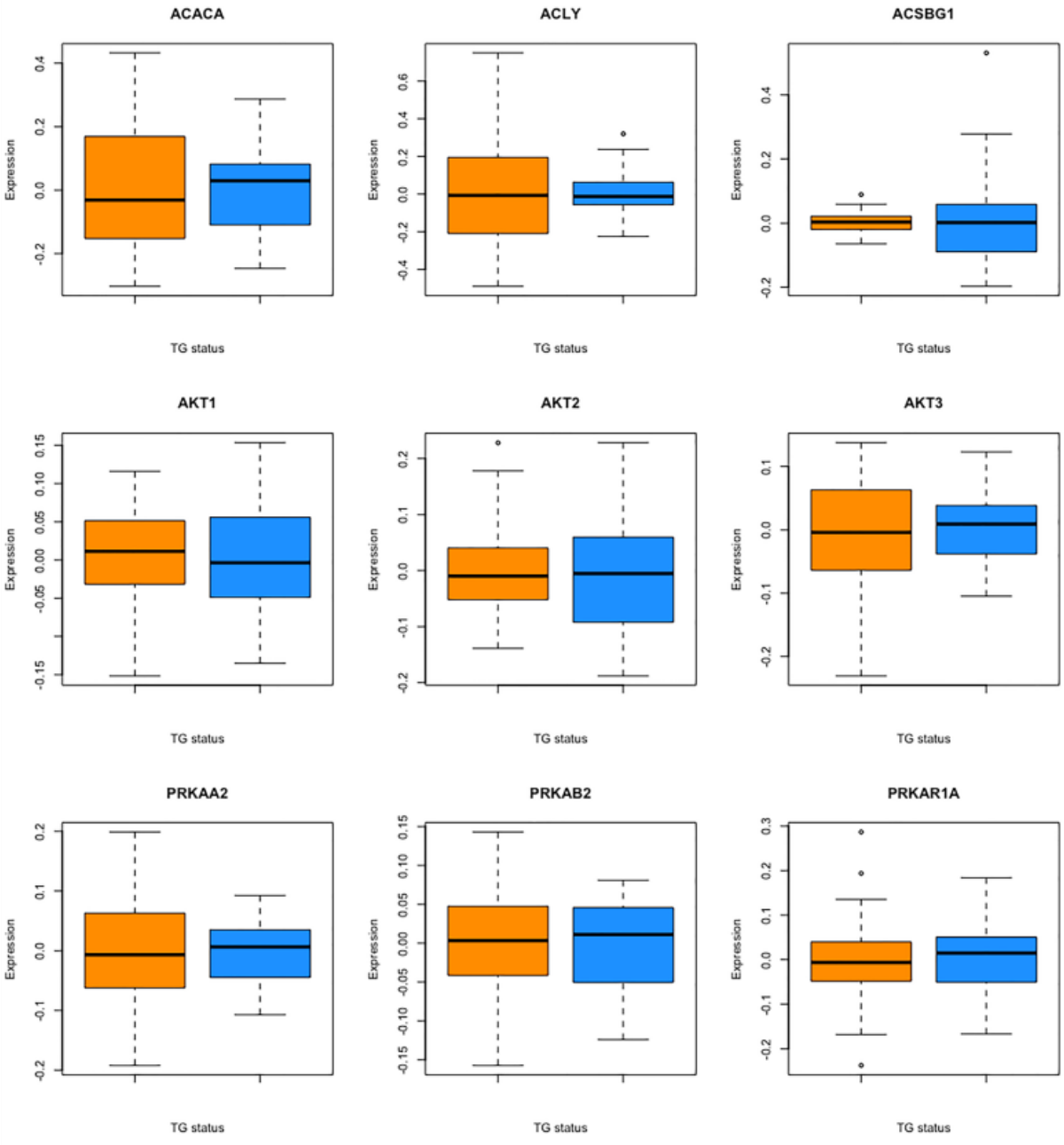
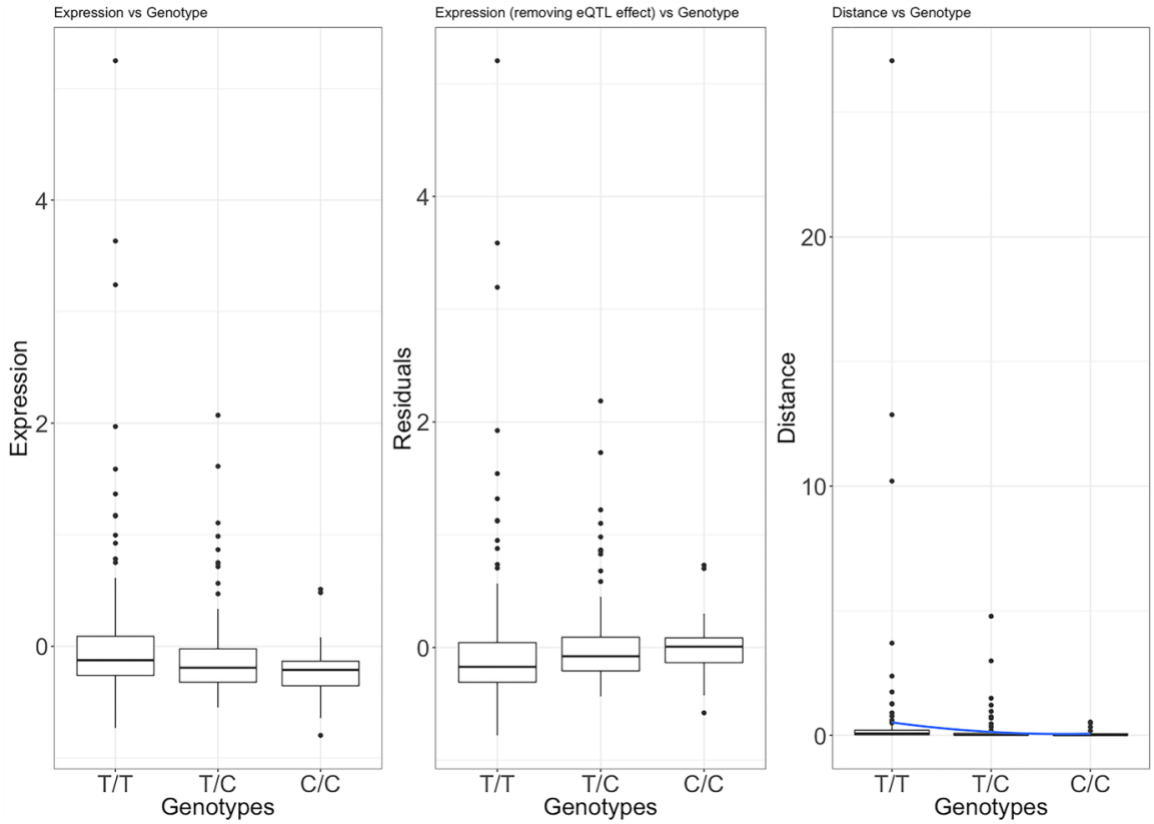


Figure 4-3 **Variant rs34272903 is a ve-QTL for the ACLY gene in the METSIM cohort**



Gene	RsID	Cor	p-value	p-value (Permuted)	FWER
ACLY	rs34272903	-0.25	4.2×10^{-6}	1.0×10^{-4}	1.0×10^{-4}

The first two columns indicate the gene ID and SNP ID. Then we report the Spearman correlation between the distance measure and genotype (Cor), p-value, permuted p-value (not controlled for multiple testing), and permuted p-value controlled for multiple testing across the SNPs (FWER).

REFERENCES

1. Plaisier, C. L. *et al.* Galanin Preproprotein Is Associated With Elevated Plasma Triglycerides. *Arterioscler. Thromb. Vasc. Biol.* **29,** 147–152 (2009).
2. Laakso, M. *et al.* METabolic Syndrome In Men (METSIM) Study: a resource for studies of metabolic and cardiovascular diseases. *J. Lipid Res.* jlr.O072629 (2017). doi:10.1194/jlr.O072629
3. Timpson, N. J., Greenwood, C. M. T., Soranzo, N., Lawson, D. J. & Richards, J. B. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nat. Rev. Genet.* **19,** 110–124 (2018).
4. Brown, A. A. veqtl-mapper: variance association mapping for molecular phenotypes. *Bioinformatics* **33,** 2772–2773 (2017).
5. Brown, A. A. *et al.* Genetic interactions affecting human gene expression identified by variance association mapping. *eLife* **3,** e01381 (2014).
6. Tronick, E. & Hunter, R. G. Waddington, Dynamic Systems, and Epigenetics. *Front. Behav. Neurosci.* **10,** 107 (2016).
7. Waddington, C. H. Canalization of Development and the Inheritance of Acquired Characters. *Nature* **150,** 563–565 (1942).
8. Wang, G. *et al.* Epistasis and decanalization shape gene expression variability in humans via distinct modes of action. *bioRxiv* 026393 (2015). doi:10.1101/026393
9. Berger, S. L. & Sassone-Corsi, P. Metabolic Signaling to Chromatin. *Cold Spring Harb. Perspect. Biol.* **8,** (2016).
10. Gibson, G. Decanalization and the origin of complex disease. *Nat. Rev. Genet.* **10,** 134–140 (2009).

11. Hulse, A. M. & Cai, J. J. Genetic variants contribute to gene expression variability in humans. *Genetics* **193,** 95–108 (2013).
12. Wang, G., Yang, E., Brinkmeyer-Langford, C. L. & Cai, J. J. Additive, epistatic, and environmental effects through the lens of expression variability QTL in a twin cohort. *Genetics* **196,** 413–425 (2014).
13. Gesta, S. & Kahn, C. White Adipose Tissue. in *Adipose Tissue Biology: Second Edition* 149–199 (2017). doi:10.1007/978-3-319-52031-5_5
14. Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinforma. Oxf. Engl.* **20,** 307–315 (2004).
15. Rodríguez, A. *et al.* Molecular characterization of the lipid GWAS signal on chromosome 18q11.2 implicates HNF4A-mediated regulation of the TMEM241 gene. *Arterioscler. Thromb. Vasc. Biol.* **36,** 1350–1355 (2016).
16. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).
17. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22,** 1760–1774 (2012).
18. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
19. Jun, G. *et al.* Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am. J. Hum. Genet.* **91,** 839–848 (2012).
20. Andrews, S. FastQC A Quality Control tool for High Throughput Sequence Data. *http://www.bioinformatics.babraham.ac.uk/projects/fastqc/* (2014).

21. Picard Tools - By Broad Institute. Available at: http://broadinstitute.github.io/picard/. (Accessed: 28th February 2018)

CHAPTER 5

CONCLUSIONS

Given the importance of understanding the genetic architecture of complex diseases, we aimed to dissect lipid GWAS loci and uncover additional sources of phenotypic heterogeneity underlying dyslipidemias. Taken together these studies explore several molecular mechanisms of gene regulation, including the effects of common variants on gene expression, transcription factor regulation, and metabolic signaling to chromatin. This latter represents a type of gene-environment interaction, first described as "decanalization" by Waddington[1] in 1942.

Since the early 2000s, GWAS has become the method of choice for the genetic analysis of complex traits. A majority of GWA variants are located in non-coding regions and likely located in regulatory elements such as transcription factor binding sites (TFBS), splice sites, and/or enhancer sites. Hence, fine-mapping studies, elucidating the associations between GWAS variants and regional genes using molecular phenotypes and *cis*-eQTL data are urgently needed. **Chapter 2, "*Molecular Characterization of the Lipid Genome-Wide Association Study Signal on Chromosome 18q11.2 Implicates HNF4A-Mediated Regulation of the TMEM241 Gene*",** describes our findings at the chromosome 18q11.2 GWAS TG locus[2]. First, we found nine variants in linkage disequilibrium ($r^2>0.7$) with the lead GWAS SNP, rs9949617[3]. We annotated all variants with biochemical evidence of molecular function using DNase I hypersensitive sites and transcription factor and chromatin states. We found rs17259126 as the top candidate variant for functional *in vitro* validation. Our luciferase assays provided evidence that the G allele exhibits a significantly lower effect on transcription ($P<0.05$) than the wild type allele, suggesting that the variant has regulatory potential. Gene expression regulations can occur through several mechanisms. Using electrophoretic mobility shift and ChIP-qPCR (chromatin immunoprecipitation coupled with quantitative polymerase chain reaction), we found that the minor G allele of rs17259126 disrupts a hepatocyte nuclear factor 4 α-binding site. Overall our *in*

vitro validation confirms that rs17259129 is a regulatory variant. However, these assays do not provide insights into the underlying gene. We performed variant-expression association analysis using *cis* expression quantitative trait locus (*cis*-eQTL) analysis and found that rs17259126 regulates the expression of the regional transmembrane protein 241 (TMEM241) gene in subcutaneous adipose RNAs from the Metabolic Syndrome In Men[4] (METSIM) cohort (p-value=$6.1x10^{-7}$-$5.8x10^{-4}$). We also replicated this result using the adipose microarray expression data from the British Multiple Tissue Human Expression Resource[5] (MuTHER; n=856) cohort.

A majority of susceptibility variants reside in non-coding regions and are likely to affect gene regulation via disruption of functional elements, such as TFBS. **In chapter 3, "*Genome-wide profiling of context-specific RORA binding sites provides a catalog of transcriptional targets associated with lipid metabolism in the human liver*",** we present preliminary results for the mapping of genome-wide RORA TFBS using chromatin immunoprecipitation followed by sequencing (ChIP-Seq). These results represent an initial attempt to map genome-wide targets of RORA in human liver cancer cell line, HepG2. RORA is an HDL-C GWAS gene in Mexicans[6] and a known regulator of the apolipoprotein genes, APOA5, APOA1, and APOC3.

Despite the successfully discovered lipid GWAS loci, it is clear that variation at these known GWAS loci explains only a small proportion of the trait heritability. Non-additive contributions to phenotypic variation in lipid traits are currently under-investigated in quantitative genetics. One example is the gene-environment interaction that may occur with toxic metabolic signaling to chromatin[7]. This process has the potential to uncover cryptic genetic variation. These kinds of gene-environment interactions can underlie the "missing heritability." **In chapter 4, "*ACLY may mediate decanalization of lipid metabolism pathways via histone acetylation*",** we investigate a specific type of the gene-environment interaction called "decanalization" and its

contribution to phenotypic variance in TG metabolism. We found that individuals with low TGs displayed a greater ATP citrate lyase (ACLY) expression variance than the individuals with high TGs (p-value= $2.0x10^{-2}$). This result was replicated in the METSIM[4] adipose RNA-Seq cohort (n=335) (p-value=$1.8x10^{-3}$). ACLY encodes the primary enzyme responsible for the synthesis of nuclear and cytosolic acetyl-CoA, which is important for the biosynthesis of fatty acids, a precursor of TGs. Using a correlation least squared (CLS) test[8] to uncover SNPs associated with ACLY expression variance, we found that the reference allele of variant rs34272903 (T/C) is associated with increased ACLY expression variance (FWER p-value=$1.0x10^{-4}$). However, we were not able to replicate this result in subcutaneous adipose tissue RNA-Seq data of the GTEx cohort[9], likely due to the genetic heterogeneity and confounding caused by the postmortem nature of this cohort.

CONCLUDING THOUGHTS

The discovery of DNA as the hereditary material was published by Avery[10] in 1944, four years after Waddington published[11] his theory of the "epigenetic landscape" and two years before he introduced the term "epigenetics" in 1942[1]. Ahead of his time, Waddington's theory of decanalization was a mechanistic hypothesis of a gene-environment (GXE) interaction.

Epigenetic mechanisms of decanalized phenotypes

Molecular epigenetics is believed to mediate the interplay between biological systems and their environment[7,12,13]. Environmentally derived metabolites can act as signaling molecules to chromatin, which may lead to chromatin remodeling and gene expression changes[7]. Under normal structural canalization, molecular epigenetics provides an organism the ability to make "quick" adaptations to the changing environments. Environmental toxicity may, however, induce initial epigenetic changes, which can also change the system's engagement with the environment[12]. This

new relationship may induce further epigenetic changes, which can lead to endogenous changes at multiple levels[12–14]. Systems biology approaches to understand complex disease should better account for these types of complex relationships that likely exist between organisms and their environment.

The connection between metabolism and epigenetically controlled gene expression

Our finding that ACLY exhibits context-specific expression variance may be an example of a metabolic signaling to chromatin[7,15,16]. ACLY mediates the conversion of citrate into nuclear acetyl-CoA, a necessary substrate for histone acetyltransferases. Changes in acetyl-CoA production are known to affect histone acetylation and gene expression[7]. ACLY, in this case, may act as a direct regulator of gene expression with or without the intervention of a local genomic variant. By modulating the levels of histone acetylation, ACLY may modulate the expression of entire networks of lipogenic enzymes (including its own expression), leading to a canalized/decanalized network and phenotypic variability (Figure 5-1). The regulatory action of ACLY on the chromatin state can help elucidate cryptic genetic variation, which may act synergistically on the phenotype. Our results show that both in the Mexican[17] and the METSIM[4] cohorts, individuals who have low serum TG levels exhibit higher variance in ACLY expression when compared to the individuals with high TG levels. This increased variability in ACLY expression may contribute to a spectrum of responses necessary for buffering environmental stressors[12–14], a robust equilibrium[18]. Similar conclusions have been reported for human neurological diseases[19,20]. Since the early 2000s, GWAS has become the method of choice for the analysis complex traits, but the effect of toxic metabolic signaling to chromatin and the effect it may have in uncovering cryptic genetic variation are under-investigated and may in part underlie the "missing heritability." Future studies can aim to identify specific signatures of acetylation in

patients with low and high serum TG levels. Characterization of ACLY acetylation signatures may thus help identify regulatory pathways and genes involved in TG metabolism.

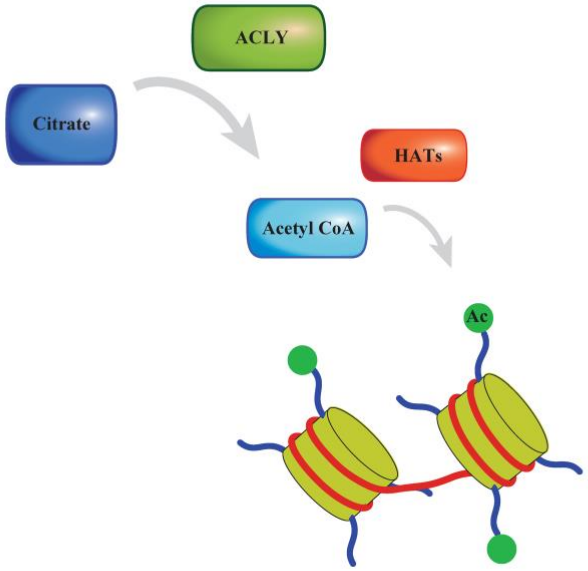


Figure 5-1 **ACLY mediates the conversion of citrate into nuclear acetyl-CoA, a cofactor for chromatin remodeling enzymes.**

FIGURE LEGENDS

Figure 5-1 **ACLY mediates the conversion of citrate into nuclear acetyl-CoA, a cofactor for chromatin remodeling enzymes.**

REFERENCES

1. Waddington, C. H. Canalization of Development and the Inheritance of Acquired Characters. *Nature* **150,** 563–565 (1942).
2. Weissglas-Volkov, D. *et al.* Investigation of variants identified in caucasian genome-wide association studies for plasma high-density lipoprotein cholesterol and triglycerides levels in Mexican dyslipidemic study samples. *Circ. Cardiovasc. Genet.* **3,** 31–38 (2010).
3. Rodríguez, A. *et al.* Molecular characterization of the lipid GWAS signal on chromosome 18q11.2 implicates HNF4A-mediated regulation of the TMEM241 gene. *Arterioscler. Thromb. Vasc. Biol.* **36,** 1350–1355 (2016).
4. Laakso, M. *et al.* METabolic Syndrome In Men (METSIM) Study: a resource for studies of metabolic and cardiovascular diseases. *J. Lipid Res.* jlr.O072629 (2017). doi:10.1194/jlr.O072629
5. Grundberg, E. *et al.* Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat. Genet.* **44,** 1084–1089 (2012).
6. Ko, A. *et al.* Amerindian-specific regions under positive selection harbour new lipid variants in Latinos. *Nat. Commun.* **5,** 3983 (2014).
7. Berger, S. L. & Sassone-Corsi, P. Metabolic Signaling to Chromatin. *Cold Spring Harb. Perspect. Biol.* **8,** (2016).
8. Brown, A. A. veqtl-mapper: variance association mapping for molecular phenotypes. *Bioinformatics* **33,** 2772–2773 (2017).
9. Ferreira, P. G. *et al.* The effects of death and post-mortem cold ischemia on human tissue transcriptomes. *Nat. Commun.* **9,** 490 (2018).

10. Avery, O. T., MacLeod, C. M. & McCarty, M. STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES. *J. Exp. Med.* **79,** 137–158 (1944).
11. Organizers and Genes. *Ann. Entomol. Soc. Am.* **34,** 517–517 (1941).
12. Tronick, E. & Hunter, R. G. Waddington, Dynamic Systems, and Epigenetics. *Front. Behav. Neurosci.* **10,** 107 (2016).
13. Baquero, F. Epigenetics, epistasis and epidemics. *Evol. Med. Public Health* **2013,** 86–88 (2013).
14. Sivanand, S. *et al.* Nuclear Acetyl-CoA Production by ACLY Promotes Homologous Recombination. *Mol. Cell* **67,** 252–265.e6 (2017).
15. Kirchner, H. *et al.* Altered DNA methylation of glycolytic and lipogenic genes in liver from obese and type 2 diabetic patients. *Mol. Metab.* **5,** 171–183 (2016).
16. Muka, T. *et al.* The role of global and regional DNA methylation and histone modifications in glycemic traits and type 2 diabetes: A systematic review. *Nutr. Metab. Cardiovasc. Dis.* **26,** 553–566 (2016).
17. Plaisier, C. L. *et al.* Galanin Preproprotein Is Associated With Elevated Plasma Triglycerides. *Arterioscler. Thromb. Vasc. Biol.* **29,** 147–152 (2009).
18. Gibson, G. Decanalization and the origin of complex disease. *Nat. Rev. Genet.* **10,** 134–140 (2009).
19. Burrows, E. L. & Hannan, A. J. Decanalization mediating gene-environment interactions in schizophrenia and other psychiatric disorders with neurodevelopmental etiology. *Front. Behav. Neurosci.* **7,** 157 (2013).

20. Mar, J. C. *et al.* Variance of gene expression identifies altered network constraints in neurological disease. *PLoS Genet.* **7,** e1002207 (2011).