INVESTIGATION OF THE GENETIC ARCHITECTURE OF CARDIOMETABOLIC DISEASE IN MEXICAN AMERICANS

By

Chuan Gao

A Dissertation Submitted to the Graduate Faculty of

WAKE FOREST UNIVERSITY SCHOOL OF ARTS AND SCIENCES

in Partial Fulfillments of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

Molecular Genetics and Genomics

May 2017

Winston Salem, North Carolina

Approved By:

Nicholette D. Allred, Ph.D., Advisor

Carl D. Langefeld, Ph.D., Chair

Donald W. Bowden., Ph.D.

John S. Parks, Ph.D.

Yongmei Liu, Ph.D.

ACKNOWLEDGEMENTS

Foremost, I would like to take this opportunity to express my sincere thanks to my advisor, Dr. Nicholette Allred. It has been my great fortune to work with her for the last four years. She introduced me to the fascinating genetics research field and guided me to become not only a young scientist, but also a thoughtful person. She encouraged me when I encountered failures, she inspired me when I faced obstacles, and she supported me when I pursued opportunities. There is no way that I could have become who I am today without her help!

Taking this opportunity, I would like to express gratitude to my committee members. Dr. Bowden, thank you so much for your valuable comments and suggestions. Your experience and visions are invaluable to my scientific research and future career. Dr. Langefeld, I want to thank you again for your help with the statistics in my thesis. You are such a good mentor and nice friend! Dr. Parks, as my first rotation mentor, your patience and suggestions have helped me transit from undergraduate to graduate student. My family and I can never thank you enough for what you have done to us! Dr. Liu, thank you for your encouragement, valuable comments, and thoughtful questions that led this thesis work to where it is today.

My great gratitude also goes to other faculty members at Wake Forest whose expertise and assistance helped me to make progress. Fang-Chi, I want to thank you for walking me through even the most basic statistical conceptions. You are such a great mentor and a trustworthy colleague. Dr. Ng (Maggie), thank you so much for your input and thoughtful questions during lab meetings and seminars. Dr. Howard, thanks for directing our molecular genetics program and creating this collaborative research environment. Dr. Hawkins, thanks for your help with my research and suggestions regarding my post graduate training. My sincere thanks also go to Linda and Debbie, thank you all for your assistance with our Genomics Center.

I also would like to thank current and previous members of the Bowden family. Pam, thank you for managing our lab and maintaining high-quality DNA. It really has been the foundation for all of my computational research. Becky, I cherish the every moment and conversation we had in lab. I feel very fortunate to have you as my best friend! Keri, I want to thank you for your help with my research and career development. You are such a nice and patient person that is not bothered by even the dumbest questions. Dr. Okut, thank you for spending so much time and walking through statistics formulas. MJ, my dear friend and roommate, thank you for accompanying me for the days and nights in the last five years. Jackie and Jake, thank you all so much for your encouragement and help with my research projects. JJ, I can't move my research forward without our servers. Thank you very much for keeping our electric brains healthy and happy. Lucho, thank you for being such a good colleague. You and your ping-pong table surely made my life productive and active. Laura, Poorva, Jeremy, and Mary, the time we spent together witnessed my growth in the lab and as an independent scientist all because of your kind-hearted care and support along the way.

I would like to express my great thanks to my family. To my parents, who love and support me with all their hearts and souls; to my grandparents, who taught me the importance of faith and altitude; to my uncles and aunts, who shared my joys and tears; to Xin, my loving wife, who is the most valuable treasure I received from graduate school. Thank you for being a good partner in research and a close family member.

Last but not least, great thanks go to Wake Forest University. Thank you for giving me the opportunity to study, to research, and to grow. Thank you for the financial and academic support that made me who I am today. Both my wife and I feel proud to be Wake Forest alumni.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

LIST OF ABBREVIATIONS

1000G	1000 Genomes Project
ADP	Adiponectin
AIR	Acute insulin response
APOB	apolipoprotein B
BMI	Body Mass Index
CHD	Coronary Heart Disease
CMS	Cardiometabolic Syndrome
CT	Computed Tomography
CV	Cross Validation
CVD	Cardiovascular Disease
DEXA	Dual-energy X-ray Absorptiometry
D_I	Disposition Index
FINS	Fasting Plasma Insulin
GFAST	Fasting Plasma Glucose
GLGC	Global Lipid Genetics Consortium
GUARDIAN	The Genetics Underlying Diabetes in Hispanics Consortium
GWAS	Genome-wide Association Study
HDL-c	High-density Lipoprotein Cholesterol
HGP	Human Genome Project
HOMA	Homeostatic Model Assessment
IBD	Identity-by-descent
INDEL	Insertion/Deletion
IRASFS	The Insulin Resistance Atherosclerosis Family Study
LDL-c	Low-density Lipoprotein Cholesterol

LOD	Logarithm of The Odds
MAF	Minor Allele Frequency
MCRI	Metabolic Clearance Rate of Insulin
NCEP	National Cholesterol Education Program
NGS	Next Generation Sequencing
PBF	Percent Body Fat
PCR	Polymerase Chain Reaction
pLOF	Predicted Loss of Function
QC	Quality Control
RNAi	RNA Interference
SAT	Subcutaneous Adipose Tissue
S_G	Glucose Effectiveness
S_I	Insulin Sensitivity
SNP	Single Nucleotide Polymorphism
SOLAR	Sequential Oligogenic Linkage Analysis Routines
T2D	Type II Diabetes
TC	Total Cholesterol
TG	Triglycerides
VAT	Visceral Adipose Tissue
VSR	Visceral-subcutaneous Ratio
WAIST	Waist Circumference
WGS	Whole Genome Sequencing
WHR	Waist-hip ratio

ABSTRACT

Gao, Chuan

INVESTIGATION OF THE GENETIC ARCHITECTURE OF CARDIOMETABOLIC DISEASE

IN MEXICAN AMERICANS

Dissertation under the direction of

Nicholette D. Allred, Ph.D., Associate Professor of Biochemistry

Cardiometabolic syndrome (CMS) is a clustering of interrelated risk factors (central obesity, hyperglycemia, hypertension, insulin resistance, and dyslipoproteinemia) that promotes the development of atherosclerotic vascular disease and type 2 diabetes. It was recognized as a disease entity by the American Society of Endocrinology, National Cholesterol Education Program (NCEP), and World Health Organization in 2003 (Castro, El-Atat, McFarlane, Aneja, & Sowers, 2003). Previous genetic studies with monozygotic and dizygotic twin pairs suggested a strongly heritable pattern for CMS risk factors, i.e. it is estimated that circulating lipid heritability ranges from 0.58 to 0.66 (h^2_{HDL}=0.61, h^2_{LDL}=0.59, h^2_{TC}=0.58, h^2_{TG}=0.66) (Knoblauch et al., 1997). However, the exact genetic mechanism of the disease is poorly understood.

In 2003, scientists from the Human Genome Project (HGP) completed sequencing of the first human genome and mapped genes from both a physical and functional standpoint. Within a few years, multiple novel techniques and platforms were developed and significantly reduced the cost of genotyping and sequencing, enabling researchers to investigate the relationship between human DNA variations and disease phenotypes at an affordable cost.

In this dissertation, we hypothesized that naturally occurring DNA variations are responsible for the heritability of CMS in Mexican Americans. To test this hypothesis, multiple genotyping and sequencing platforms (GWAS, Exome Chip, and Exome Sequencing) were assessed in individuals from the Insulin Resistance Atherosclerosis Family Study (IRASFS). In addition, CMS related phenotypes including obesity, type 2 diabetes (T2D), blood pressure, and plasma lipid levels were measured at clinical centers. To investigate genotype-phenotype associations, we performed association and linkage analyses. Specifically, four scientific questions were addressed: 1. What is the genetic basis of obesity in Mexican Americans? Can genetic signals explain the increasing prevalence of the epidemic? (Chapter II) 2. What is the genetic architecture of adipose tissue distribution? Can genetic variations explain the gender-specific pattern of fat distribution? (Chapter III) 3. What is the genetic component of plasma lipid levels in Mexican Americans? How can exome sequencing add to current knowledge of lipid metabolism? (Chapter IV) 4. What is the role of insertion and deletion (INDEL) variants in regulating cardiometabolic phenotypes? Can INDELs be imputed with confidence? (Chapter V)

In summary, genetic association and linkage analyses were performed for cardiometabolic phenotypes in Mexican Americans from the IRASFS. Multiple genetic variants were identified with significant signals to relevant phenotypes. These results revealed genetic mechanisms of the disease, added evidence to the heritability, and suggested directions for future functional studies.

Chapter I

INTRODUCTION

Chuan Gao prepared this chapter. Dr. Nicholette Allred acted in an advisory and editorial capacity.

Cardiometabolic Syndrome

Cardiometabolic syndrome (CMS) is a constellation of maladaptive cardiovascular, renal, metabolic, prothrombotic, and inflammatory abnormalities. These interrelated factors contribute to atherogenic dyslipidemia, elevated blood pressure, insulin resistance, and proinflammatory and prothrombic states. In 2003, CMS was recognized as a disease entity by the American Society of Endocrinology, National Cholesterol Education Program (NCEP), and World Health Organization (Castro, et al., 2003).

Obesity

Obesity is one of the major risk factors for CMS (Cannon, 2007). It is a growing epidemic affecting 16.9% of individuals 2-19 years old and 34.9% of adults in the United States (Ogden, Carroll, Kit, & Flegal, 2014). It is estimated that the annual cost attributable to obesity is $190.2 billion dollars. Among this expense, about half is attributed to direct medical costs, representing 21% of the total annual medical spending in the United States (Cawley & Meyerhoefer, 2012; Wolf & Colditz, 1998). Previous research has shown obesity to be strongly associated with an increased risk for multiple metabolic diseases including cardiovascular disease (CVD), type 2 diabetes, metabolic syndrome, certain types of cancer, stroke, and hypertension (Frayling et al., 2007; Haslam & James, 2005; Narkiewicz, 2005). Recently, studies have shown extreme obesity (Body Mass Index (BMI) of 40-59 kg/m^2) may shorten life expectancy up to 14 years (Kitahara et al., 2014). According to the new guidelines released by the American Heart Association, American College of Cardiology, and The Obesity Society in November 2013, obesity is considered as a disease and doctors should provide more active treatments for obese patients (Jensen et al., 2014).

Obesity is a disease condition with multiple risk factors including genetics, lifestyle, diet, and additional health conditions (Heggeness, Esses, & Kostuik, 1991; Keith et al., 2006). Among these, genetic factors have been shown as the strongest risk factors for obesity with an explanation over 70% of variation in BMI (Willyard, 2014). Genetic research of obesity has been successful. To date, more than 100 loci have been identified with significant associations with obesity and obesity-related phenotypes (Berndt et al., 2013; Fox, White, et al., 2012; Heid et al., 2010; Hindorff LA; Kilpelainen et al., 2011; Liu et al., 2013; Monda et al., 2013; Randall et al., 2013; Speliotes et al., 2010).

As with many complex diseases, the prevalence of obesity varies by ethnicity, with Mexican Americans having a higher estimated prevalence than non-Hispanic whites, i.e. 40.4% compared to 34.3%, respectively (Flegal, Carroll, Kit, & Ogden, 2012). Despite the higher risk of obesity in the Mexican population compared to non-Hispanic whites, most obesity genetics studies have been conducted in European populations, i.e. studies of the genetic contributors have been few in number and limited in scope in the Mexican population. Until now, VIVA LA FAMILIA was the only cohort with published genome-wide significant obesity signals specific to the Mexican population (Comuzzie et al., 2012).

Recent research has suggested that obesity is not a homogeneous condition and that regional fat distribution affects glucose and lipid metabolism beyond total body adiposity (Wajchenberg, 2000). For example, visceral adipose tissue (VAT) has been shown to be responsible for the increased mortality and risk for metabolic disorders while subcutaneous adipose tissue (SAT) is thought to be benign (P. Cohen et al., 2014; Lapidus et al., 1984; Larsson et al., 1984; Wajchenberg, 2000). Anthropometric measures, e.g. waist circumference (WAIST) and waist-hip ratio (WHR), have been widely used to measure adipose tissue distribution. However, these measures are usually biased by age and skeletal structure (Taylor, Jones, Williams, & Goulding,

2000). Currently, computed-tomography (CT) is considered as the gold standard for the measurement of adipose tissue deposition (Blaak, 2001; Norris et al., 2005). However, due to cost and accessibility reasons, limited studies have been performed to date.

Cardiovascular disease

Cardiovascular disease (CVD), mainly coronary heart disease (CHD) and stroke, is the leading cause of death in the United States (Moore & Tabas, 2011). According to the CDC about 610,000 people die of heart disease in the United States every year and more than 33% of American adults have one or more types of CVD (CDC, 2017). It is estimated that by 2030, 41% of the US population will have some form of CVD (Benjamin et al., 2017). While the exact mechanism of disease remains unclear, lipid concentrations are well-accepted as major risk factors as well as clinical indicators for CVD, e.g. total cholesterol (TC), low density lipoprotein cholesterol (LDL), high density lipoprotein cholesterol (HDL), and triglycerides (TG).

Genetic studies have suggested a strong heritability for circulating lipid levels, e.g. TC, LDL, HDL, and TG. Based on a European twin-pairs study, it was estimated that circulating lipid heritability ranges from 0.58 to 0.66 (h^2_{HDL}=0.61, h^2_{LDL}=0.59, h^2_{TC}=0.58, h^2_{TG}=0.66) (Knoblauch, et al., 1997). Given the public health relevance as well as the strong genetic component, numerous GWAS have been performed to investigate the genetic architecture of circulating lipid levels. The most recent Global Lipid Genetics Consortium (GLGC) analyzed 188,577 individuals from four ethnicities (Europeans, East Asians, South Asians, and African Americans) and identified 157 loci associated with plasma lipid traits (Willer et al., 2013).

As with many complex diseases, the prevalence of CVD varies by ethnicity, i.e. Mexican Americans have been observed to have higher risk compared to non-Hispanic whites (32.1% versus 23.8%) (CDC, 2013; Go et al., 2013). Until now, the largest lipid GWAS in Mexicans was

performed by Below et al. in 2015 with 4,383 individuals of Mexican ancestry. However, all genome-wide significant regions identified in the Mexican-based meta-analysis have been previously associated with lipid traits (Below et al., 2016; Surakka et al., 2015; Willer et al., 2008; Willer, et al., 2013).

Genome-wide association studies (GWAS)

An important goal for genetics research is to identify genetic polymorphisms that are risk factors of complex disease or disease related phenotypes. Traditional biological approaches offer a potentially powerful screening yet only a small number of genes can be studied at a time(Hirschhorn & Daly, 2005). A genome-wide association study (GWAS) is an examination of a genome-wide set of genetic variants and detect variants at genomic loci that are associated with complex traits in the population, i.e. GWAS searches single nucleotide polymorphisms (SNPs) across the genome for increased frequency in people with a disease or disease related phenotype. Compared to hypothesized candidate gene studies, GWAS comprehensively screen all variants and have the potential to identify novel mechanisms.

The first successful GWAS was published in 2005 with the identification of two SNPs with associations with macular degeneration (Klein et al., 2005). To date, thousands of genetic loci have been identified with significant association signals and were reliably replicated in hundreds of independent studies (Berndt, et al., 2013; Fox, White, et al., 2012; Heid, et al., 2010; Hindorff LA; Kilpelainen, et al., 2011; Liu, et al., 2013; Monda, et al., 2013; Randall, et al., 2013; Speliotes, et al., 2010; Willer, et al., 2013). Based on these findings, numerous novel disease mechanisms, signaling pathways, and drug targets were identified. One famous example is the identification of the transcription factor 7-like 2 (TCF7L2) gene, which is by far the most significant genetic

marker associated with Type 2 Diabetes (T2D) and has been validated in multiple independent studies (Grant et al., 2006).

Despite the large number of signals identified, GWAS have also been criticized with major limitations. For example, among the variants discovered by GWAS, over 80% fall outside protein coding regions with no biological relevance to disease and no clinical utility has been established (Manolio et al., 2009). Also, despite the large numbers of significant signals that predispose to or protect against certain disease conditions, the majority of the genetic heritability is still unexplained, i.e. combining the genome-wide significant BMI signals identified to date, only 2.7% of the phenotypic variance can be explained, providing very limited information for disease risk prediction (Choquet & Meyre, 2011; Locke et al., 2015).

Another limitation of GWAS is that it primarily focuses on common SNPs, which was supported by the "common disease, common variant" hypothesis, i.e. common complex diseases are caused by a large number of common variants and each of the variants contributes to a small proportion of the phenotypic variance (Gibson, 2011; Manolio, et al., 2009). Recent studies using next generation sequencing technologies have identified large numbers of rare variants with predicted loss of function (pLOF) effects. For example, based on an exome sequencing study with more than 50 thousand individuals, it is estimated that each individual had a median of 21 rare pLOF variants and more than 92% of the genes have one or more heterozygous loss of function carriers (Dewey et al., 2016). Therefore, rare variants, particularly functional rare variants, could also contribute to common disease susceptibility.

Traditional GWAS have extensive coverage on SNPs. However, SNPs are only one class of genetic variation and insertions, deletions, and structural variations also contribute to a significant proportion of genome sequence variations yet they have been rarely evaluated in the GWAS

setting (Weischenfeldt, Symmons, Spitz, & Korbel, 2013). Insertions and deletions (INDELs) are defined as the addition and loss, respectively, of one or more nucleotide base pairs of a DNA sequence. It is estimated that INDELs, especially small INDELs (with 16 bases or less), are abundant in human genomes and constitute 21% of the total variants identified (Mullaney, Mills, Pittard, & Devine, 2010). Similar to SNPs, INDELs also play an important role in disease susceptibility. Genetic studies have confirmed associations between INDELs and multiple disease and non-disease related phenotypic variations including dietary starch consumption, autism, schizophrenia, Crohn's disease, rheumatoid arthritis, type 1 diabetes, and obesity (Cantsilieris & White, 2013; Craddock et al., 2010; Jacquemont et al., 2011; Malhotra & Sebat, 2012; Perry et al., 2007; Pinto et al., 2014). However, compared to SNP based association analyses, very few studies have included INDELs. Traditional discovery and genotyping approaches for INDELs largely rely on polymerase chain reaction (PCR) and several other modified PCR-based techniques (Dhawan & Padh, 2009). These techniques are either expensive or inconvenient and therefore not ideal for large-scale studies (Almal & Padh, 2012). Short-read DNA sequencing data from the 1000 Genomes Project (1000G) has enabled a better constructed set of INDELs across different ethnicities with enhanced size and breakpoint resolution (Mills et al., 2011; Sudmant et al., 2010). The release of the 1000G reference panel has enabled researchers, for the first time, to accurately impute large numbers of INDELs from array-based genotypes (Abecasis et al., 2012).

In addition to GWAS, linkage analysis has also been proposed as a powerful tool to detect the chromosomal location of disease genes. Different from association analysis, which focuses on the phenotypic distribution in the population, linkage analysis detects the co-segregation of genetic markers and disease phenotypes across multiple generations within families. The metric used in linkage analysis is a logarithm of the odds (LOD score), e.g., a LOD score of three means

the odds are a thousand to one in favor of genetic linkage. Traditionally, a LOD score of three or more is considered as a significant linkage signal. Microsatellites have been proposed as excellent genetic markers for linkage analysis, known as the multipoint linkage analysis, due to their heterozygosity. i.e. they are highly mutable markers often with 15 or more alleles in any given population. A microsatellite is a repetitive DNA sequence in which certain DNA bases (ranging in length from 2–5 base pairs) are repeated, typically 5–50 times (Van Oosterhout, Hutchinson, Wills, & Shipley, 2004). With the development of genotyping and sequencing techniques, researchers are able to perform two-point linkage analysis using individual genetic markers, e.g. SNPs (Browning et al., 2004). Compared to SNPs, microsatellites are more polymorphic and therefore a relatively small number genotyping is sufficient to capture a great deal of the genomic variation (recombination events) (Bahram & Inoko, 2007). On the other hand, with genotyping arrays and next generation sequencing techniques, SNPs are much easier to genotype in large numbers, giving two-point approach a higher resolution.

Linkage has been shown to be very successful towards Mendelian disorders, i.e. diseases that are caused by single gene mutations (Botstein & Risch, 2003). However, when linkage analysis was applied for common complex disease, the limitations start to emerge. For example, linkage analysis lacks the sensitivity for variants with small effects, i.e. linkage signals are largely dampened when a carrier fails to exhibit the phenotype, which is known as the incomplete penetrance (Wiltshire, Morris, McCarthy, & Cardon, 2005). In addition, the power of linkage analysis is dependent on the number of segregation events. Therefore, linkage analysis can only be applied to family studies with complete pedigree structures.

Hypotheses and Objectives

This thesis aims to explore the hypotheses that genetic markers (SNPs and INDELs) are responsible for the susceptibility of CMS. To test the hypotheses, genetic association and linkage analyses were performed with CMS risk factors (obesity, plasma lipid levels). To cover a broad allelic spectrum of genetic markers, multiple genotyping and sequencing platforms were used and compared; novel machine based measures were evaluated as a comprehensive and direct measures of disease related phenotypes. The results have the potential to establish novel models for genetic association analysis and provide directions for future mechanistic research of CMS.

Chapter II

A Comprehensive Analysis of Common and Rare Variants to Identify Adiposity Loci in Hispanic Americans: The IRAS Family Study (IRASFS)

Chuan Gao, Nan Wang, Xiuqing Guo, Julie T. Ziegler, Kent D. Taylor, Anny H. Xiang, Yang Hai, Steven Kridel, Jerry L. Nadler, Fouad Kandeel, Leslie J. Raffel, Yii-Der I. Chen, Jill M. Norris, Jerome I. Rotter, Richard M. Watanabe, Lynne E. Wagenknecht, Donald W. Bowden, Elizabeth K. Speliotes, Mark O. Goodarzi, Carl D. Langefeld, Nicholette D. Palmer

The following manuscript was published on PlosOne. Stylistic variations are due to the requirements of the journal. Chuan Gao performed the analyses and prepared the manuscript. Dr. Nicholette Allred acted in an advisory and editorial capacity.

Abstract

Obesity is growing epidemic affecting 35% of adults in the United States. Previous genome-wide association studies (GWAS) have identified numerous loci associated with obesity. However, the majority of studies have been completed in Caucasians focusing on total body measures of adiposity. Here we report the results from genome-wide and exome chip association studies focusing on total body measures of adiposity including body mass index (BMI), percent body fat (PBF) and measures of fat deposition including waist circumference (WAIST), waist-hip ratio (WHR), subcutaneous adipose tissue (SAT), and visceral adipose tissue (VAT) in Hispanic Americans (n_{max}=1263) from the Insulin Resistance Atherosclerosis Family Study (IRASFS). Five SNPs from two novel loci attained genome-wide significance ($P<5.00x10^{-8}$) in IRASFS. A missense SNP in the isocitrate dehydrogenase 1 gene (*IDH1)* was associated with WAIST (rs34218846, MAF=6.8%, $P_{DOM}=1.62x10^{-8}$). This protein is postulated to play an important role in fat and cholesterol biosynthesis as demonstrated in cell and knock-out animal models. Four correlated intronic SNPs in the Zinc finger, GRF-type containing 1 gene (*ZGRF1*; SNP rs1471880, MAF=48.1%, $P_{DOM}=1.00x10^{-8}$) were strongly associated with WHR. The exact biological function of *ZGRF1* and the connection with adiposity remains unclear. SNPs with p-values less than $5.00x10^{-6}$ from IRASFS were selected for replication. Meta-analysis was computed across seven independent Hispanic-American cohorts (n_{max}=4156) and the strongest signal was rs1471880 ($P_{DOM}=8.38x10^{-6}$) in *ZGRF1* with WAIST. In conclusion, a genome-wide and exome chip association study was conducted that identified two novel loci (*IDH1 and ZGRF1*) associated with adiposity. While replication efforts were inconclusive, when taken together with the known biology, *IDH1* and *ZGRF1* warrant further evaluation.

Keywords: Genome-Wide Association Study; exome chip; Hispanic Americans; obesity

Introduction

Obesity is a global health problem closely associated with an increased risk for multiple metabolic diseases (Frayling, et al., 2007; Haslam & James, 2005; Narkiewicz, 2005). Body mass index (BMI) has been widely used in studies to estimate total body adiposity. However, BMI is derived from total body weight which possesses inter-individual variability attributed to muscle mass, i.e. BMI is not a direct measure of fat deposition, which is closely linked to health outcomes. Waist-hip ratio (WHR) and waist circumference (WAIST) have been well-recognized as complementary approaches to estimate fat deposition. However, they are often skewed by age and skeletal structure (Taylor, et al., 2000). In addition to anthropometric measures, computed tomography (CT) has been recognized as the gold standard for measuring regional fat deposition (Blaak, 2001). Visceral adipose tissue (VAT) and subcutaneous adipose tissue (SAT) can be estimated by CT scans with both being strong risk factors for metabolic disturbances (Lafontan & Berlan, 2003; Montague & O'Rahilly, 2000; Wagenknecht et al., 2003). Alternatively, dual-energy X-ray absorptiometry (DEXA) can provide a direct measurement of total body fat volume (Going et al., 2011) by partitioning total body mass into bone, lean, and fat soft tissue components.

Genome-wide association studies (GWAS) have been successful in identifying obesity-related loci with more than 100 loci identified to date (Berndt, et al., 2013; Fox, White, et al., 2012; Heid, et al., 2010; Hindorff LA; Kilpelainen, et al., 2011; Liu, et al., 2013; Monda, et al., 2013; Randall, et al., 2013; Speliotes, et al., 2010). However, over 80% of GWAS variants fall outside protein coding regions, which impairs causal inference (Manolio, et al., 2009). In addition, associated variants possess small effect sizes providing limited information for disease risk prediction (Choquet & Meyre, 2011). More recent evidence suggests low frequency and rare variants (minor allele frequency (MAF) <5%) also play a role in susceptibility to disease (Gibson, 2011). In

addition, although the overall risk of obesity is much higher in Hispanic populations compared to non-Hispanic whites, i.e. 40.4% versus 34.3%, respectively (Flegal, et al., 2012), studies of the genetic contributors have been few in number and limited in scope in the Hispanic population. Until now, VIVA LA FAMILIA was the only cohort with published genome-wide significant obesity signals specific to the Hispanic population (Comuzzie, et al., 2012).

In this study, we hypothesized that genetic factors are responsible for the increased obese status in the Hispanic population. By combining more refined adiposity measures and genotypic information from GWAS and exome chip, we are able to conduct a comprehensive scan of the genome with the potential to identify ethnic specific causal variants.

Materials and Methods:

Ethics Statement

Participants included in this study were recruited from clinical centers in San Antonio, TX and San Luis Valley, CO. The Institutional Review Board of each clinical (UT Health Science Center San Antonio Review Board and Colorado Multiple Institutional Review Board, respectively) and analysis (Wake Forest School of Medicine) site approved the study protocol and all participants provided their written informed consent.

Study Participants

Study design and recruitment for the Insulin Resistance Atherosclerosis Family Study (IRASFS) have been described (Henkin et al., 2003). Briefly, the IRASFS was designed to identify the genetic and environmental basis of insulin resistance and adiposity. Hispanic Americans included in this report (n=1417 individuals, 90 pedigrees) were recruited from clinical centers in San

Antonio, TX and San Luis Valley, CO. While a diagnosis of diabetes was not required for participation, about 12.7% of genotyped individuals had diabetes. A detailed description of the phenotypes can be found in supplemental materials.

Genotyping and Quality Control

GWAS genotyping was supported through the Genetics Underlying Diabetes in Hispanics (GUARDIAN) Consortium. Genotyping was attempted for 1039 Hispanic Americans plus 13 quality control (QC) duplicates using the Illumina OmniExpress Array (Illumina Inc.; San Diego, CA, USA; n=730,525 markers) with an additional 14 external controls included to verify reproducibility across genotyping runs. Exome chip genotyping was carried out on the Illumina HumanExome Array v1.0 (n=560) and v1.1 (n=864) in the Center for Genomics and Personalized Medicine Research at Wake Forest School of Medicine, Winston-Salem, NC, USA. A detailed description of the quality control procedures can be found in supplemental materials. Overall, 687,094 polymorphic autosomal SNPs from the OmniExpress and 81,599 SNPs from the exome chip were analyzed in 1034 and 1263 individuals, respectively. Among them, 18,289 SNPs were overlapping between the two platforms. Genotype concordance rate was over 99.9%.

Phenotypes

Anthropometric measures of adiposity were obtained using standard methods including height, weight, waist circumference (minimum between 10^{th} rib and iliac crest), and hip circumference (maximum circumference at the buttocks). BMI was calculated as weight in kilograms divided by height in meters squared. A CT scan was performed to estimate visceral and subcutaneous fat area (cm^2). This procedure consisted of a single scout of the abdomen followed by a 10-mm thick axial image at the L4-L5 disc space using a standard protocol. CT images were read

centrally at the University of Colorado Health Sciences Center. VAT and SAT were computed as previously described (Norris, et al., 2005). Percent body fat (PBF) was measured using DEXA at a 5 year follow-up exam, thus a reduced sample size as compared to other measures was available. A whole body DEXA scan uses the differential attenuation of two low dose x-ray beams to partition total body mass into bone, lean, and fat soft tissue components based on established mass-attenuation constants for bone mineral and lipid. Percent body fat (PBF) was calculated using total fat mass divided by measured weight x 100.

Statistical Analysis for GWAS and Exome Chip

Phenotypes were transformed to best approximate the distributional assumptions of conditional normality and homogeneity of variance. Specifically, BMI, WAIST, and WHR were natural log transformed, SAT and VAT values were square-root transformed and PBF required no transformation. Admixture estimates were calculated using maximum likelihood estimation of individual ancestries using ADMIXTURE (Alexander, Novembre, & Lange, 2009). Specifically, the largest set of uncorrelated markers (r^2<0.1) for K populations yielding the lowest cross validation (CV) error was used for unsupervised calculation of ancestral proportions. Representative ancestral populations from HapMap (CEU, YRI, CHN, and MEX) were included in the analysis. For GWAS, 117,347 LD-pruned SNPs for K=5 populations (CV error=0.48) were used. For exome chip, 10,566 uncorrelated SNPs for K=5 populations (CV error=0.52) were used. Three admixture estimates explained the largest amount of variation within the data and were highly correlated (r^2>0.93) across platforms. Tests of association between individual variants and quantitative traits were computed using the Wald test from the variance component model implemented in Sequential Oligogenic Linkage Analysis Routines (SOLAR) (Almasy & Blangero, 1998). Genetic models of association were calculated adjusting for age, gender, recruitment center, and

admixture estimates. The primary inference was the additive genetic model. A lack of fit to the additive model was also tested using the orthogonal contrast (-1, 2,-1). If that lack of fit was significant ($P<0.05$), the model with the "best" p-value is the minimum of the dominant, additive, and recessive. Overall, the results were modestly inflated with inflation factors ranging from 1.04 to 1.08. For robust estimation purposes, the additive and recessive genetic models were not computed if there were less than 10 and 20 individuals homozygous for the minor allele, respectively (similar to a minimal MAF of 1% and 2%). Conditional analysis was performed by adding the SNP with the strongest statistical significance to the model as a covariate.

Power analysis

Power was computed using QUANTO (http://hydra.usc.edu/GxE). Simulations suggest that for these pedigrees the effective sample size equivalent to unrelated individuals for a quantitative trait is 92%. Thus, power calculations were based on a sample size of 951 for GWAS and 1162 for exome chip. The statistical power of our study to detect SNP-trait associations was computed assuming a type 1 error rate of $\alpha=5.0\times10^{-8}$. Overall, the OmniExpress had power of 0.70, 0.80, and 0.90 to detect SNP-trait associations that explain 3.7%, 4.1% and 4.7% of the trait variation, respectively. Similarly, the exome chip had power of 0.70, 0.80, and 0.90 to detect SNP-trait associations that explain 3.0%, 4.1% and 4.7% of the trait variation, respectively.

De novo Genotyping in IRASFS and IRAS

In an effort to directly replicate the top association signals observed from exome chip and to search for potential causal SNPs at the *IDH1* and *ZGRF1* loci, a total of 76 SNPs were genotyped using the Sequenom MassARRAY Genotyping System (Sequenom, San Diego, CA, USA). Among these, 51 SNPs from the exome chip were chosen for genotyping in IRAS (n=184) for replication

($P<5.0x10^{-5}$). Another 25 SNPs (including 13 missense SNPs) within the *IDH1* and *ZGRF1* loci which were not covered by GWAS or exome chip were chosen for genotyping in IRASFS. Overall, genotyping efficiency was greater than 95%. To evaluate genotyping accuracy, 12 and 72 blind duplicate samples were included in IRAS and IRASFS, respectively. For all SNPs, genotyping was 99% concordant. PedCheck was computed for IRASFS genotype data and resulted in zeroing of 24 genotypes due to Mendelian inconsistencies (O'Connell & Weeks, 1998). Association analysis in IRASFS was computed using SOLAR as described. Analysis of data from IRAS was computed using QSNPGWA (https://www.phs.wakehealth.edu/public/home.cfm). Overall, 38 of the 51 SNPs were polymorphic and all SNP genotypes conformed to Hardy-Weinberg expectation ($P>0.05$).

Replication and Meta-analysis

Six cohorts participating in the GUARDIAN consortium (Goodarzi et al., 2014) provided in silico replication data: the Insulin Resistance Atherosclerosis Study (IRAS, $n_{max}=184$) (Wagenknecht et al., 1995), BetaGene ($n_{max}=1218$) (Black et al., 2008; Li et al., 2009; Shu et al., 2009; Watanabe et al., 2007), the Troglitazone in Prevention of Diabetes Study (TRIPOD, $n_{max}=125$) (Buchanan et al., 2000; Buchanan et al., 2002), the Hypertension-Insulin Resistance Family Study (HTN-IR $n_{max}=666$) (Cheng et al., 2001; Xiang et al., 2001), the Mexican-American Coronary Artery Disease Study (MACAD, $n_{max}=749$) (Goodarzi et al., 2004; Goodarzi et al., 2003; Goodarzi et al., 2005) and the NIDDM-Atherosclerosis Study (NIDDM-Athero, $n_{max}=179$) (Y.-P. Wang et al., 2000). A detailed description of the replication cohorts can be found in the supplemental materials. All cohorts, including IRASFS, were genotyped centrally as described above. All study protocols were approved by the local institutional review committees and all participants gave their informed consent.

A total of 71 GWAS SNPs ($P<5.00x10^{-6}$) from the six adiposity phenotypes were selected for replication in the six cohorts in the GUARDIAN consortium. Meta-analysis of BMI, WAIST, and WHR was computed using the fixed effect model implemented in METAL (www.sph.umich.edu/csg/abecasis/metal/) as well as a random effect model in Metasoft (Han & Eskin, 2011) (http://genetics.cs.ucla.edu/meta/). For PBF, only IRASFS, BetaGene, MACAD, and HTN were included. For SAT and VAT, as they were not available in replication cohorts, a weighted meta-analysis of the p-values and samples sizes using surrogate phenotypes was performed. For example, BMI was used as the surrogate for PBF in IRAS, TRIPOD, and HTN-IR; BMI for SAT in all six replication cohorts; and WAIST for VAT in all six replication cohorts.

Evaluation of previously identified signals

A total of 127 independent signals ($r^2<0.8$) associated with adiposity and adiposity-related traits with genome-wide significance from previously published studies were evaluated (Hindorff LA). A complete list of phenotypes used for the query can be found in supplemental material. Proxy SNPs ($r^2>0.8$) for each of the 127 tag SNPs were also identified using SNAP Proxy Search (Johnson et al., 2008) under the 1000 Genomes Pilot 1 SNP data set with a distance limit of 500kb. Association analysis was computed for all proxy SNPs with the six adiposity traits in IRASFS. Imputation of targeted variants not present on the OmniExpress Array was performed using IMPUTE2 (B. N. Howie, Donnelly, & Marchini, 2009). All IRASFS samples genotyped on the OmniExpress Array (n=1034) were imputed together using the 1000 Genomes Integrated Reference Panel (March 2012). In addition, 67 SNPs with associations to BMI and obesity from the 127 SNPs were selected for risk score analysis. The risk score was generated based on the number of risk alleles of the 67 SNPs. Associations of the risk score with six obesity phenotypes was conducted using SOLAR adjusting for age, gender, center, and admixtures.

Results

Characteristics of the study samples are shown in **Table 1**. Across all studies there was a higher proportion of females. On average, individuals were overweight with a mean BMI greater than 28kg/m^2. The IRASFS exome chip analysis included an additional 229 samples (n=1263) compared to GWAS (n=1034), of which 161 were individuals with T2D. This resulted in modestly increased means in adiposity-related traits.

In IRASFS, 687,094 polymorphic autosomal SNPs from the OmniExpress and 81,599 SNPs from the exome chip were analyzed in 1034 and 1263 individuals, respectively. A summary of the association results are shown in **Figure 1** and **Table 2**. In total, five SNPs from two loci reached genome-wide significance ($P<5.00x10^{-8}$). Among these were four highly correlated SNPs (rs13144672, rs7696816, rs1471880, rs12054518; $r^2>0.96$) associated with WHR in the Zinc finger, GRF-type containing 1 gene (*ZGRF1*). SNP rs1471880 (MAF=48.1%), an intronic variant, showed the strongest signal of association under a dominant genetic model (WHR, $P_{DOM}=1.00x10^{-8}$) and explained 2.7% of the variance in WHR. On average, minor allele carriers have 2.3% lower of WHR (0.84±0.082 as compared to 0.86±0.085 in non-carriers). The second genome-wide significant signal was rs34218846 (MAF=6.8%) with WAIST ($P_{DOM}=1.62x10^{-8}$). This SNP explains 2.1% of the phenotypic variance and marks a valine to isoleucine change (V178I) in the Isocitrate Dehydrogenase 1 gene (*IDH1*) on chromosome 2. De novo genotyping of additional, putatively functional SNPs at these loci in the IRASFS cohort did not identify additional statistically significant variants.

Replication of signals from the IRASFS GWAS (n=71 SNPs with $P<5.00x10^{-6}$) was attempted through meta-analysis with six additional Hispanic-American cohorts. Overall, no SNP attained genome-wide significance after meta-analysis **(Table 3)**. The most significant signal remained to be rs1471880 ($P_{DOM}=8.38x10^{-6}$) at the *ZGRF1* locus associated with WAIST, which was also the

strongest signal identified by GWAS (WHR, $P_{DOM}=1.00x10^{-8}$, WAIST $P_{DOM}=6.47x10^{-7}$). Among replication cohorts, similar allele frequencies and a consistent direction of effect were observed in five of the larger cohorts while the two smaller cohorts, TRIPOD (n=125) and NIDDM-Athero (n=179), had an opposite direction of effect. For *IDH1*, the top SNP rs34218846 was identified from exome chip and was not available for in silico replication among the additional cohorts. Analysis of two GWAS proxy SNPs, rs6435435 ($r^2=0.91$ with rs34218846, $P_{DOM}=1.73x10^{-6}$ for BMI) and rs6734788 ($r^2=0.37$ with rs34218846, $P_{ADD}=7.33x10^{-7}$ for WAIST), near *IDH1* resulted in decreased significance (rs6435435 $P_{DOM}=0.11$ with BMI and rs6734788 $P_{ADD}=7.98x10^{-4}$ with WAIST) with inconsistent directions of effect. *De novo* genotyping of variants at the *IDH1* locus in IRAS (n=187) revealed five nominally associated SNPs (P<0.05), of which two SNPs, rs12105636 (BMI $P_{ADD}=0.046$) and rs16840781 (BMI $P_{DOM}=0.030$), were significant with a consistent direction of effect. However, the top *IDH1* missense SNP (rs34218846) was not significant (WAIST $P_{ADD}=0.45$) with an opposite direction of effect. SNP rs12105636 and rs16840781 had nominal association signals in the IRASFS GWAS (WAIST $P_{DOM}=3.94x10^{-3}$ and $P_{DOM}=2.33x10^{-3}$, respectively) and were poorly correlated with rs34218846 ($r^2=0.34$).

In addition to the search for novel adiposity variants, 127 independent signals ($r^2<0.8$) associated with adiposity and adiposity-related traits with genome-wide significance from previously published studies were evaluated in the IRASFS. Among these, 116 SNPs were directly genotyped or successfully imputed in IRASFS. Overall, 71 SNPs showed nominal association (P<0.05) with consistent direction of effect for at least one of the six adiposity traits. These included 23 SNPs for BMI, 17 SNPs for WAIST, 13 SNPs for WHR, 31 SNPs for SAT, 21 SNPs for VAT, and 13 SNPs for PBF. A two-sided nonparametric sign test was computed for the p-value thresholds of 0.10, 0.05, 0.01, and $4.31x10^{-4}$ (based on a Bonferroni correction of 116 variants). In brief, significantly higher replication signal concordance was observed with SAT and VAT

(P<0.05). However, no replication signal survived Bonferroni correction. The strongest signal observed was rs2820464 located intergenically between lysophospholipase-like 1 gene (*LYPLAL1*) and solute carrier family 30, member 10 gene (*SLC30A10*) associated with SAT ($P_{ADD}=7.06x10^{-4}$). This variant was identified in a European cohort for an association with WHR ($P=7.00x10^{-9}$) (Berndt, et al., 2013). Risk score analysis of the 67 previously identified obesity SNPs showed the strongest signal for SAT ($P=5.9x10^{-4}$). BMI, WAIST, and PBF were nominally associated with P-values $2.2x10^{-3}$, $7.7x10^{-3}$, and $3.2x10^{-3}$, respectively. Not surprising, VAT (P=0.22) and WHR (P=0.83) were not associated with the risk score as they are measures of adiposity depositions instead of total fat volumes.

Discussion

Here we present a combined study of genome-wide and exome chip arrays to investigate the genetic determinants of adiposity measures in the Hispanic-American population. The complementary approach of using GWAS and exome chip enabled a broader coverage of both common and rare functional variants, resulting in an increased chance to identify causal mutations. Obesity-related traits evaluated included anthropometric (WAIST, WHR, and BMI), CT (SAT and VAT), and DEXA (PBF) measures. The assessment of CT and DEXA scans provided more accurate estimates of regional and total adiposity, respectively. We evaluated associations among Hispanic Americans from IRASFS ($n_{max}=1263$) using GWAS and exome chip analysis with replication in six independent Hispanic cohorts ($n_{max}=4155$). Association studies revealed *ZGRF1* and *IDH1* as two possible novel adiposity-related loci: *ZGRF1* was associated with waist-hip ratio ($P_{DOM}=1.00x10^{-8}$) and *IDH1* was associated with waist circumference ($P_{DOM}=1.62x10^{-8}$).

Overall, three intronic variants and one missense SNP in *ZGRF1* were identified above genome-wide significance for WHR **(Table 2)**. The missense mutation (rs7696816) marks an asparagine to

serine amino acid change with a benign effect predicted by PolyPhen (Adzhubei et al., 2010). The specific function of this gene remains unclear. The overall expression of *ZGRF1* in the human body is relatively low with the exception in brain and testis (Kapushesky et al., 2010). Direct replication of the *ZGRF1* signals was performed across six cohorts and the strongest signal from meta-analysis was rs1471880, $P_{DOM}=8.38x10^{-6}$ **(Table 3)**. A consistent direction of effect was observed across the five larger cohorts ($n_{max}=3645$) However, the statistical significance decreased. Examination of this region in the GIANT (Genetic Investigation of Anthropometric Traits) Consortium for BMI and class 1 obesity (BMI>30) failed to reveal significant signals of association at the *ZGRF1* locus (P>0.01) (Berndt, et al., 2013; Yang et al., 2012). Interestingly, previous studies have identified *ALPK1* (rs4833407), 100kb proximal to *ZGRF1*, to be associated with obesity in European populations (Bradfield et al., 2012). However, the two SNPs in *ALPK1* and *ZGRF1* were poorly correlated in both CEU and IRASFS Hispanic Americans ($r^2=0.005$ and 0.013, respectively). In IRASFS, most association signals centered around the *ZGRF1* locus with a few in *NEUROG2* and very weak signals in *ALPK1* **(Figure 2)**. *NEUROG2* is a proneural protein neurogenin and has been shown to control cortical neuron migration through the regulation of small GTP-binding protein Rnd2 (Heng et al., 2008) and no direct link with adiposity has been established. Conditional analysis of this region with rs1471880 as a covariate abolished all association signals in *ZGRF1* as well as the signals in nearby *NEUROG2* without changes in *ALPK1* **(Figure 2)**.

IDH1 encodes cytosolic NADP+ dependent isocitrate dehydrogenase (IDPc) which has been proposed as a key enzyme for supplying cytosolic NADPH (Nam, Park, & Park, 2012). The most significant association signal observed was SNP rs34218846 (MAF=0.068; $P_{DOM}=1.62x10^{-8}$) encoding a missense mutation from valine to isoleucine in exon 6 and was predicted as "probably damaging" by PolyPhen (Adzhubei, et al., 2010). This mutation is located at the

subunit dimerization interface, suggesting a potential regulatory role in gene function. Previous genetic studies have suggested a strong correlation between *IDH1* mutations and cancer (Reitman & Yan, 2010). A biological link between *IDH1* and adiposity has been postulated using cell models. Specifically, stable transfection of *IDH1* cDNA positively correlated with adipogenesis of 3T3-L1 cells whereas decreased IDPc expression using an antisense IDPc vector retarded 3T3-L1 adipogenesis (Koh et al., 2004). A more recent study reported knockdown of IDPc expression by RNA interference (RNAi) which inhibited adipocyte differentiation and lipogenesis in 3T3-L1 preadipocytes. In addition, in diet-induced obese mice transduced with IDPc short-hairpin RNA, a loss of body weight and reduction of triglyceride levels were observed (Nam, et al., 2012). The evaluation of serum triglyceride levels in IRASFS revealed carriers of rs34218846 T allele (adiposity protective allele) had a 20mg/dL decrease in triglyceride levels compared to non-carriers ($P_{DOM}=7.79x10^{-3}$). Taken together, *IDH1* appears to play an important role in fat metabolism. SNP rs34218846 was not directly genotyped among the replication cohorts. Therefore, two proxy SNPs, rs6435435 ($P_{DOM}=1.73x10^{-6}$ for BMI, $r^2=0.91$ with rs34218846) and rs6734788 ($P_{ADD}=7.33x10^{-7}$ for WAIST, $r^2=0.37$ with rs34218846), were selected for meta-analysis. However, these proxies failed to replicate (rs6435435 $P_{DOM}=6.74x10^{-5}$ for WAIST and rs6734788 $P_{ADD}=9.02x10^{-5}$ for WHR). Lack of association was similarly observed in IRAS (n=184) with direct genotyping of rs34218846 (P= 0.45), which could be attributed reduced power given the small sample size. To search for additional putatively causal variants in *IDH1*, we conducted de novo genotyping in IRASFS which revealed an intronic SNP (rs59684347) showing stronger evidence of association ($P_{ADD}=7.42x10^{-9}$; WAIST) **(Figure 3)**. However, rs34218846 and rs59684347 were highly correlated ($r^2=1.00$) and all evidence of association in the region was abolished after inclusion of rs34218846 as a covariate in the analysis **(Figure 3)**. Overall, *IDH1* represents a promising locus with evidence of association to adiposity-related

traits, especially waist circumference. Notably, larger cohorts from European-derived populations in the GIANT Consortium have identified BMI associated signals in *CRYGD* (rs10932241), which is 100kb proximal to *IDH1*. However, there was no signal of association for the *IDH1* locus in GIANT and rs10932241 was poorly correlated with rs34218846 (r^2=0.057) and only nominally associated with BMI (p-value=0.057) in IRASFS despite a similar minor allele frequency observed in European populations (MAF=5.31%) (Berndt, et al., 2013; Yang, et al., 2012).

In summary, although encouraging results have been revealed, there are several study limitations. Like most minority studies, sample size largely limited the power, especially for rare variants assessed on the exome chip. In addition, the utility of the Illumina HumanExome Array in Hispanic Americans is not optimal as only 81,559 out of 242,901 SNPs on the array were polymorphic, likely attributable to a design based on findings in Caucasians and African Americans. The application of Illumina OmniExpress BeadChip has similar concerns: the SNPs on the chip may not tag the LD structure as well in Hispanic Americans. Another issue is the lack of replication signals: all signals fell below the significance threshold after meta-analysis. There are several possible reasons: first, the replication cohorts were limited to directly genotyped GWAS variants and we were unable to replicate signals from the exome chip among all cohorts. Second, some replication cohorts did not have CT and DEXA measures for replication, necessitating the use of surrogate phenotypes. Third, while all cohorts were of Hispanic ancestry, different ascertainment criteria were used. For example, BetaGene recruited participants at high risk of gestational diabetes while HTN-IR recruited participants at high risk of hypertension. This differs from IRASFS which is a population-based study recruited based on large family size. Additionally, the sample sizes for IRAS, TRIPOD, and NIDDM-Athero were relatively small. This may explain why the more significant associations, e.g. rs1471880 demonstrated an opposite direction of

effect in TRIPOD (n=125) and NIDDM-Athero (n=179). Another concern is the large effects of *IDH1* (2.1%) and *ZGRF1* (2.7%) in this study are weak from previous European population studies. One explanation is the potential for ethnic-specific variants or that the signals are the result of gene-environment effects. It is also possible that the signals observed are not causal and they were detected due to a long range LD with other loci.

Until now, VIVA LA FAMILIA was the only cohort with published genome-wide significant obesity-related signals specific to the Hispanic population (Comuzzie, et al., 2012). Further evaluation of the obesity-related loci from VIVA LA FAMILIA in IRASFS revealed nominal association for rs2823615 ($P_{DOM}=7.86x10^{-3}$ with SAT), an intronic SNP in the Family with Sequence Similarity 222 Member A gene (*FAM222A*). This SNP has been shown to be associated with increased respiratory quotient in VIVA LA FAMILIA and increased SAT in IRASFS.

In summary, we computed a combined study of genome-wide and exome chip arrays in the IRASFS Hispanic-American population. Six obesity related traits were analyzed for association. *ZGRF1* and *IDH1* attained genome-wide significance in IRASFS and replication of significant signals was evaluated in six additional Hispanic cohorts ($n_{max}=4155$). Meta-analysis suggested decreased levels of significance (*ZGRF1* rs1471880, $P_{DOM}=8.38x10^{-6}$; *IDH1* rs6435435, $P_{DOM}=6.74x10^{-5}$). These results highlight the importance of GWAS and exome chip research in minority populations where an increased prevalence of adiposity-related diseases may be associated with a differential genetic architecture than in European-derived populations.

Acknowledgements

This research was supported the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). Computational resources were provided, in part, by the Wake Forest School of Medicine Center for Public Health Genomics. The authors would like to acknowledge the

members of the GUARDIAN Consortium. The authors thank the other investigators, the staff, and the participants of the studies for their valuable contributions.

Table 1. Demographic characteristics of the study populations.

	IRASFS		Replication Cohorts					
	GWAS	Exome Chip	IRAS	TRIPOD	BetaGene	HTN-IR	MACAD	NIDDM-Athero
n	1034	1263[c]	184	125	1218	666	749	179
Male (%)	41.1	41.1	41.1	0	28.4	40.6	43.3	41.8
Age (years)[b]	40.6±13.7	42.8±14.6	54.0±8.2	34.8±6.3	34.5±8.2	37.4±14.2	34.5±8.8	31.8±9.69
Body Mass Index (BMI; kg/m^2)[b]	28.3±5.8	28.9±6.1	28.2±5.1	30.6±5.4	29.5±6.1	28.8±5.5	28.9±5.1	28.6±6.3
Waist circumference (WAIST; cm)[b]	88.2±13.6	89.4±15.5	90.1±12.2	91.4±12.6	94.2±14.0	90.7±14.5	92.7±12.5	91.3±14.8
Waist Hip Ratio (WHR)[b]	0.85±0.083	0.86±0.085	0.87±0.088	0.86±0.061	0.89±0.073	0.88±0.084	0.89±0.076	0.84±0.087
Subcutaneous Adipose Tissue (SAT; cm^2)[b]	329.8±151.7	339.1±154.7	NA	NA	NA	NA	NA	NA
Visceral Adipose Tissue (VAT; cm^2)[b]	106.6±57.0	114.0±61.3	NA	NA	NA	NA	NA	NA
Percent Body Fat (PBF)[b]	33.5±8.7	34.0±8.7	NA	NA	34.4±8.4	32.6±9.1[a]	32.2±8.3	NA

[a]HTN-IR has only 139 individuals with PBF measurements; [b]Values are expressed as the mean ± standard deviation; [c]Includes 161 diabetics.

Abbreviations: IRASFS, Insulin Resistance Atherosclerosis Family Study; IRAS, Insulin Resistance Atherosclerosis Study; TRIPOD, Troglitazone in Prevention of Diabetes Study; BetaGene, Family-based study of obesity, insulin resistance, and beta-cell dysfunction; HTN-IR, Hypertension-Insulin Resistance Family Study; MACAD, Mexican-American Coronary Artery Disease Study; NIDDM-Athero, NIDDM-Atherosclerosis Study.

Table 2. Significant signals of association from genome-wide and exome chip association analyses in IRASFS.

SNP	Chr: Position (hg19)	Gene	Alleles[f]	N	MAF[e]	Beta+/-SE	P-Value
Body Mass Index (BMI)							
rs34218846[b]	2:209108317	*IDH1*	T/C	1253	0.070	-0.10±0.020	4.81E-08[c]
Waist Circumference (WAIST)							
rs34218846[b]	2:209108317	*IDH1*	T/C	1257	0.068	-0.080±0.010	1.62E-08[c]
Waist-Hip Ratio (WHR)							
rs1471880[a]	4:113546107	*ZGRF1*	C/A	1034	0.48	-0.027±0.0047	1.00E-08[c]
rs13144672[a]	4:113472958	*ZGRF1*	C/T	1034	0.47	0.027±0.0048	3.15E-08[d]
rs12054518[a]	4:113549989	*ZGRF1*	A/G	1034	0.48	0.027±0.0048	3.23E-08[d]
rs7696816[a]	4:113539969	*ZGRF1*	C/T	1034	0.48	0.027±0.0048	4.35E-08[d]

[a]SNP identified from GWAS; [b]SNP identified from exome chip; [c]Dominant Model; [d]Recessive Model; [e]Minor allele frequency based on the entire population; [f]Minor/Major allele on the positive strand.

Table 3. Fixed-effect meta-analysis results ($P<2.0x10^{-3}$) for significant signals of association ($5.00x10^{-6}$) from IRASFS.

SNP	Chr:Position (hg19)	Alleles[b]	Effect	SE	Weight	Zscore	Direction[f]	*Gene Name*	P-value
Waist Circumference (WAIST)[a]									
rs1471880	4:113546107	A/C	0.0221	0.005			+++-++-	*ZGRF1*	8.38E-06[d]
rs6435435	2:209112551	A/G	0.0307	0.0077			+-++-++	*IDH1*	6.74E-05[d]
rs13144672	4:113472958	A/G	-0.0131	0.0033			---+--+	*ZGRF1*	6.90E-05[c]
rs7696816	4:113539969	A/G	-0.0126	0.0033			---+--+	*ZGRF1*	1.01E-04[c]
rs12054518	4:113549989	A/G	0.0125	0.0033			+++-++-	*ZGRF1*	1.24E-04[c]
rs1869479	11:44343856	A/C	0.0176	0.0047			+-+-++-	*HSD17B12/CD82*	2.01E-04[d]
rs6734788	2:209093069	A/G	-0.0194	0.0058			-+-++++	*CCNYL1/IDH1*	7.98E-04[c]
rs7937515	11:71841325	A/G	0.0141	0.0046			+-+-++-	*FAM86C1/FOLR3*	2.00E-03[c]
Waist-Hip Ratio (WHR)[a]									
rs7696816	4:113539969	A/G	-0.0116	0.0028			---++--	*ZGRF1*	4.34E-05[e]
rs13106629	4:113459416	A/G	0.0116	0.0028			+++--++	*C4orf32/ZGRF1*	4.39E-05[e]
rs12054518	4:113549989	A/G	0.0115	0.0028			+++--++	*ZGRF1*	5.43E-05[e]
rs2129405	4:113447137	A/G	0.0113	0.0028			+++--++	*C4orf32/ZGRF1*	6.24E-05[e]
rs10770244	12:17848331	A/G	-0.0091	0.0023			---++-+	*MIR3974/Y_RNA*	6.67E-05[d]
rs13144672	4:113472958	A/G	-0.0111	0.0028			---++--	*ZGRF1*	7.22E-05[e]
rs6734788	2:209093069	A/G	-0.0109	0.0028			-+-++-+	*CCNYL1/IDH1*	9.02E-05[c]

[a]Meta-analysis was computed based on beta and SE; [b] Reference/alternate allele; [c]Additive model; [d]Dominant model; [e]Recessive model; [f]Direction follows as: IRASFS, IRAS, BetaGene, TRIPOD, HTN-IR, MACAD, NIDDM-Athero.

Figure 1. Manhattan plots for genome-wide and exome chip association analysis in IRASFS Hispanic Americans: **A**. Body Mass Index (BMI), **B**. Waist Circumference (WAIST), **C**. Waist-Hip Ratio (WHR), **D**. Subcutaneous Adipose Tissue (SAT), **E**. Visceral Adipose Tissue (VAT), and **F**. Percent Body Fat (PBF). Results were adjusted for age, gender, recruitment center (San Antonio, TX or San Luis Valley, CO), and admixture estimates. P-values are shown under the best fit model. The blue line at $-\log_{10}(PVAL)= 4$ represents a best P-value=10^{-4} and the red line at $-\log_{10}(PVAL)=5.5$ represents a best P-value=$3.16x10^{-6}$.

Figure 1

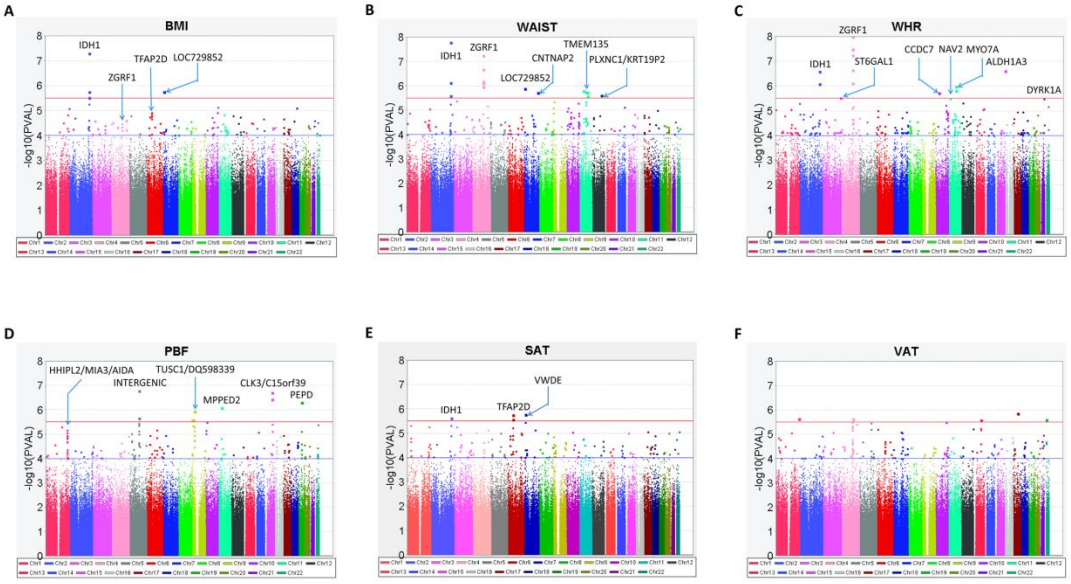

A
BMI
IDH1
TFAP2D
LOC729852
ZGRF1
B
WAIST
ZGRF1
IDH1
TMEM135
CNTNAP2
PLXNC1/KRT19P2
LOC729852
C
WHR
ZGRF1
IDH1
ST6GAL1
CCDC7
NAV2
MYO7A
ALDH1A3
DYRK1A
D
PBF
HHIPL2/MIA3/AIDA
INTERGENIC
TUSC1/DQ598339
MPPED2
CLK3/C15orf39
PEPD
E
SAT
IDH1
TFAP2D
VWDE
F
VAT
-log10(PVAL)

Figure 2. Regional plot of *ZGRF1* (*C4orf21*) for association with waist-hip ratio. A. Analysis results in IRASFS for SNPs from genome-wide and exome chip datasets; **B**. Conditioned on the most significant variant (rs1471880). $-\log_{10}$(p-values) under the best fit model are indicated on the left-hand Y axis. Association analyses were computed with adjustment for age, gender, recruitment center, and admixture estimates with SNP rs1471880 as an additional covariate in panel B. The recombination rates are indicated on the right-hand Y axis based on HapMap. The color of each SNP annotates its correlation (r^2) with the index SNP and was taken from the 1000 Genomes AMR population. A circle denotes intronic and intergenic SNPs, a triangle denotes a missense SNP, and a square denotes a SNP in the untranslated region (UTR).

Figure 2

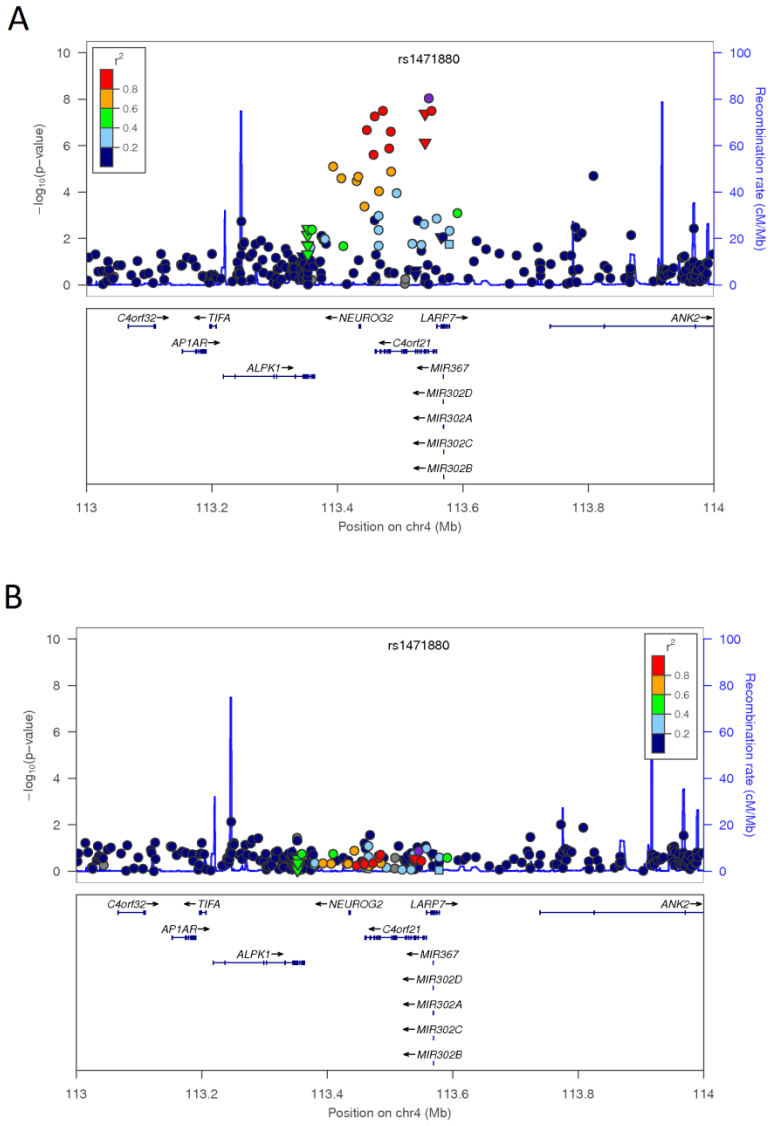
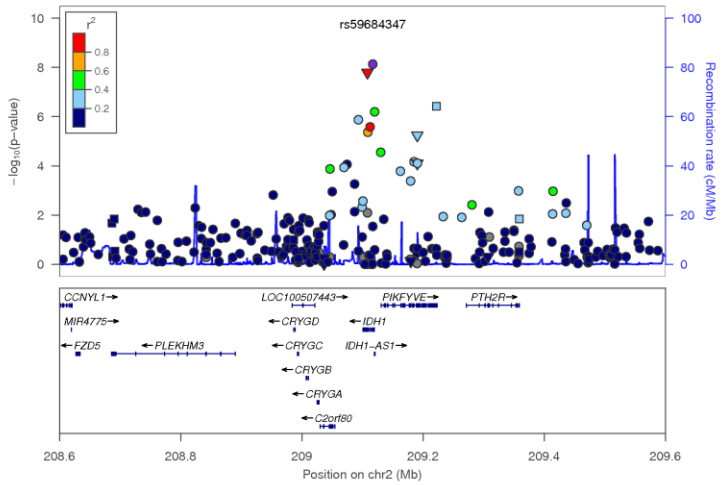

A
rs1471880
r^2
0.8
0.6
0.4
0.2
$-\log_{10}$(p-value)
Recombination rate (cM/Mb)
C4orf32
TIFA
NEUROG2
LARP7
ANK2
AP1AR
C4orf21
ALPK1
MIR367
MIR302D
MIR302A
MIR302C
MIR302B
113
113.2
113.4
113.6
113.8
114
Position on chr4 (Mb)
B
rs1471880
r^2
0.8
0.6
0.4
0.2
$-\log_{10}$(p-value)
Recombination rate (cM/Mb)
C4orf32
TIFA
NEUROG2
LARP7
ANK2
AP1AR
C4orf21
ALPK1
MIR367
MIR302D
MIR302A
MIR302C
MIR302B
113
113.2
113.4
113.6
113.8
114
Position on chr4 (Mb)

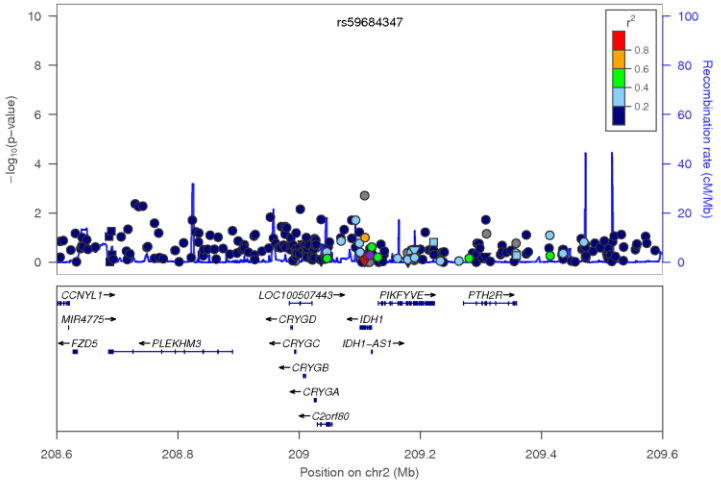
Figure 3. Regional plot of *IDH1* for association with waist circumference. **A**. Analysis results in IRASFS for SNPs from genome-wide and exome chip datasets as well as de novo genotyping of the region; **B**. Conditioned on rs34218846. $-\log_{10}$(p-values) under the best fit model are indicated on the left-hand Y axis. Association analyses were computed with adjustments for age, gender, recruiting center, and admixtures with SNP rs34218846 as an additional covariate in panel B. The recombination rates are indicated on the right-hand Y axis based on HapMap. The color of each SNP annotates its correlation (r^2) with the index SNP and was taken from the 1000 Genomes AMR population. A circle denotes intronic and intergenic SNPs, a triangle denotes a missense SNP, and a square denotes a SNP in the untranslated region (UTR).

Figure 3

A


rs59684347
r²
0.8
0.6
0.4
0.2
−log10(p-value)
Recombination rate (cM/Mb)
CCNYL1
LOC100507443
PIKFYVE
PTH2R
MIR4775
CRYGD
IDH1
FZD5
PLEKHM3
CRYGC
IDH1-AS1
CRYGB
CRYGA
C2orf80
208.6
208.8
209
209.2
209.4
209.6
Position on chr2 (Mb)

B


rs59684347
r²
0.8
0.6
0.4
0.2
−log10(p-value)
Recombination rate (cM/Mb)
CCNYL1
LOC100507443
PIKFYVE
PTH2R
MIR4775
CRYGD
IDH1
FZD5
PLEKHM3
CRYGC
IDH1-AS1
CRYGB
CRYGA
C2orf80
208.6
208.8
209
209.2
209.4
209.6
Position on chr2 (Mb)

Chapter III

Genome-wide study of subcutaneous and visceral adipose tissue reveals novel sex-specific adiposity loci in Mexican Americans: The Insulin Resistance Atherosclerosis Family Study

Chuan Gao, Carl D. Langefeld, Julie T. Ziegler, Kent D. Taylor, Jill M. Norris, Yii-Der I. Chen, Jacklyn N. Hellwege, Elizabeth K. Speliotes, Jerome I. Rotter, Donald W. Bowden, Lynne E. Wagenknecht, Nicholette D. Palmer

The following manuscript will be submitted to Journal of Human Genetics. Stylistic variations are due to the requirements of the journal. Chuan Gao performed the analyses and prepared the manuscript. Dr. Nicholette Allred acted in an advisory and editorial capacity.

Abstract:

Regional fat deposition is a strong risk factor for metabolic diseases beyond total adiposity. However, the exact genetic mechanisms remain unclear. Here we performed a genome-wide association study of ~8 million SNPs in 983 Mexican Americans (N_{male}=403, N_{female}=580) from the insulin resistance atherosclerosis family study (IRASFS). Association analyses were performed with and without sex stratification for subcutaneous adipose tissue (SAT), visceral adipose tissue (VAT), VAT adjusted for body mass index (VAT_BMI), and visceral-subcutaneous ratio (VSR) obtained from computed tomography (CT). The strongest signal was rs2185405 (MAF=40%, $P_{VAT}=1.98x10^{-8}$), an intronic variant of a well-characterized diabetes gene *GLIS3* (GLIS family zinc finger 3 gene). In addition, SNP rs12657394 (MAF=19%) was associated with VAT_BMI in males ($P_{male}=2.39x10^{-8}$; $P_{female}=2.5x10^{-3}$) and locates intronically in *SRFBP1* (serum response factor binding protein 1 gene). On average, male carriers of the variant have 24.6cm^2 increased VAT compared to non-carriers. Subsequently, genome-wide SNP-sex interaction analysis was performed. SNP rs10913233 (MAF=14%, $P_{int}=3.07x10^{-8}$) in *PAPPA2* and rs10923724 (MAF=38%, $P_{int}=2.89x10^{-8}$) upstream of *TBX15* were strongly associated with the interaction effect for VSR. In summary, six loci were identified with genome-wide significant associations with fat deposition and interactive effects and provided genetic evidence for a differential basis of fat deposition between genders.

Keywords: Genome-Wide Association Study; Exome Chip; Mexican Americans; Adiposity Deposition; Sex Heterogeneity

Introduction

Obesity is a global health epidemic affecting more than 500 million individuals worldwide and responsible for nearly three million deaths each year (Scully, 2014). Previous studies have confirmed obesity as a strong risk factor for many metabolic diseases including cardiovascular disease (CVD), type 2 diabetes (T2D), metabolic syndrome, certain types of cancer, stroke, and hypertension (Frayling, et al., 2007; Haslam & James, 2005; Narkiewicz, 2005). However, the exact mechanisms underlying these associations have been poorly elucidated.

Recent research has suggested obesity is not a homogeneous condition and that regional fat distribution affects glucose and lipid metabolism beyond total body adiposity (Wajchenberg, 2000). For example, visceral fat (VAT) has been shown to be responsible for the increased mortality and risk for metabolic disorders while subcutaneous fat (SAT) is thought to be benign (P. Cohen, et al., 2014; Lapidus, et al., 1984; Larsson, et al., 1984; Wajchenberg, 2000). Genetic studies have been successful in identifying genetic loci responsible for regional fat distribution using measures including waist circumference (WAIST) and waist-hip ratio (WHR). However, anthropometric measures can be impacted by skeletal structure and aging (Taylor, et al., 2000) and cannot differentiate between regional fat depots, e.g. visceral and subcutaneous adipose tissue, and therefore bias studies. Currently, computed tomography (CT) is considered as the gold standard for the measurement of adipose tissue deposition (Blaak, 2001; Norris, et al., 2005). However, due to cost and accessibility reasons, only one genome-wide association study (GWAS) has been published focusing on directly measured SAT and VAT with genome-wide significant signals (Fox, Liu, et al., 2012).

Numerous evidence have suggested strong sex specificity for regional adipose tissue distributions with females having a higher proportion of gluteal-femoral body fat whereas males

have more in the abdominal (visceral) region (Blaak, 2001; Zillikens et al., 2008). This observation suggests a potentially different mechanism for fat deposition in different genders. In 2015, the genetic investigation of anthropometric traits GIANT consortium published a genetic study of adipose tissue deposition using WHR and identified 49 (33 new) signals associated with 20 loci demonstrating sex-specific effects (Shungin et al., 2015). Until now, no formal genome-wide SNP-sex interaction analysis has been performed in Mexican Americans.

Here we report a genetic study of sex-specific adipose tissue deposition using CT measures including SAT and VAT in the Insulin Resistance Atherosclerosis Family Study (IRASFS). Genome-wide and exome chip association studies were combined to provide a more comprehensive scan of both common and rare variants. As adiposity deposition differs between genders, sex-stratified analyses as well as genome-wide SNP-sex interaction analyses were performed.

Materials and Methods

Insulin Resistance Atherosclerosis Family Study (IRASFS)

The study design, recruitment, and phenotyping for the IRASFS have been previously described (Gao et al., 2015; Henkin, et al., 2003). In brief, the IRASFS was designed to investigate the genetic and environmental basis of insulin resistance and adiposity. Individuals included in this cohort (N=1,417 individuals, 90 pedigrees) were Mexican Americans recruited from clinical centers in San Antonio, TX and San Luis Valley, CO. Since a diagnosis of diabetes was not required for participation, about 12.7% of individuals had diabetes. The study protocol was approved by the Institutional Review Board of each participating clinical and analysis site and all participants provided their written informed consent.

Phenotypes

Measures of adiposity were obtained using a standardized protocol. BMI was calculated as weight in kilograms divided by height in meters squared. Computed tomography (CT) scans were performed to estimate visceral and subcutaneous fat area (VAT and SAT, respectively; cm^2). This procedure consisted of a single scout of the abdomen followed by a 10-mm thick axial image. Axial images were obtained at L4-L5 disc space using a standard protocol. CT images were sent to a centralized reading center at the University of Colorado Health Sciences Center. VAT and SAT were computed from these data as previously described (Norris, et al., 2005). Visceral-subcutaneous ratio (VSR) was computed as the ratio of VAT and SAT.

Genotyping and Quality Control

Genotyping

GWAS genotyping was supported through the Genetics Underlying Diabetes in Hispanics (GUARDIAN) Consortium (Goodarzi, et al., 2014) using the Illumina OmniExpress and 1S Array (Illumina Inc.; San Diego, CA, USA) and Exome chip genotyping was carried out on the Illumina HumanExome Array . A detailed description of genotyping platforms and quality controls can be found in supplemental materials.

Imputation

Imputation was performed using IMPUTE2 (B. N. Howie, et al., 2009) and the 1000G phase I integrated reference panel. All IRASFS samples genotyped on the OmniExpress and 1S array were imputed together. Imputed variants were filtered (confidence score >0.90 and an information score >0.50) resulting in 8,000,632 polymorphic SNPs for analysis. Imputation

quality was evaluated using 10,000 SNPs with both exome chip and imputation coverage randomly selected from 32,729 overlapping SNPs.

Statistical Analysis

GWAS and Exome Chip

Phenotypes were transformed to best approximate the distributional assumptions of conditional normality and homogeneity of variance. Specifically, VSR was natural log transformed and SAT and VAT were square-root transformed. Admixture estimates were calculated using maximum likelihood estimation of individual ancestries as implemented in ADMIXTURE (Alexander, et al., 2009). Specifically, the largest set of uncorrelated markers (r^2<0.1) for K populations yielding the lowest cross-validation (CV) error was used for unsupervised calculation of ancestral proportions. Representative ancestral populations from HapMap (CEU, YRI, CHN, and MEX) were included in the analysis. For GWAS, 117,347 LD-pruned SNPs for up to K=5 populations (CV error=0.48) were used. For exome chip, 10,566 uncorrelated SNPs for K=5 populations (CV error=0.52) were used. Three admixture estimates explained the largest amount of variation within the data and were highly correlated (r^2>0.93) across platforms. For SNPs available in both GWAS imputation and exome chip, exome chip genotypes were always used for analysis.

Tests of association between individual variants and quantitative traits were computed using the Wald test from the variance component model implemented in Sequential Oligogenic Linkage Analysis Routines (SOLAR) (Almasy & Blangero, 1998). Genetic associations were calculated adjusting for age, sex, recruitment center (San Antonio, TX or San Luis Valley, CO), and admixture estimates. The primary inference was the additive model. A lack of fit to the additive model was tested using the orthogonal contrast. If the lack-of-fit test was significant

(P<0.05), the model with the "best" p-value as the minimum of the dominant, additive, and recessive genetic models was selected; this is estimated to have an inflation factor of 1.3. For robust estimation purposes, the additive and recessive genetic models were not computed if there were less than 10 and 20 individuals homozygous for the minor allele, respectively (this threshold is not applicable for SNPs with the non-significant lack-of-fit test). Interaction analysis was performed to test the beta-coefficient of the interaction variable using the same genetic model as the main effect.

Results

A demographic summary of the study samples is shown in **Table 1**. Overall, individuals were overweight with an average body mass index (BMI) greater than 28.3kg/m^2. In total, 983 and 1205 individuals with available phenotypes were analyzed for GWAS imputed and exome chip SNPs, respectively. Compared to GWAS, an additional 222 samples were included in the analysis of the exome chip data, of which 150 were individuals with T2D. This resulted in modestly increased means in age and adiposity traits (P<0.001). In addition, mean trait values of adiposity phenotypes were significantly different between females and males, i.e. females had significantly larger amounts of SAT (P<0.0001) while males had larger amounts of VAT (P<0.0001). The ratio between VAT and SAT (VSR) was two times greater in males than females (P<0.0001), suggesting males have the predisposition to store fat viscerally. In total, 7,708,309 SNPs with minor allele frequency greater than 1% from the IRASFS Mexican Americans were analyzed for comprehensive adipose tissue deposition measures including SAT, VAT, VSR, and VAT_BMI.

Imputation quality

The majority of the SNPs analyzed were derived from statistical imputation as opposed to direct genotyping. To evaluate imputation quality, kappa statistics were computed for 10,000 SNPs captured by both exome chip and imputation. Overall, SNP genotypes were well-matched between the two platforms with an r^2>0.95. However, for rare variants with MAF less than 1%, the imputation quality varied and therefore were excluded from the analysis.

Association Results

Association analyses were computed for SAT, VAT, VAT adjusted for BMI (VAT_BMI), and VSR adjusting for age, sex, recruitment center, and admixture estimates. A summary of the association results is shown in **Figure 1, Table 2**. The genome-wide significant signal was SNP rs2185405 (P_{dom}=1.98x10^{-8}, MAF=40%) with VAT. It is an intronic SNP located in the GLIS family zinc finger 3 gene (*GLIS3*). In addition, an intergenic SNP rs1504143 (12q21.33, P_{dom}=8.35x10^{-8}, MAF=12.5%) on chromosome 12 showed suggestive association with VSR.

Sex-stratified Association Analysis

Results of the sex-stratified association analysis are summarized in **Table 2**. Overall, six SNPs from three loci reached genome-wide significance (5x10^{-8}). Two correlated (r^2=0.98) intronic SNPs (rs12657394, MAF=18.9%, P_{male}=2.39x10^{-8}, P_{female}=4.10x10^{-3}; rs2914610, MAF=19.2%, P_{male}=4.55x10^{-8}, P_{female}=5.58x10^{-3}) within gene *SRFBP1* (serum response factor binding protein 1) on chromosome 5 were strongly associated with VAT_BMI in males. Three correlated SNPs (rs1002945, rs13247968, and rs1830005, r^2>0.91) downstream of *SNX13* (sorting nexin 13 gene) were significantly associated with VAT_BMI in males but not females (rs13247968, MAF=42.9%, P_{male}=7.63x10^{-9}, P_{female}=0.36).

SNP-sex Interaction Analysis

Results of the SNP-sex interaction are summarized in **Figure 2 and Table 2**. SNP rs9289345 was strongly associated with the interaction variable for VAT_BMI ($P_{int}=3.73x10^{-8}$, MAF=1.3%). It is an intronic SNP locating within gene *TMCC1* (transmembrane and coiled-coil domain family 1). An intronic SNP within gene *PAPPA2* (pappalysin 2) (rs10913233, $P_{int}=3.07x10^{-8}$, MAF=13.8%) was associated with the interaction variable for VSR. SNP rs10923724, located upstream of *TBX15* (T-box 15 gene), was strongly associated with the interaction variable for VSR ($P_{int}=2.89x10^{-8}$, MAF=38.1%).

Assessment of Previously Identified Signals

From the literature, three previously identified loci were associated with CT measures in European-derived populations (Fox, Liu, et al., 2012) and were evaluated herein. The *FTO* (fat mass and obesity-associated gene) was modestly associated with SAT in IRASFS (rs9922619, $P=4.01x10^{-3}$). However, the trait associations observed in Europeans at the *LYPLAL1* (lysophospholipase-like 1 gene) and *THNSL2* (threonine synthase-like 2 gene) loci were not significant in IRASFS (rs11118316, $P_{VSR}=0.68$; rs1659258, $P_{VAT\_female}=0.12$). Instead, rs11118316 was nominally associated with the interaction variable with VSR ($P_{VSR\_INT}=1.19x10^{-2}$). Further analysis of the region (between recombination spots, 219-219.7Mb) in IRASFS did not reveal additional strong association signals. Overall, 21 signals were significant ($P<0.05$) associated with at least one of the four CT phenotypes and 18 of the 21 signals exhibited gender-specific effects ($P<0.05$ for gender stratified or interaction analysis). The strongest signals observed was SNP rs2645294 ($P_{VSR\_INT}=1.20x10^{-5}$). It is located in the 3'-UTR region of the tryptophanyl TRNA synthetase 2 mitochondrial gene (*WARS2*) and has been previously reported to be associated with WHR (Shungin, et al., 2015). However, the previous study revealed no sex specificity for

SNP rs2645294 (Shungin, et al., 2015), suggesting potentially different biology were represented by WHR and VSR.

Discussion

Here we present a study combining genome-wide and exome chip arrays to investigate the genetic determinants of adipose tissue deposition in Mexican Americans from the Insulin Resistance Atherosclerosis Family Study (IRASFS). The assessment of CT measures enabled differentiation between SAT and VAT and provided more refined adiposity phenotypes compared to traditional anthropometric measures. In addition, as SAT and VAT distributions are well known to vary by sex, sex-stratified as well as formal SNP-sex interaction analyses were performed and signals with significant sex-specific effects were identified.

SNP rs2185405, an intronic SNP within gene *GLIS3* (GLIS family zinc finger 3), was found to be genome-wide significant for the main effect analysis with VAT without sex stratification (P_{dom}=1.98x10^{-8}, MAF=40%) (**Figure 3**). *GLIS3* is a member of the GLI-similar zinc finger protein family and encodes a nuclear protein with five C2H2-type zinc finger domains. It functions as both a repressor and activator of transcription and is specifically involved in the development of pancreatic beta-cells, the thyroid, eye, liver and kidney (Tatusova, Ciufo, Fedorov, O'Neill, & Tolstoy, 2015). Previous studies have shown this gene is associated with neonatal and type 1 and 2 diabetes in multiple ethnicities (Awata et al., 2013; Rees et al., 2011; Senee et al., 2006). *In vitro* experiments suggested *GLIS3* was able to modulate pancreatic beta-cell apoptosis via the regulation of a splice variant of the BH3-only protein Bim(Nogueira et al., 2013). Regional fat distribution affects glucose metabolism beyond total body adiposity (Lapidus, et al., 1984; Larsson, et al., 1984) and VAT has been shown to be a risk factor for metabolic disorders (P. Cohen, et al., 2014; Wajchenberg, 2000). Therefore, it is possible that visceral fat is involved in

the modulation of beta-cell function through *GLIS3*. Further evaluation of the SNP using glucose homeostasis traits in IRASFS revealed a significant association with insulin clearance (P=0.03) and a suggestive association signal with fasting insulin (P=0.09) (data not shown).

For sex-stratified analyses, five SNPs from two loci (*SRFBP1*, 7p21.1) reached genome-wide significance. *SRFBP1*, also named as *p49/STRAP*, was associated with VAT_BMI in males. Regional plots indicate long-range linkage disequilibrium (LD) covering a 200kb region (**Figure 4**). The strongest signal in the region was an intronic SNP rs12657394 under a dominant model ($P_{male}=2.39x10^{-8}$; $P_{female}=0.001$; MAF=18.9%). In males, the minor allele carriers on average had a $24.6cm^2$ increase in VAT compared to common allele homozygous individuals. Conditional analysis using rs12647394 as a covariate demolished the association signals (data not shown), suggesting only one independent signal exists in *SRFBP1*. Previous studies have shown that the protein encoded by *SRFBP1* specifically interacts with an acidic amino acid motif in the N-terminus of GLUT4 in adipose cells, suggesting a possible role in biosynthesis and/or processing of GLUT4 in adipocytes (Lisinski et al., 2006). However, no biological evidence has been identified to explain a possible mechanism for sex specificity. Further evaluation of glucose homeostasis traits in IRASFS with and without sex stratification revealed modest association signals for rs12657394 with fasting insulin ($P_{add}=7.92x10^{-3}$), fasting glucose ($P_{rec}=0.025$), $HOMA_{IR}$ (homeostatic model of assessment of insulin resistance) ($P_{add}=0.03$) and $HOMA_B$ (homeostatic model of assessment of β-cell function) ($P_{dom}=0.03$). However, no sex-specific patterns were observed (data not shown). Extended examination of this region in the GIANT consortium with waist circumference (WC), BMI, and WHR stratified by sex failed to replicate this signal (Berndt, et al., 2013; Randall, et al., 2013). This could be due to ethnic heterogeneity as well as the limitations of anthropometric measures compared to CT-derived fat deposition.

On 7p21.1 SNP rs13247968 (MAF=42.5%, $P_{male}=7.63x10^{-9}$, $P_{female}=0.36$) as well as two other highly correlated SNPs ($r^2 \geq 0.9$) were strongly associated with VAT after adjusting for BMI under the dominant model in males. On average, male carriers of the rs13247968 minor allele (G) had a 16.4 cm^2 increase of VAT compared to males homozygous for the major allele. The regional plot (**Figure 5c**) shows a tight cluster of signals located downstream of *SNX13* (Sorting nexin 13 gene, also known as *RGS-PX1*). The RGS (Regulator of G protein) domain of the encoded protein can function as a GTPase-activating protein for G alpha subunits of heterotrimeric G proteins while the PX (Phox) domain works as a sorting nexin protein involved in intracellular trafficking (Teasdale, Loci, Houghton, Karlsson, & Gleeson, 2001; Zheng et al., 2001). Studies have shown that the protein can target lysosomes and delay lysosomal degradation of the EGFR (Epidermal growth factor receptor) (Zheng, et al., 2001). Further examination of this region in the GIANT consortium with WC, BMI, and WHR stratified by sex identified rs1990467, 100kb proximal to the locus, associated with WC_BMI in males ($P=2.6x10^{-4}$) (Berndt, et al., 2013; Randall, et al., 2013). However, the correlation between rs13247968 and rs1990467 was poor ($r^2=0$ in IRASFS,). Previous genetic studies have suggested *SNX13* is strongly associated with HDL cholesterol in European ancestry individuals (Willer, et al., 2013). Evaluation of SNP rs13247968 with circulating HDL, LDL, triglyceride, cholesterol levels and BMI in IRASFS failed to detect significant associations ($P>0.05$; data not shown). Interestingly, this locus is 300kb upstream of histone deacetylase 9 gene (*HDAC9*), which encodes an important histone deacetylase that regulates transcriptional regulation, cell cycle progression, and developmental events (X. Zhou, Marks, Rifkind, & Richon, 2001). Genetic studies have suggested *HDAC9* is associated with multiple phenotypes including coronary artery disease (CAD), BMI, and vigorous physical activity in European and Hispanic populations (Comuzzie, et al., 2012; Dichgans et al., 2014).

Genome-wide SNP-sex interaction analyses revealed three genome-wide significant signals.

PAPPA2 (pappalysin 2 or pregnancy-associated plasma protein A2) located on chromosome 1 (rs10913233, $P_{Add}=3.07x10^{-8}$, MAF=13.8%) was associated with the SNP-sex interaction effect for VSR (**Figure 6**). The protein encoded by *PAPPA2* is proposed to be a biomarker for preeclamptic placentae in pregnant women (Crosley et al., 2014). It works as a protease specifically cleaving insulin-like growth factor binding protein 5 (IGFBP5) and thus plays an important role in regulation IGFBP5 levels (Overgaard et al., 2001). Diseases related to this gene include HELLP-syndrome and developmental dysplasia of hip (Buimer et al., 2008; Jia et al., 2012). Human population genetic studies have suggested this gene is associated with height (Lango Allen et al., 2010; Lettre et al., 2008). Previous mice studies have identified this gene to be associated with body size and weight, bone size and shape, postnatal growth retardation with more pronounced phenotypes in females compared to males (Blake JA, 2014; Christians, de Zwaan, & Fung, 2013; Conover et al., 2011). Further examination of this region in the GIANT Consortium with WC, BMI, and WHR stratified by sex identified modest signals for WHR and WHR_BMI in females (rs10913190, $P_{WHR}=5.6x10^{-4}$, rs10913282 $P_{WHR\_BMI}=8.6x10^{-4}$; $r^2<0.01$ with rs12090061 in IRASFS).

SNP rs10923724 was significantly associated with the SNP-sex interaction variable with VSR ($P_{add}=2.89x10^{-8}$, MAF=38%) (**Figure 7**). Sex-stratified analysis revealed rs10923724 was nominally associated in males and females with opposite directions of effect ($P_{male}=3.84x10^{-4}$, $\beta_{male}=-0.97$; $P_{female}=2.31x10^{-3}$, $\beta_{female}=0.10$). It is an intergenic SNP locating at 1p12, 14kb upstream of *TBX15* and 27kb downstream of *WARS2*. *WARS2* (tryptophanyl tRNA synthetase 2) is one of the two isoforms of mitochondrial aminoacyl-tRNA synthetases that catalyzes the aminoacylation of tRNA (Tatusova, et al., 2015). *TBX15*, or T-box 15, is a member of the T-box family. The family encodes a phylogenetically conserved family of transcription factors that regulate a variety of developmental processes (Tatusova, et al., 2015). Diseases related to the gene product include Cousin syndrome and congenital heart malformations (Dikoglu et al., 2013; Hu et al., 2013).

TBX15 has also been indicated to be involved in the pathophysiology of placental diseases (Chelbi et al., 2011). Interestingly, this locus has been one of the top WHR genetic signals. SNP rs2645294 ($P=1.7x10^{-19}$) and rs984222 ($P=8.69x10^{-25}$) have been previously identified with associations with WHR with no sex-specific patterns(Heid, et al., 2010; Shungin, et al., 2015). In IRASFS, the two SNPs were highly correlated ($r^2=0.87$) with strong associations with the interaction variable of VSR (rs2645294, $P_{VSR\_int}=1.2x10^{-5}$; rs984222, $P_{VSR\_int}=2.63x10^{-7}$). However, they were not associated with WHR (P>0.1). SNP rs10923724 has nominal correlations with the previous two SNPs ($r^2<0.58$) and was not associated with WHR (P=0.6). Analysis of rs10923724 conditioned by rs2645294 and rs984222 revealed nominal association (P=0.04) in males and no association in females (P=0.98). These results suggest there are multiple signals in the region with and without sex-specific heterogeneity. Taken together, 1p12 is an interesting locus with complicated signals combing sex-specific and non-sex-specific mechanisms.

Although significant signals have been identified, study limitations do exist. First, the majority of the SNPs analyzed were from statistical imputation as opposed to direct genotyping. Although SNP genotypes were highly concordant between platforms, the imputation quality was reduced for rare variants with a MAF less than 1%, adding uncertainties to the low-frequency SNP results. Therefore, these results focus on only SNPs with MAF greater than 1%. Another limitation was the small number of available individuals. This largely limited the study power, especially for the analysis of rare variants, i.e. exome chip, and sex-stratified analyses. This may explain why no significant signals were detected with the exome chip platform. In addition, limited biological evidence was found to support sex-specific signals identified by genetic studies. It is likely attributed to the small number of biological studies that have been focused on sex-specific mechanisms of adipose deposition. Also the application of Illumina HumanExome Array in Mexican Americans is debatable as this platform may not be optimal for Mexican ancestry as

only one third of the SNPs on the array were polymorphic, likely attributable to an array design biased towards Caucasians and African Americans. The application of Illumina Omni array has similar concerns in that SNPs on the array may not tag the LD structure as well in Mexican Americans.

Increasing evidence has supported disease susceptibility heterogeneity related to adipose tissue depots: subcutaneous adipose tissue is more benign while visceral adipose is more correlated with risk of metabolic disease (P. Cohen, et al., 2014). The use of CT scans has enabled a more direct estimate of regional adiposity, i.e. differentiate between subcutaneous and visceral adipose tissue, compared to anthropometric measures. In addition, CT scans are less prone to user bias as all CT scans were conducted under the same protocol and results were sent to a centralized reading center. In contrast, anthropometric measures can often be biased by factors including age, sex, clinical site, etc. (Taylor, et al., 2000). However, CT measures can include bias as well. For example, females predominantly store body fat in the gluteal-femoral region while males store fat in abdominal region (Blaak, 2001). However, CT measures were obtained by axial slices at the L4-L5 disc space in IRASFS and therefore no gluteal adipose tissue was measured. This may explain why our study observed fewer signals in females as compared to males.

In summary, we computed a combined study of genome-wide and exome chip arrays in the IRASFS Mexican cohort. Adiposity phenotypes included were SAT, VAT, VSR, and VAT_BMI. Sex stratification and formal SNP-sex interaction analyses were conducted to search for sex-specific signals. The conclusions supported a genetic basis to the differential mechanism of adipose tissue distribution by sex. Moreover, these results highlighted the importance of using more refined measures of adiposity (CT scans) as well as the added utility of research in minority populations where an increased prevalence of adiposity-related diseases may be associated with a different genetic architecture.

Acknowledgments and funding information:

This research was supported by DK097524 from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) and HG007112 from the national Human Genome Research Institute (NHGRI). The provision of GWAS genotyping data was supported in part by UL1TR000124 (CTSI) and DK063491 (DRC). Exome chip genotyping was supported by DK081350, HG007112, DK087914, and the Department of Internal Medicine at the University of Michigan. Computational resources were provided, in part, by the Wake Forest School of Medicine Center for Public Health Genomics. The authors would like to acknowledge the members of the GUARDIAN Consortium with research supported by DK085175 from NIDDK and from the IRASFS grants (HL060944 and HL061019).

Ethical statement:

Participants included in this study were recruited from clinical centers in San Antonio, TX and San Luis Valley, CO. The Institutional Review Board of each clinical (UT Health Science Center San Antonio Review Board and Colorado Multiple Institutional Review Board, respectively) and analysis (Wake Forest School of Medicine) site approved the study protocol and all participants provided their written informed consent.

Table 1. Demographic characteristics of the study population.

	GWAS			Exome Chip		
	overall	male	female	overall	male	female
N	983	403	580	1205	498	707
BMI (kg/m^2)	28.3±5.8	27.9±5.1	28.5±6.2	28.9±6.1	28.4±5.4	29.2±6.6
Age (years)	40.6±13.6	39.5±14.1*	41.4±13.3*	42.6±14.6	41.6±14.9*	43.4±14.3*
SAT (cm^2)	329.7±151.7	263.7±129.2†	375.7±149.4†	339.1±154.7	272.9±132.2†	386.1±152.1†
VAT (cm^2)	106.6±57.0	120.3±59.9†	97.1±52.9†	114.0±61.3	127.3±63.7†	104.8±57.8†
VSR	0.35±0.21	0.49±0.22†	0.25±0.13†	0.36±0.22	0.50±0.23†	0.27±0.15†

Significant difference was observed between sexes: * ($P<0.001$), † ($P<0.0001$).

Table 2. Genome-wide significant signals from SNP association analyses

SNP[a]	Chr:Pos	Gene	Alleles[e]	RAF[f]	P_overall (N=983)	Beta	P_female (N=580)	Beta	P_male (N=403)	Beta	P_interaction (N=983)
Visceral Adipose Tissue (VAT)											
rs2185405	9:4078851	*GLIS3*	T/C	0.40	**1.98E-08[d]**	0.90	1.36E-04[d]	0.76	1.10E-05[d]	1.11	0.69[d]
Visceral Adipose Tissue adjusted by BMI (VAT_BMI)											
rs12657394	5:121308199	*SRFBP1*	A/G	0.19	3.68E-07[c]	0.62	4.10E-03[b]	0.35	**2.39E-08[c]**	1.09	1.10E-03[c]
rs2914610	5:121314168	*SRFBP1*	A/G	0.19	9.78E-07[c]	0.59	5.58E-03[b]	0.34	**4.55E-08[c]**	1.07	1.53E-03[c]
rs1002945	7:17796659	*AHR-SNX13*	A/T	0.43	2.80E-05[c]	0.50	0.33[b]	0.10	**1.17E-08[c]**	1.09	1.37E-04[c]
rs13247968	7:17806817	*AHR-SNX13*	G/T	0.42	1.30E-05[c]	0.53	0.36[b]	0.10	**7.63E-09[c]**	1.11	1.59E-04[c]
rs1830005	7:17807563	*AHR-SNX13*	C/T	0.43	6.72E-06[c]	0.55	0.18[b]	0.15	**3.67E-08[c]**	1.07	8.70E-04[c]
rs9289345	3:129579506	*TMCC1*	G/A	0.01	0.51[b]	0.35	0.017[c]	1.40	2.90E-04[c]	-4.59	**3.73E-08[b]**
Visceral Subcutaneous Adipose Ratio (VSR)											
rs10913233	1:176625427	*PAPPA2*	T/A	0.14	0.68[b]	-0.011	4.94E-03[b]	-0.087	4.66E-03[c]	0.13	**3.07E-08[b]**
rs10923724	1:119546842	*TBX15/WARS2*	C/T	0.38	0.68[b]	-0.0078	2.30E-03[c]	0.099	3.84E-04[b]	-0.098	**2.89E-08[b]**

[a]SNP in build GRCh37/hg19; [b]Additive model; [c]Dominant model; [d]Recessive model; [e]Minor/Major allele; [f]Reference allele based on minor allele.

Figure 1. Manhattan plots for association analysis in IRASFS Mexican Americans: **A**. Subcutaneous adipose tissue (SAT), **B**. Visceral adipose tissue (VAT), **C**. Visceral adipose tissue adjusted for body mass index (VAT_BMI), **D**. Visceral-subcutaneous adipose tissue ratio (VSR). Results were adjusted for age, sex, recruitment center (San Antonio, TX or San Luis Valley, CO), and admixture estimates. P-values are shown under the best fit model. The lower line at $-\log_{10}(PVAL)=5$ represents the best P-value=1.00×10^{-5} and the upper line represents the genome-wide significance threshold ($P=5.00 \times 10^{-8}$).

Figure 1a.

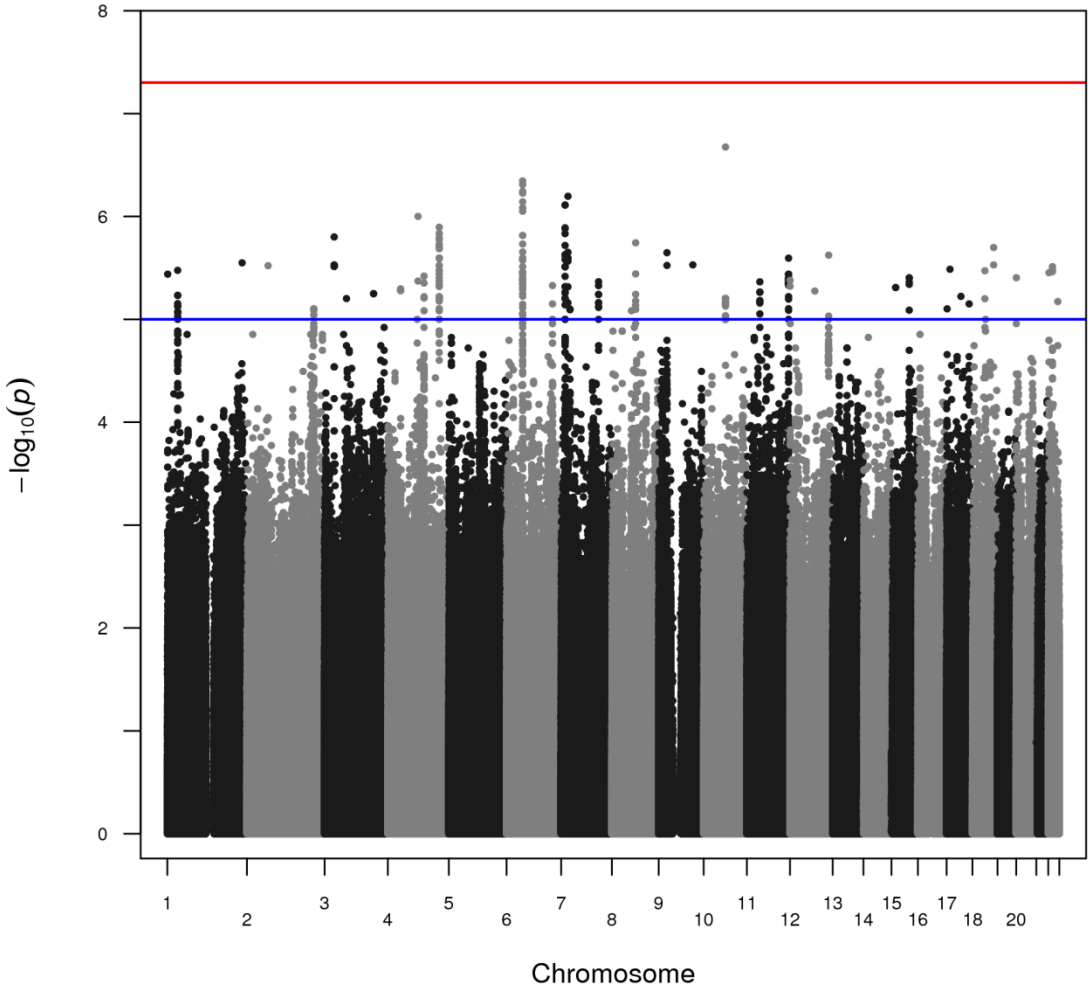

8
6
4
2
0
$-\log_{10}(p)$
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
20
Chromosome

Figure 1b.

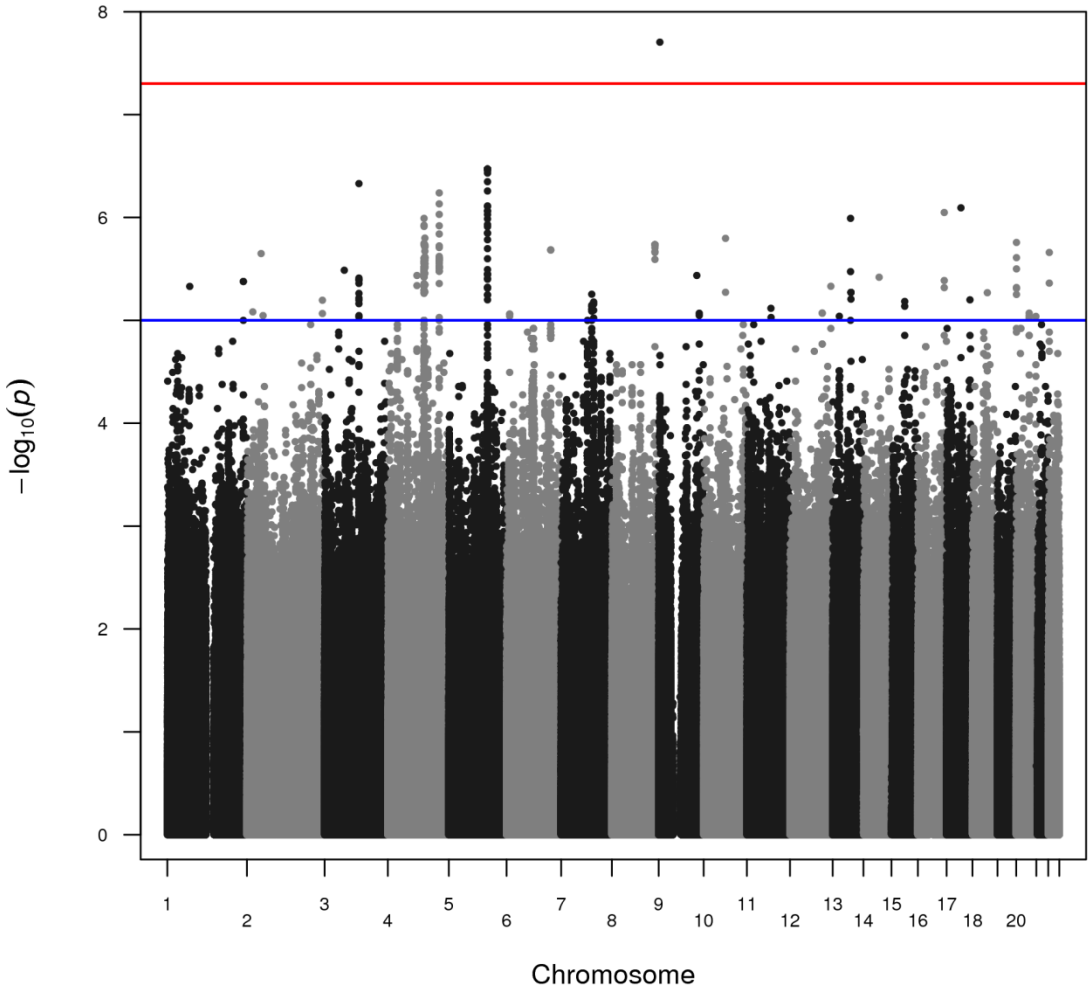

8
6
4
2
0
$-\log_{10}(p)$
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
20
Chromosome

Figure 1c.

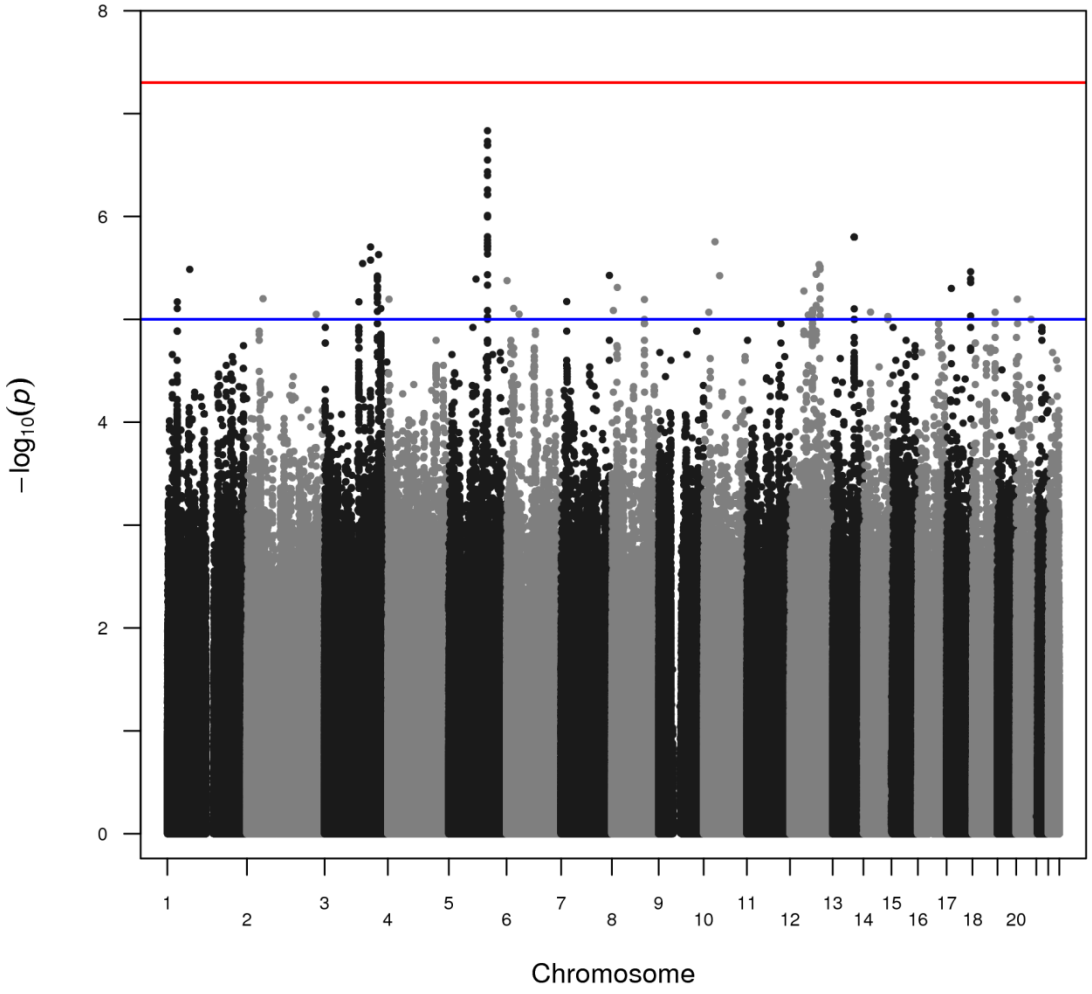

8
6
4
2
0
$-\log_{10}(p)$
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
20
Chromosome

Figure 1d.

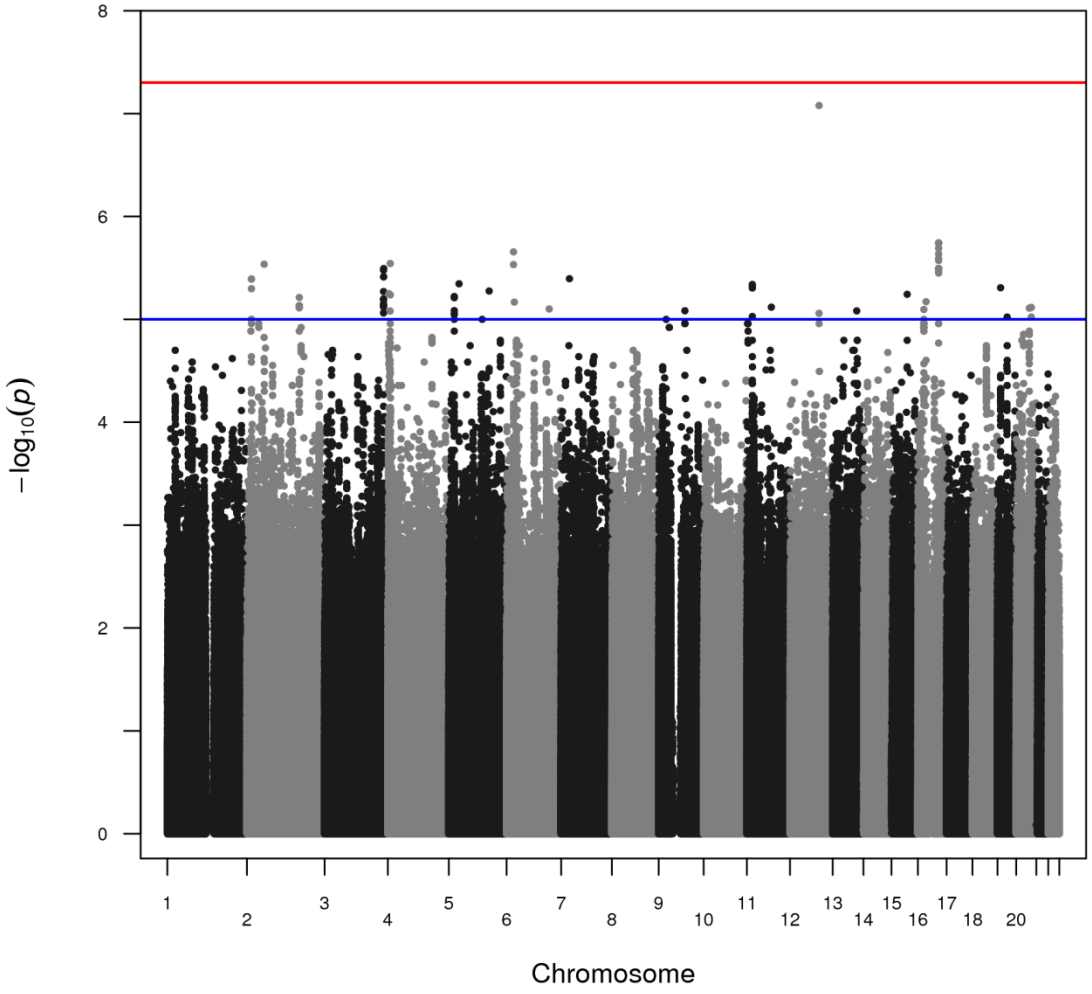

8
6
4
2
0
$-\log_{10}(p)$
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
20
Chromosome

Figure 2. Manhattan plots for genome-wide SNP-sex interaction analysis in IRASFS Mexican Americans: **A**. Subcutaneous adipose tissue (SAT), **B**. Visceral adipose tissue (VAT), **C**. Visceral adipose tissue adjusted for body mass index (VAT_BMI), **D**. Visceral-subcutaneous adipose tissue ratio (VSR).

Figure 2a.

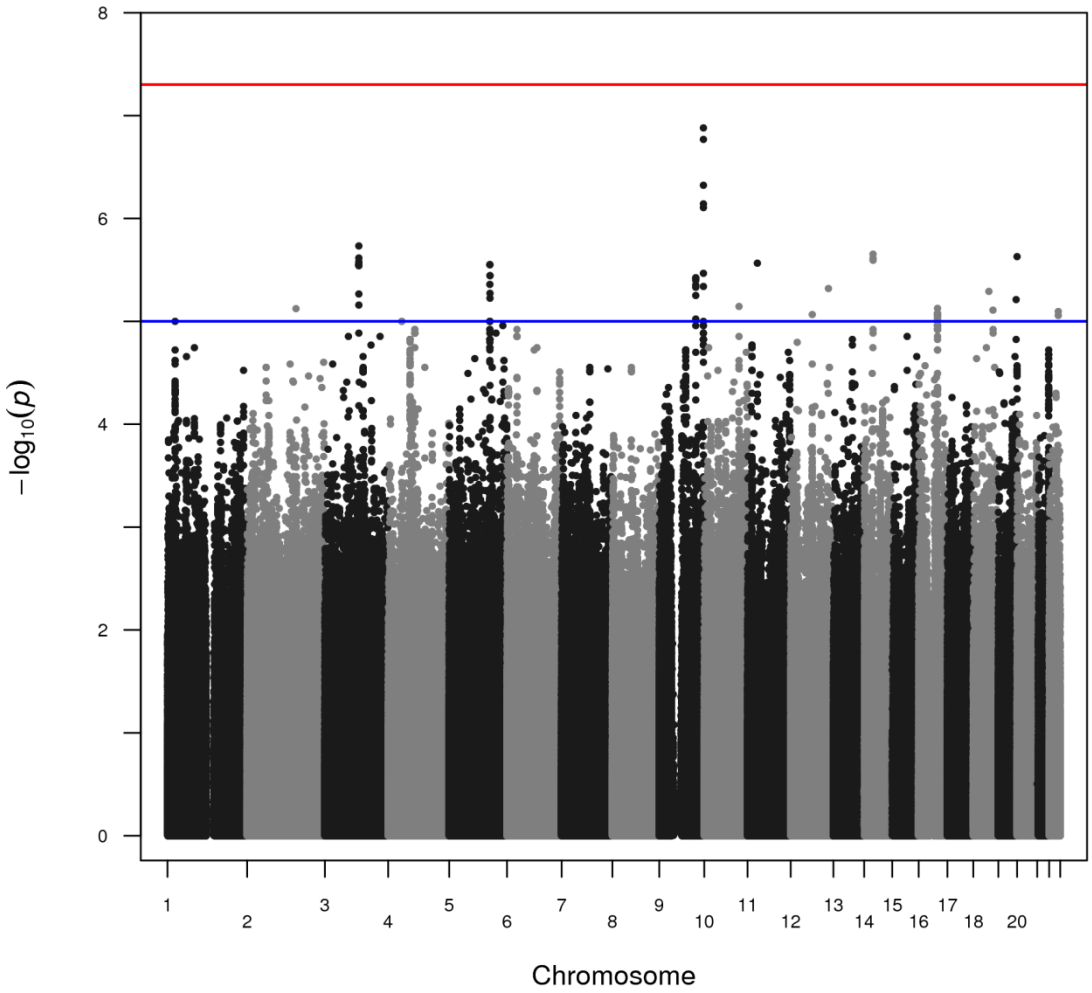

8
6
4
2
0
$-\log_{10}(p)$
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
20
Chromosome

Figure 2b

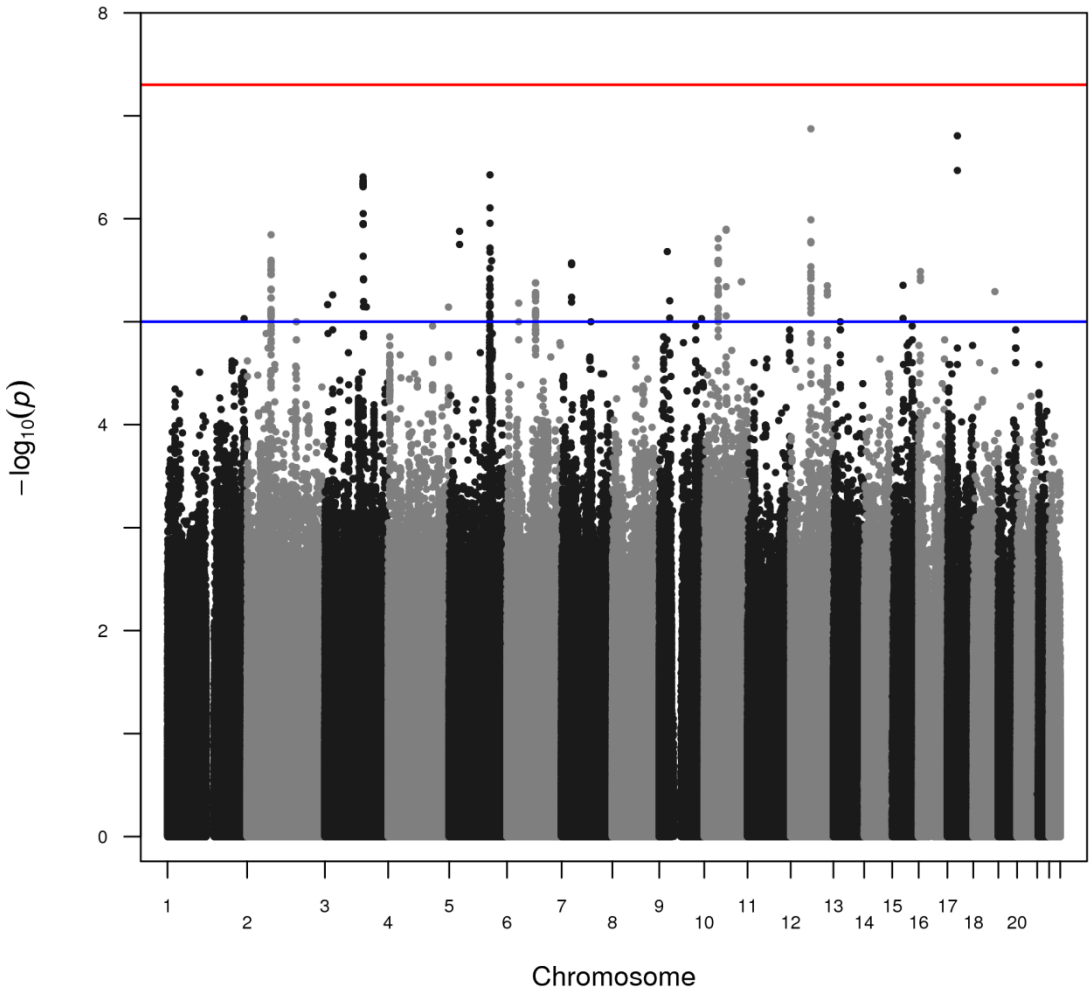

8
6
4
2
0
$-\log_{10}(p)$
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
20
Chromosome

Figure 2c

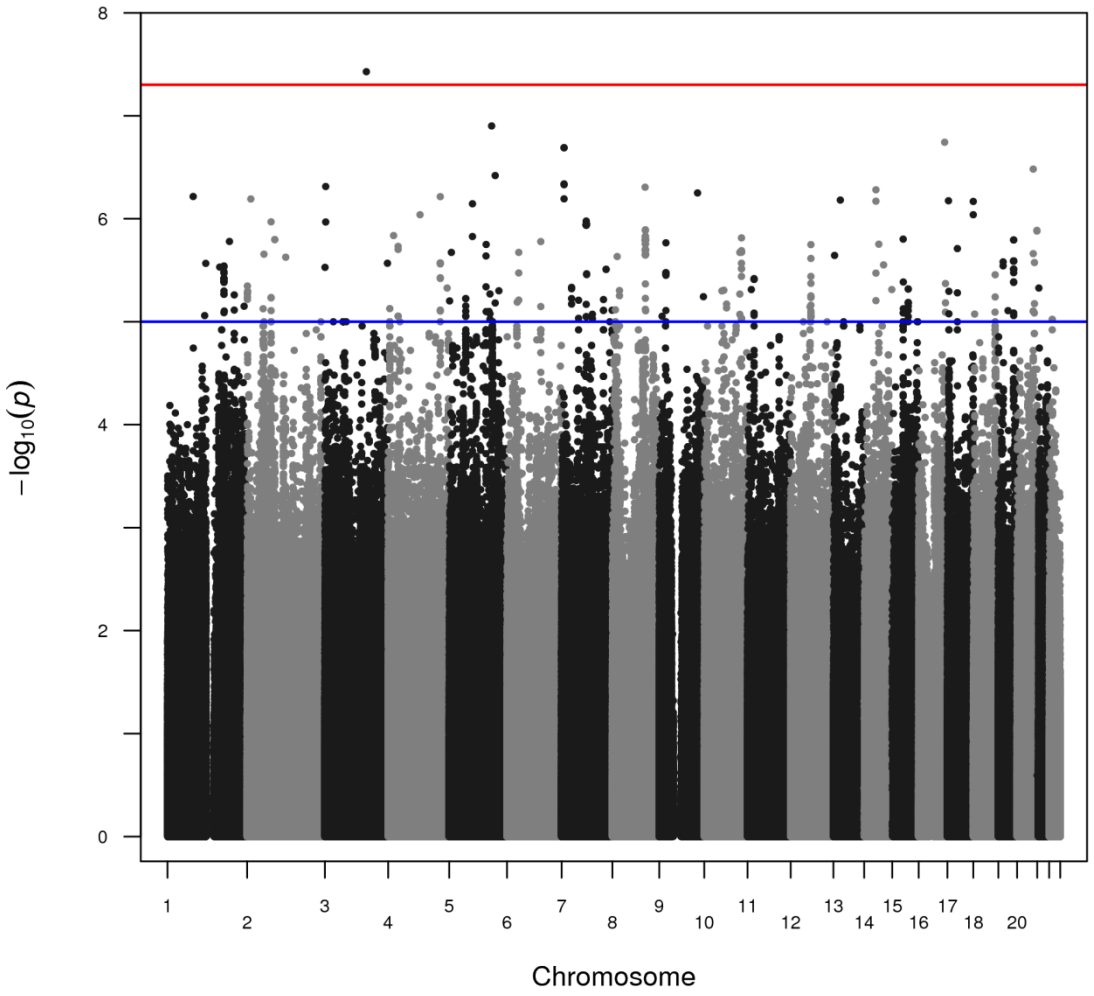

8
6
4
2
0
$-\log_{10}(p)$
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
20
Chromosome

Figure 2d

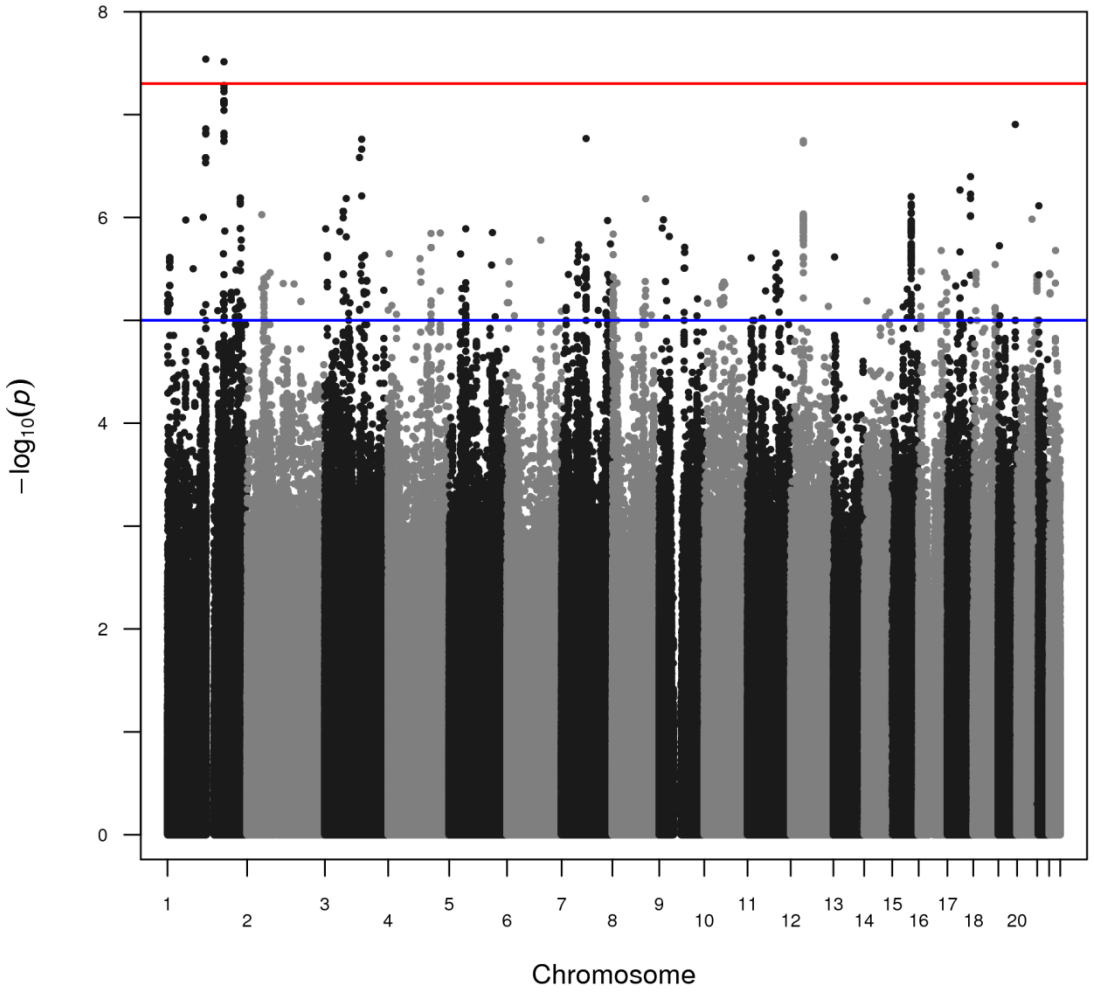

8
6
4
2
0
$-\log_{10}(p)$
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
20
Chromosome

Figure 3. Regional plots of *GLIS3* for association with visceral adipose tissue (VAT) in IRASFS Mexican Americans combining genome-wide and exome chip datasets.

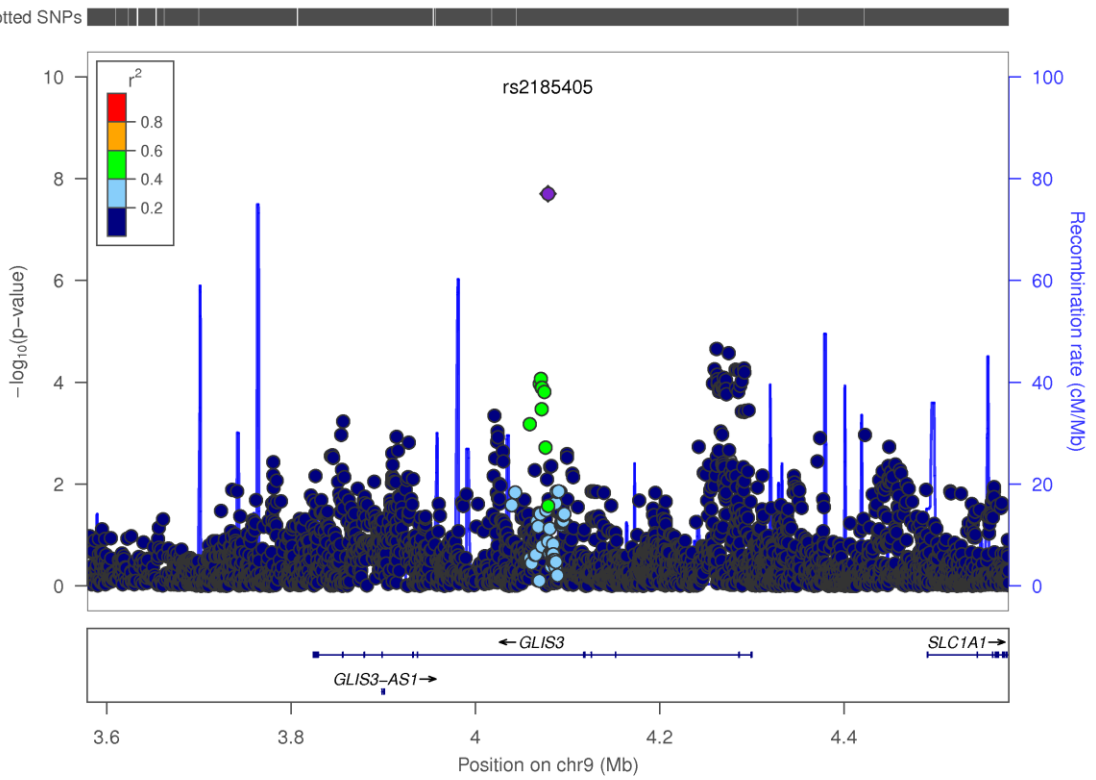


Figure 4. Regional plots of *SRFBP1* for association with visceral adipose tissue adjusted by BMI in IRASFS Mexican Americans combining genome-wide and exome chip datasets. **A.** without sex stratification, **B.** females only, **C.** males only.

Figure 4a

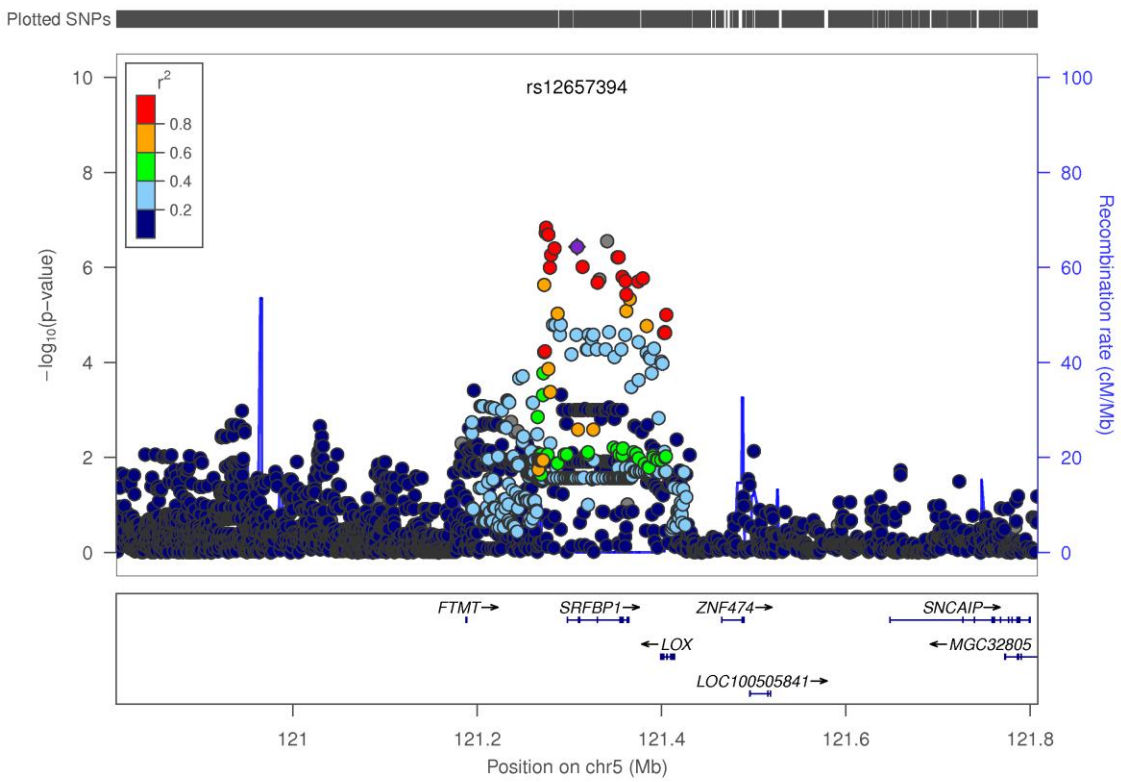

Plotted SNPs
rs12657394
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
FTMT
SRFBP1
ZNF474
SNCAIP
LOX
MGC32805
LOC100505841
Position on chr5 (Mb)

Figure 4b

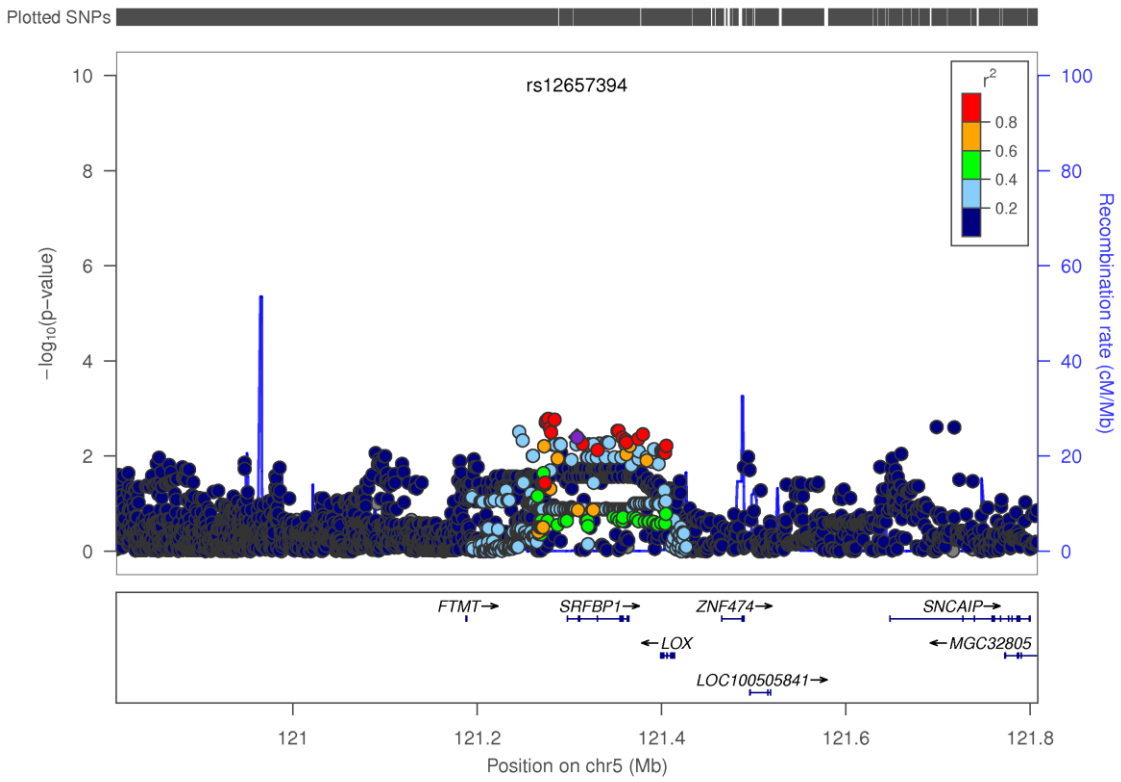

Plotted SNPs
rs12657394
r2
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
FTMT
SRFBP1
ZNF474
SNCAIP
LOX
MGC32805
LOC100505841
Position on chr5 (Mb)

Figure 4c

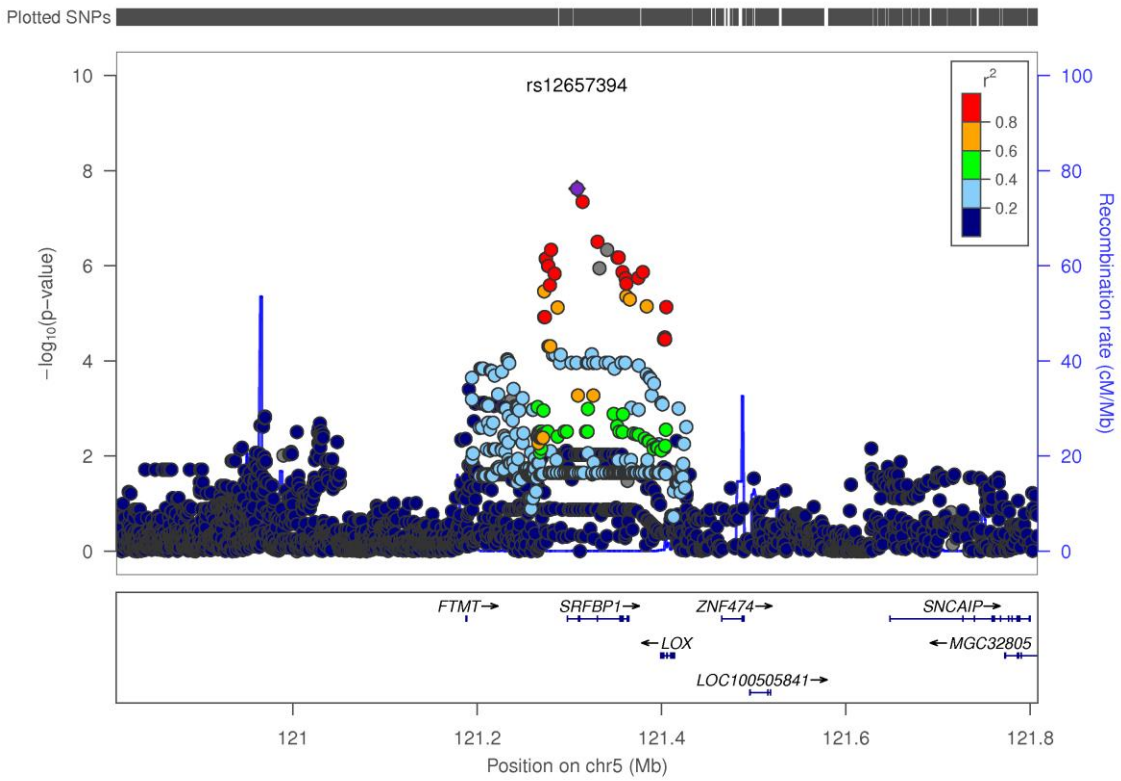

Plotted SNPs
rs12657394
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
FTMT
SRFBP1
ZNF474
SNCAIP
LOX
MGC32805
LOC100505841
Position on chr5 (Mb)

Figure 5. Regional plots of the rs13247968 locus for association with visceral adipose tissue adjusted by BMI in IRASFS Mexican Americans combining genome-wide and exome chip datasets. **A.** without sex stratification, **B.** females only, **C.** males only.

Figure 5a

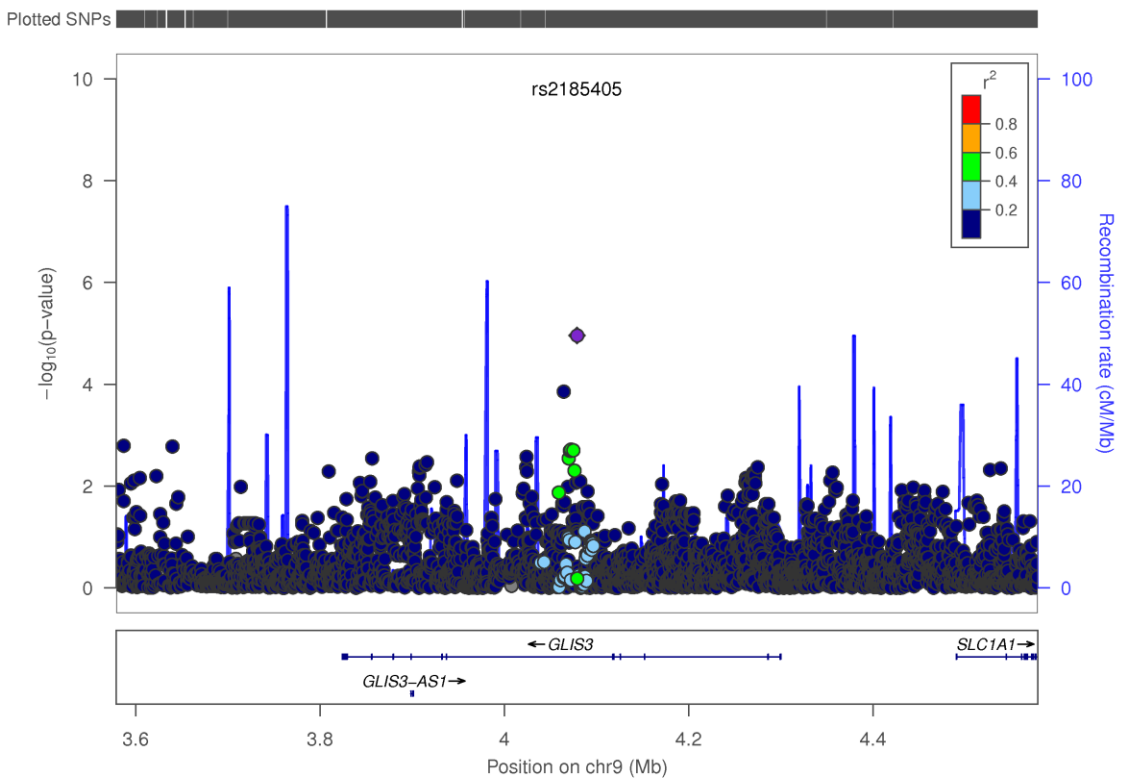

Plotted SNPs
rs2185405
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
GLIS3
SLC1A1
GLIS3−AS1
Position on chr9 (Mb)

Figure 5b

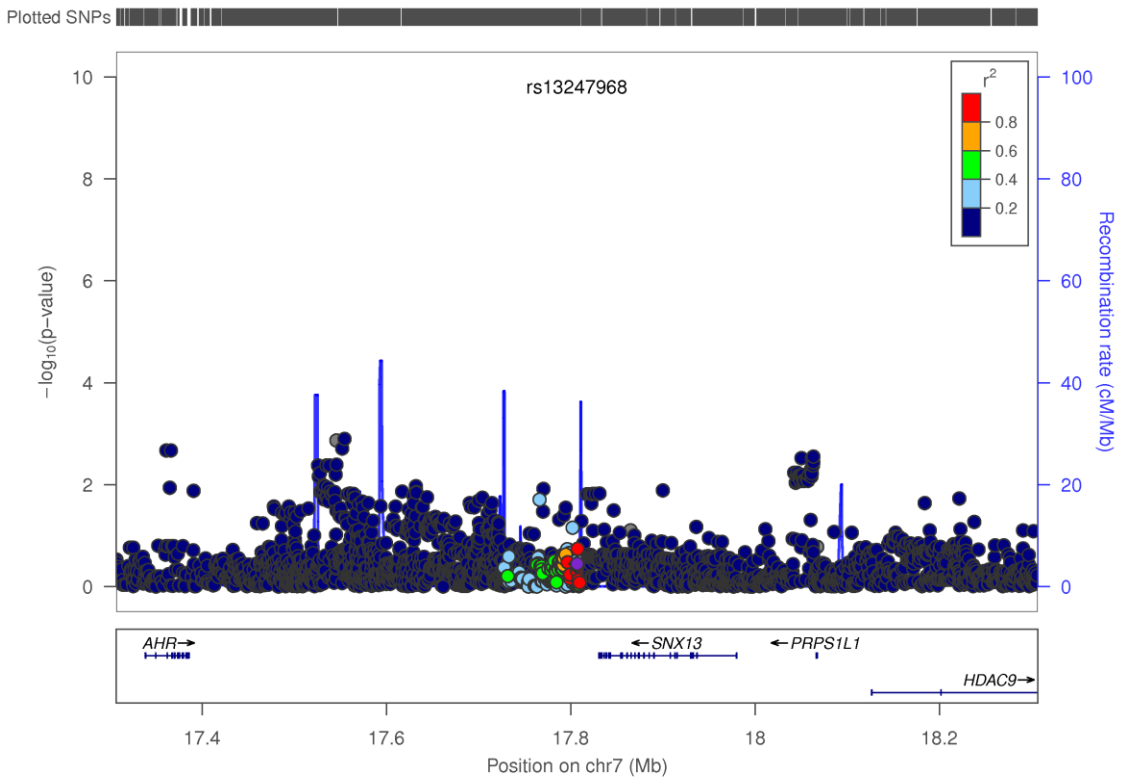

Plotted SNPs
rs13247968
r2
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
AHR
SNX13
PRPS1L1
HDAC9
Position on chr7 (Mb)

Figure 5c

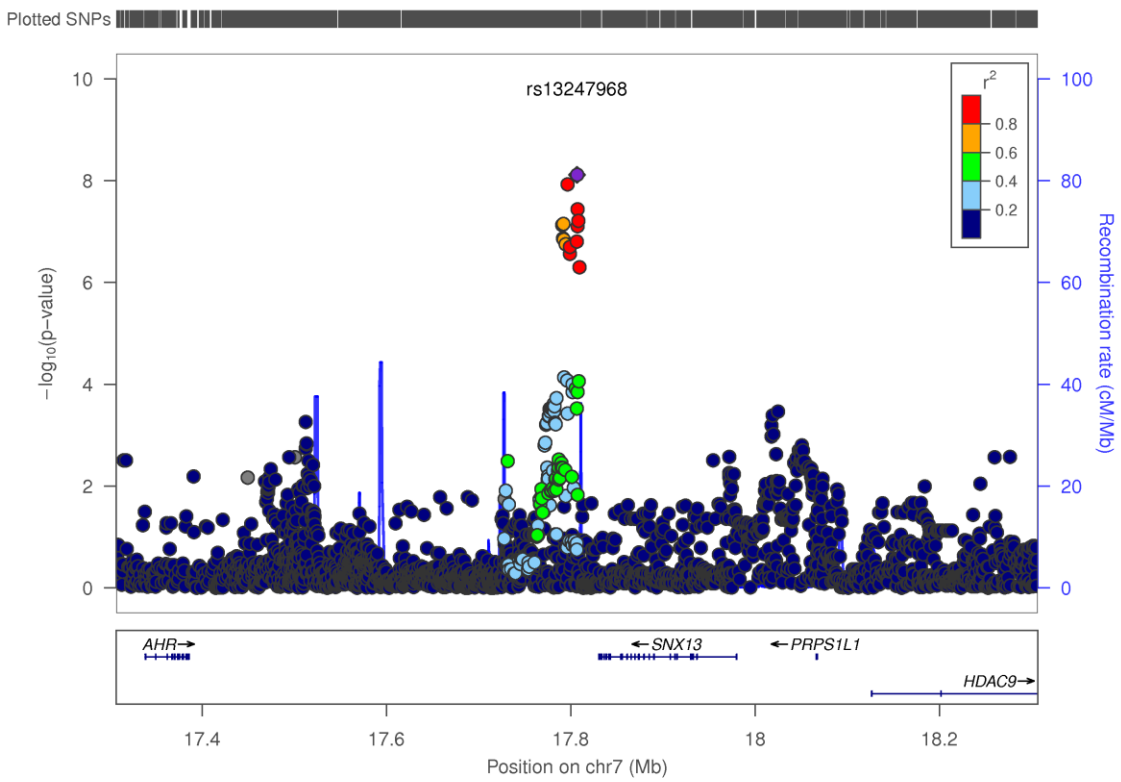

Plotted SNPs
rs13247968
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
AHR
SNX13
PRPS1L1
HDAC9
Position on chr7 (Mb)

Figure 6. SNP rs10913233 was associated with the SNP-sex interaction variable for VSR. **A.** Regional plots of the rs10913233 locus for interaction analysis, **B.** genotypic means of rs10913233 SNP-sex interaction analysis results stratified by sex.

Figure 6a

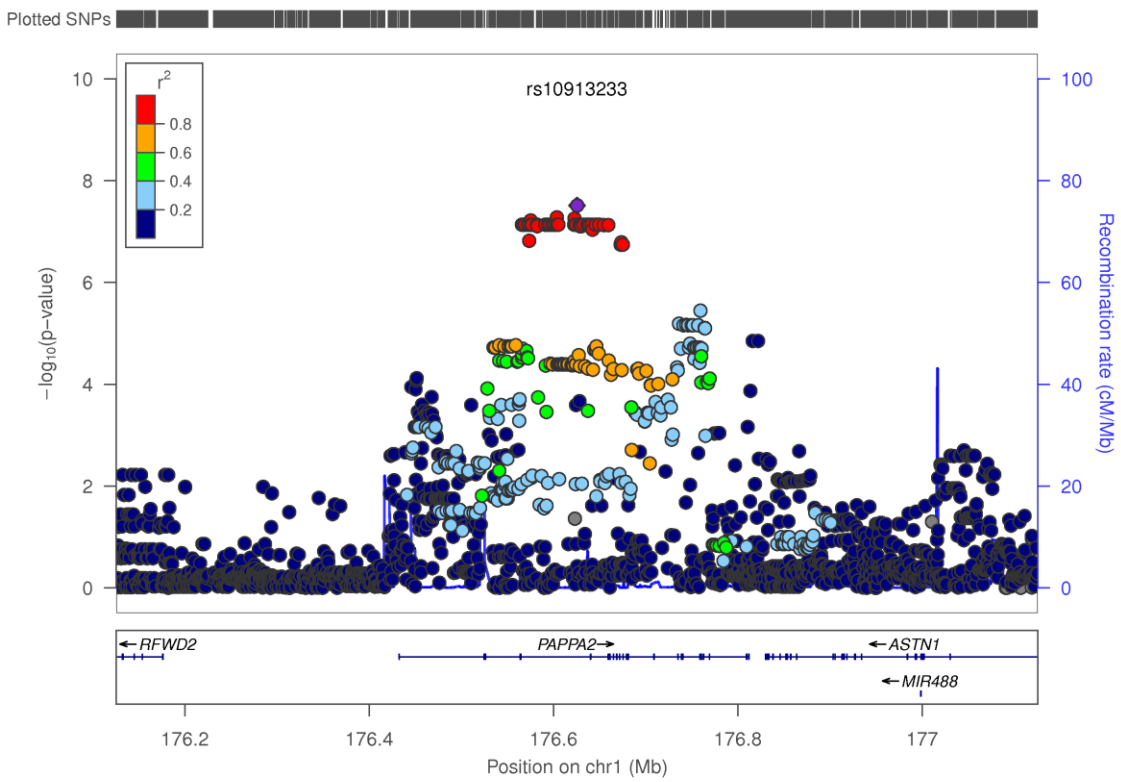

Plotted SNPs
rs10913233
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
RFWD2
PAPPA2
ASTN1
MIR488
176.2
176.4
176.6
176.8
177
Position on chr1 (Mb)

Figure 6b

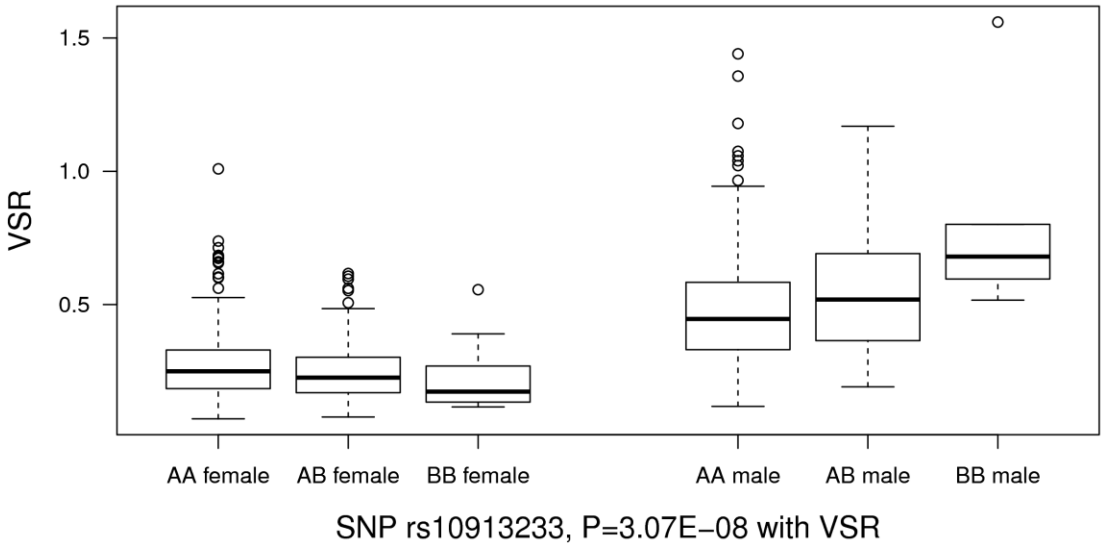

VSR
1.5
1.0
0.5
AA female
AB female
BB female
AA male
AB male
BB male
SNP rs10913233, P=3.07E−08 with VSR

Figure 7. SNP rs10923724 was associated with the SNP-sex interaction variable for VSR. **A.** Regional plots of the rs10923724 locus for interaction analysis, **B.** genotypic means of rs10923724 SNP-sex interaction analysis results stratified by sex.

Figure 7a

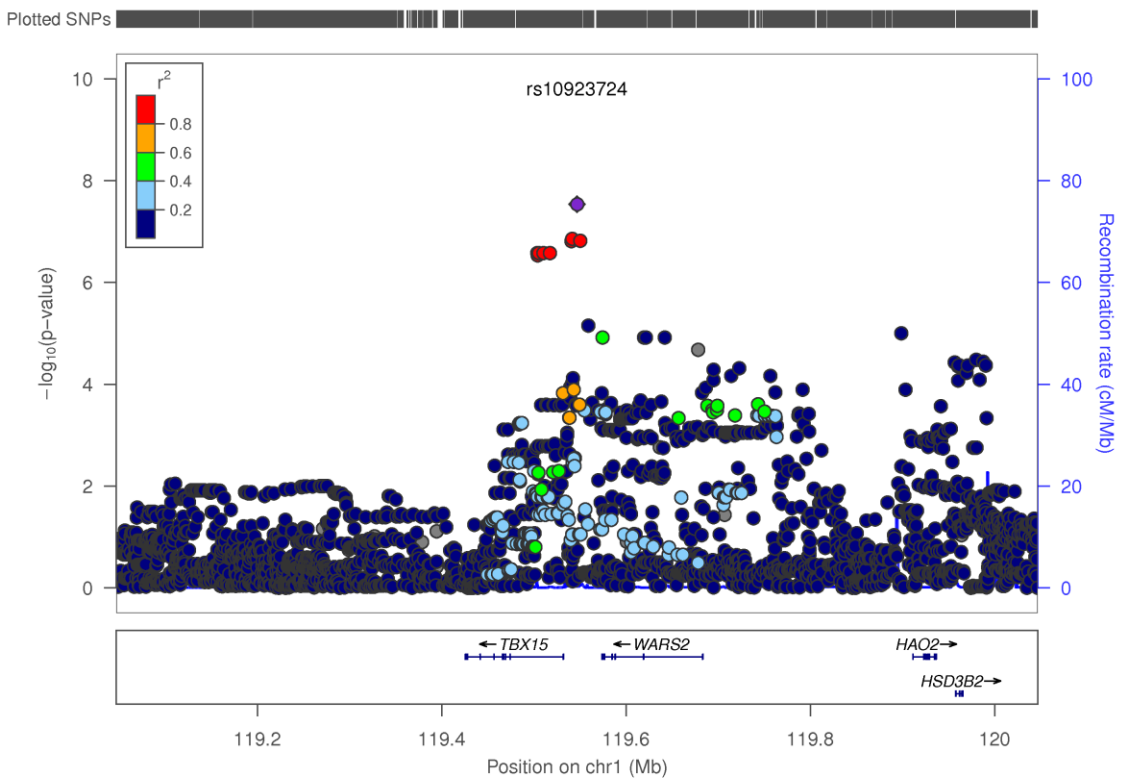
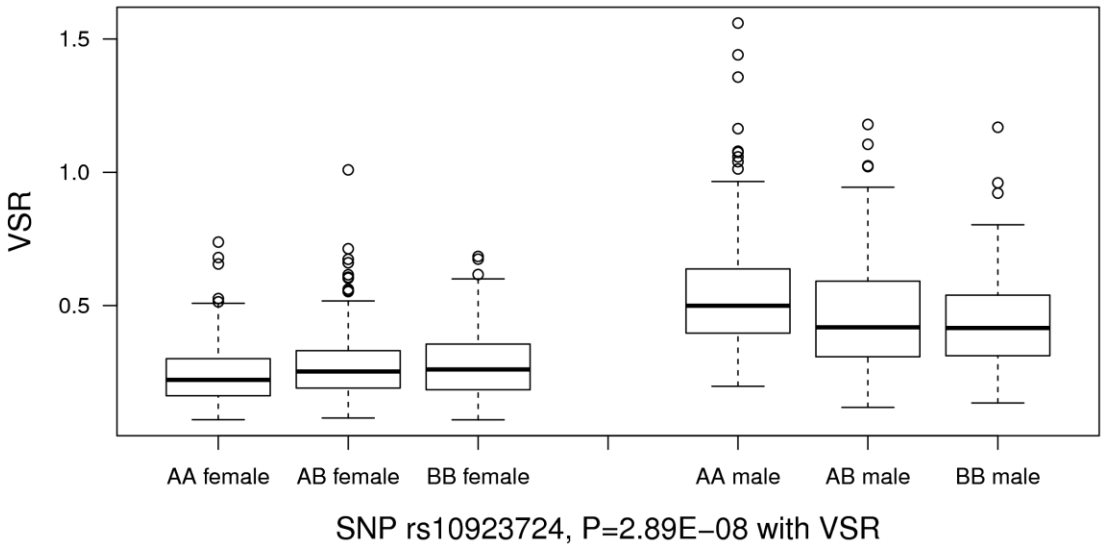

Plotted SNPs
rs10923724
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
TBX15
WARS2
HAO2
HSD3B2
Position on chr1 (Mb)

Figure 7b


1.5
1.0
0.5
VSR
AA female
AB female
BB female
AA male
AB male
BB male
SNP rs10923724, P=2.89E−08 with VSR

Chapter IV

Exome Sequencing Identifies Genetic Variants Associated with Circulating Lipid Levels in Mexican Americans: The Insulin Resistance Atherosclerosis Family Study (IRASFS)

Chuan Gao, Keri L. Tabb, Ph.D., Latchezar M. Dimitrov, MSc, Kent D. Taylor, Ph.D., Nan Wang, Ph.D., Xiuqing Guo, Ph.D., Jirong Long, Ph.D., Jerome I. Rotter, M.D., Richard M. Watanabe, Ph.D., Joanne E. Curran, Ph.D., John Blangero, Ph.D., Carl D. Langefeld, Ph.D., Donald W. Bowden, Ph.D., Nicholette D. Palmer, Ph.D.

The following manuscript will be submitted to Circulation Cardiovascular Genetics. Stylistic variations are due to the requirements of the journal. Chuan Gao performed the analyses and prepared the manuscript. Dr. Nicholette Allred acted in an advisory and editorial capacity.

Abstract:

Background:

Genome-wide association studies have identified numerous variants associated with lipid levels; yet, the majority are located in non-coding regions with unclear mechanisms.

Methods and Results:

Exome sequencing of 1,205 Mexican Americans (90 pedigrees) from the Insulin Resistance Atherosclerosis Family Study (IRASFS) was evaluated with four lipid measures. Heritability estimates in IRASFS suggested a strong genetic basis: low-density lipoprotein (LDL, h^2=0.50), high-density lipoprotein (HDL, h^2=0.57), total cholesterol (TC, h^2=0.53), and triglyceride (TG, h^2=0.42) levels. Association and linkage analyses were performed for 548,889 variants. One genome-wide significance signal was detected intronically in the apolipoprotein A-V gene (*APOA5*) with TG (rs651821, $P_{TG}=3.67x10^{-10}$, LOD_{TG}=2.36, MAF=14.2%). In addition, two correlated SNPs (r^2=1.0) rs189547099 ($P_{TG}=6.31x10^{-08}$, LOD_{TG}=3.12, MAF=0.50%) and chr4:157997598 ($P_{TG}=6.31x10^{-08}$, LOD_{TG}=3.12, MAF=0.50%) reached exome-wide significance ($P<9.11x10^{-08}$). rs189547099 is an intronic SNP in the folliculin interacting protein 2 gene (*FNIP2*). SNP chr4:157997598 is intronic in the glycine receptor beta gene (*GLRB*). Linkage analysis revealed 46 SNPs with a LOD>3 with the strongest signal at rs1141070 (LOD_{LDL}=4.30, P_{LDL}=0.33, MAF=21.6%) in the DNA fragmentation factor subunit beta gene (*DFFB*). In addition, fifty three nominally associated variants ($P<5.00x10^{-05}$, MAF≥1.0%) were selected for replication in six Mexican-American cohorts (N=3,280). The strongest signal observed was a synonymous variant (rs1160983, $P_{LDL}=4.44x10^{-17}$, MAF=2.7%) in the translocase of outer mitochondrial membrane 40 gene (*TOMM40*). Beyond primary findings, previously reported lipid loci were fine-mapped using exome sequencing in IRASFS.

Conclusions:

Multiple association and linkage signals were identified with lipids. These results support that exome sequencing complements and extends insight into the genetics of lipid levels.

Key words:

Genome-wide association study; exome sequencing; Mexican Americans; Lipids

Introduction

Cardiovascular disease (CVD) is the leading cause of death worldwide (Libby, 2002). In the United States, CVD accounts for more deaths than any other major cause of death and, on average, 2,200 Americans die of CVD each day (Mozaffarian et al., 2016). While the exact mechanism of disease remains unclear, lipid concentrations are well-accepted as major risk factors as well as clinical indicators for CVD. Genetic studies have suggested strong heritability for circulating lipid levels, e.g. total cholesterol (TC), LDL cholesterol (LDL), HDL cholesterol (HDL), and triglycerides (TG). Based on a European twin-pairs study, it is estimated that circulating lipid heritability ranges from 0.58 to 0.66 (H_{HDL}=0.61, H_{LDL}=0.59, H_{TC}=0.58, H_{TG}=0.66) (Knoblauch, et al., 1997).

Given the public health relevance as well as the strong genetic component, numerous genome-wide association studies (GWAS) have been performed to investigate the genetic architecture of circulating lipid levels. The most recent Global Lipid Genetics Consortium (GLGC) analyzed 188,577 individuals from four ethnicities (Europeans, East Asians, South Asians, and Africans) and identified 157 loci associated with plasma lipid traits (Willer, et al., 2013). While well-powered, these efforts have largely overlooked the fastest growing US minority population, Hispanic Americans. Compared to non-Hispanic whites, Hispanics suffer even higher risks for CVD, i.e. 32.1% versus 23.8% (CDC, 2013; Go, et al., 2013). Until now, the largest lipid GWAS in Mexican Hispanics was performed by Below et al. in 2015 with 4,383 Mexican ancestry individuals. However, all genome-wide significant regions identified in the Mexican meta-analysis have been previously associated with lipid traits (Below, et al., 2016; Surakka, et al., 2015; Willer, et al., 2008; Willer, et al., 2013).

GWAS were designed with a focus on common variants, partly supported by the "common disease, common variant" hypothesis (Manolio, et al., 2009). However, despite the large number of genetic signals identified by GWAS, over 80% fall outside of protein coding regions, which complicates causal inference (Manolio, et al., 2009). Among genes with evidence of association, only a small proportion of variance is explained, providing limited information for disease risk prediction (Choquet & Meyre, 2011; Willer, et al., 2013). Sequencing of the exome, rather than the entire human genome, has been shown to be an efficient strategy to search for novel variants with larger effects (Ng et al., 2010). Previous exome sequencing studies have identified multiple rare variants associated with CVD in non-Hispanic populations (J. C. Cohen, Boerwinkle, Mosley, & Hobbs, 2006; Do et al., 2015).

To search for functional coding variants regulating circulating lipid levels in Mexicans, whole exome sequencing was performed in 1,205 Mexican Americans from the Insulin Resistance Atherosclerosis Family Study (IRASFS). Association and family-based linkage analyses were performed for 548,889 variants. We hypothesized that a family cohort would have increased power to detect rare variants due to their transmission in multigenerational pedigrees. With the complimentary approaches of linkage and association, exome sequencing has the potential to identify ethnic-specific variants regulating lipid levels.

Methods

Insulin Resistance Atherosclerosis Family Study (IRASFS)

The study design, recruitment, and phenotyping for the IRASFS has been previously described (Henkin, et al., 2003). In brief, the IRASFS was designed to investigate the genetic and environmental basis of insulin resistance and adiposity. Mexican Americans included in this cohort (N=1,417 individuals, 90 pedigrees) were recruited from clinical centers in San Antonio,

TX and San Luis Valley, CO. Since recruitment was based on reported family size and not on affection (diabetes) status, only about 12.7% of genotyped subjects had diabetes. After removing individuals with diabetes and incomplete phenotypes, 1,205 individuals from 90 pedigrees were included in this report. Phenotype acquisition and variable calculations have been previously described (Henkin, et al., 2003; Wing et al., 2011). In brief, TC, TG, and HDL were measured from fasting plasma with standards and LDL was calculated using Friedewald formula. The study protocol was approved by the Institutional Review Board of each participating clinical and analysis site and all participants provided their written informed consent.

Exome Sequencing

Exome sequencing was performed at Texas Biomedical Research Institute using the Illumina Nextera Exome Enrichment System in conjunction with the Illumina HiSeq 2500 sequencer. All sequence reads passed through the Illumina Data Analysis Pipeline, and those from samples passing QC criteria were mapped to the human genome reference sequence (hg19). A detailed description of the sequencing platform and analysis pipeline has been published (Tabb et al., 2016).

Statistical Analysis

To ensure normality, high density lipoprotein (HDL), total cholesterol (TC), and triglycerides (TG) were natural log-transformed; low density lipoprotein (LDL) was square root-transformed. While lipid medications were carefully evaluated, the majority of the participants in IRASFS were metabolically healthy with a lipid medication rate of 3.9%. Therefore, only a single regression model was performed without accounting for lipid medication status. Heritability of the lipid levels were estimated using Sequential Oligogenic Linkage Analysis Routines (SOLAR) (Almasy &

Blangero, 1998) adjusting for age, sex, body mass index (BMI), recruitment center. Tests of association between individual variants and quantitative traits were computed using the Wald test from the variance component model implemented in SOLAR. Variants with Mendelian inconsistencies were removed (N=4,024), resulting in a final number of 548,889 variants (O'Connell & Weeks, 1998). Each variant was coded to an additive model based on the minor allele (reference allele). Genetic model of association were calculated adjusting for age, sex, body mass index (BMI), recruitment center, and admixture estimates. Admixture estimates were calculated as described previously (Gao, et al., 2015) using maximum likelihood estimation of individual ancestries as implemented in ADMIXTURE (Alexander, et al., 2009). For family-based linkage analysis, variant-specific identity-by-descent (IBD) probabilities were computed using the Monte Carlo method implemented in SOLAR (Hellwege et al., 2014). Two-point linkage was performed using the variance components method implemented in SOLAR, with adjustment of age, gender, recruitment center, and BMI (Hellwege, et al., 2014). Variant annotation was performed using ANNOVAR (K. Wang, Li, & Hakonarson, 2010).

Replication and Meta-analysis

Six cohorts participating in the Genetics Underlying Diabetes in Hispanics (GUARDIAN) Consortium (Goodarzi, et al., 2014) provided in silico replication data: the Insulin Resistance Atherosclerosis Study (IRAS (Wagenknecht, et al., 1995)), BetaGene (Black, et al., 2008; Li, et al., 2009; Shu, et al., 2009; Watanabe, et al., 2007), the Troglitazone in Prevention of Diabetes Study (TRIPOD (Buchanan, et al., 2000; Buchanan, et al., 2002)), the Hypertension-Insulin Resistance Family Study (HTN-IR (Cheng, et al., 2001; Xiang, et al., 2001)), the Mexican-American Coronary Artery Disease Study (MACAD (Goodarzi, et al., 2004; Goodarzi, et al., 2003; Goodarzi, et al., 2005)) and the NIDDM-Atherosclerosis Study (NIDDM-Athero (Y.-P. Wang, et al., 2000)). All study protocols were approved by the local institutional review committees and all participants

gave their informed consent. GWAS genotyping was supported through the GUARDIAN Consortium (Goodarzi, et al., 2014) using the Illumina OmniExpress array (Illumina Inc.; San Diego, CA, USA) and imputation was performed centrally using IMPUTE2 (B. N. Howie, et al., 2009) and the 1000G phase l integrated reference panel (March 2012). Variants included in analysis had confidence scores >0.90 and information scores >0.50. A detailed description of quality control can be found in our previous publication (Gao, et al., 2015).

A total of 56 nominally associated SNPs ($P<5.00x10^{-05}$) with minor allele frequency (MAF) ≥1.0% as well as two rare SNPs that reached exome-wide significance ($P<9.11x10^{-08}$, MAF<1.0%) were selected for replication in the GUARDIAN Consortium. Meta-analysis was performed among IRASFS (n_{max}=1,205), IRAS (n_{max}=181), BetaGene (n_{max}=1,218), TRIPOD (n_{max}=125), HTN-IR (n_{max}=763), MACAD (n_{max}=749), and NIDDM-Athero (n_{max}=244) using the 1000G imputation dataset. Overall, 49 SNPs were directly tagged in replication cohorts and four SNPs were tagged by a proxy SNP (r^2>0.6). However, no available proxies were found for the remaining five SNPs, which were excluded, resulting in a total of 53 SNPs in the meta-analysis. The meta-analysis was computed using METAL (http://csg.sph.umich.edu/abecasis/metal). Considering the differential study designs, a weighted meta-analysis of the p-values and samples sizes accounting for direction of effect was performed.

Previously identified signals

Previously identified lipid loci (N=157) from the recent GLGC (Willer, et al., 2013) were extracted and all exome sequencing variants within ±100kb of the reported index SNPs were selected for fine-mapping. Conditional association and linkage analyses were performed with GLGC index SNP as an adjusting covariate.

ENCODE RNA sequencing data

Four human primary hepatocytes RNA sequencing results were plotted using the UCSC genome browser (Bashliev, 1987; "An integrated encyclopedia of DNA elements in the human genome," 2012). The four liver biopsy samples included were derived from four European individuals: GSM2072386 (20-week female), GSM2072387 (22-week male), GSM2072372 (32-year male), and GSM2072373 (6-year female).

Results

A total of 1,205 individuals were included in association and linkage analyses. Characteristics of the study individuals are shown in **Table 1**. Overall, individuals were predominantly female (59%) and were overweight with an average BMI of 29 kg/m^2. Since the recruitment was based on family size rather than diagnosis, e.g. CVD was not required for participation, the participants were metabolically normal, with average HDL (43.61 mg/dl), LDL (109.41 mg/dl), TC (178.06 mg/dl), and TG (124.80 mg/dl) within desirable or near-desirable ranges ("Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report," 2002). According to the National Cholesterol Education Program (NCEP) ("Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report," 2002), 183 individuals (15%) within the study had undesirable high TG levels (TG >150 mg/dl), 121 (10%) individuals had an LDL level greater than 160 mg/dl, 521 (43%) individuals had an HDL level less than 40 mg/dl, and 290 (24%) individuals had a TC level greater than 200 mg/dl.

Heritability analysis in IRASFS suggested a strong genetic component for lipid levels. HDL had the strongest heritability (H_{HDL}=0.57) with 17.2% of the variance explained by covariates (age, sex,

center, bmi) ($P=5.87x10^{-26}$), the heritability of TC was 0.53 with 10.3% of the variance explained by covariates ($P=3.12x10^{-23}$), the heritability of LDL was 0.50 with 7.3% of the variance explained by covariates ($P=1.22x10^{-21}$), and TG had the lowest heritability ($H_{TG}=0.42$) with 15.9% of the variance explained by covariates ($P=5.37x10^{-14}$).

From exome sequencing data, a total of 548,889 variants were successfully analyzed for association and linkage. Among them, 30.0% (164,591) were extremely rare variants with only one or two observations and 82.5% (452,807) were low frequency variants as defined by a MAF <5%; 6.8% (N=37,157) of the variants were insertions/deletions; 33.1% (N=182,020) and 15% (N=83,874) of the variants marked an non-synonymous and synonymous amino acid change in coding genes, respectively.

Association and linkage results are shown in **Figure 1**. The strongest evidence of association attaining genome-wide significance ($P<5.0x10^{-08}$) was observed at the apolipoprotein A-V gene (*APOA5*) with two highly correlated SNPs ($r^2=0.92$) rs651821 ($P=3.67x10^{-10}$, LOD=2.36, MAF=14%) and rs2072560 ($P=5.14x10^{-10}$, LOD=2.05, MAF=13%) with TG (**Figure 1d**; **Table 2**) on chromosome 11. On average, individuals had 23% more TG for each risk allele carried at rs651821 (TT: 117.59 mg/dl, TC: 142.62 mg/dl, CC: 174.41 mg/dl). SNP rs651821 is a 5'-UTR (untranslated region) variant located 3 bases upstream of the first coding exon of *APOA5* and rs2072560 is an intronic SNP located between exon 3 and 4.

Additional evidence of association which attained exome-wide significance ($P<9.11x10^{-08}$, based on a conservative Bonferroni correction for 548,889 variants) included two correlated SNPs ($r^2=1.0$) on chromosome 4: rs189547099 ($P_{TG}=6.31x10^{-08}$, $LOD_{TG}=3.12$, MAF=0.5%) and chr4:157997598 ($P_{TG}=6.31x10^{-08}$, $LOD_{TG}=3.12$, MAF=0.5%) (**Table 2**). Notably, these variants were both significantly associated and linked with TG. SNP rs189547099 is an intronic SNP located in

the folliculin interacting protein 2 gene (*FNIP2*) and the chromosome 4 open reading frame 45 (*C4orf25*). SNP chr4:157997598 is located 1,817 kb upstream of SNP rs189547099 in the first intron of the glycine receptor beta gene (*GLRB*). On average, risk allele (T) heterozygous carriers for chr4:157997598 had 2.9 times TG compared to non-carriers (CC: 124 mg/dl vs. TC: 358 mg/dl). No risk allele homozygotes were found.

Two-point linkage analysis was performed for 548,889 variants. Of these, two SNPs had a LOD score greater than 4 (**Table 2**), 46 SNPs reached a LOD greater than 3. Among these, 14 variants were significantly linked (LOD>3) with TC, 13 variants were significantly linked with HDL and TG respectively, and 8 variants were significantly linked with LDL. The strongest linkage signal was observed at SNP rs1141070 (LOD_{LDL}=4.30 LOD_{TC}=3.93, MAF=22%) with LDL and TC levels. rs1141070 is located in exon 5 of the DNA fragmentation factor gene (*DFFB*) on chromosome 1 and marks a synonymous amino acid substitution. In addition, SNP rs11648905 (LOD_{LDL}=4.28, LOD_{TC}=3.77, MAF=41%) was strongly linked with LDL and TC levels. This variant is an intronic SNP located between exon 7 and 8 of the transmembrane protein 8A gene (*TMEM8A*). No significant association signals were observed for the two linked signals: rs1141070, P_{LDL}=0.33, P_{TC}=0.74; rs11648905, P_{LDL}=0.78, P_{TC}=0.84 (**Table 2**).

Meta-analysis

Meta-analysis with six additional independent cohorts was computed for the 53 selected SNPs. Overall, six variants reached genome-wide significance after meta-analysis (**Table 3**). The strongest signal was observed at rs1160983 (P_{LDL}=4.44x10^{-17}, MAF=2.7%), which is a synonymous coding variant in the translocase of outer mitochondrial membrane 40 gene (*TOMM40*), with LDL. The two *APOA5* variants, which attained genome-wide significance in IRASFS, were also successfully replicated (meta-analysis p-values: rs2072560, P_{TG}=5.67x10^{-16}; rs651821,

P_{TG}=2.66x10^{-15}) with a consistent direction across all cohorts. In addition, strong meta-analysis signals were detected for *APOA1* (rs2070665, P_{TG}=7.03x10^{-09}) and *CETP* (rs1532625, P_{HDL}=7.72x10^{-14}, rs11076176, P_{HDL}=2.15x10^{-08}) with TG and HDL, respectively. SNP rs72685601 was selected as the proxy SNP (r^2=0.59) for the two variants that reached exome-wide significance (chr4:157997598, rs189547099). It was nominally associated with TG (P_{TG}=3.69x10^{-03}) with consistent direction of effects across six of the seven cohorts.

Previously identified lipid loci

Fine-mapping of the previously identified 157 loci +/-100kb resulted in a total of 14,232 exome sequencing variants which were analyzed for association and linkage in IRASFS. In addition, conditional analyses based on the locus-specific GLGC index SNP were performed. For association, 1,231 of the variants were identified with a P-value less than 0.05 with at least one of the reported lipid traits. Among the 1,231 significant variants, 646 remained to be significant (P<0.05) after conditional analysis based on the loci-specific index SNP. The strongest association signal was observed at rs651821 (P_{TG}=3.67x10^{-10}, LOD_{TG}=2.36) in *APOA5*. After conditioning on SNP rs964184, it remained nominally associated and linked with TG ($P_{TG|rs964184}$=1.76x10^{-04}, $LOD_{TG|rs964184}$=0.99). In addition, SNP rs1532625 also survived the stringent Bonferroni correction (P<3.51x10^{-06} for 14,232 variants): P_{HDL}=2.46x10^{-07}, LOD_{HDL}=1.42. This SNP is an intronic variant located in the cholesteryl ester transfer protein gene (*CETP*). However, conditional analysis with the index SNP rs3764261 totally abolished the signal ($P_{HDL|rs3764261}$=0.61, $LOD_{HDL|rs3764261}$=0.02). For linkage analysis, 14 and 127 SNPs reached a LOD score greater than two and one, respectively, for at least one reported lipid traits. Among them, one (rs1134760, $LOD_{HDL|rs16942887}$=2.46) and 28 (LOD>1) remained to be linked after conditional analysis, respectively.

Discussion

In this study, heritability estimates for four lipid phenotypes are reported in Mexican Americans from IRASFS. Subsequently, exome-wide association analysis was performed using exome sequencing data derived from 1,205 Mexican Americans from IRASFS. Multiple significant association signals were identified and top signals were evaluated in six additional independent cohorts (N=3,280). In addition, 157 previously identified lipid loci were fine-mapped using exome sequencing data in IRASFS. As a complementary approach to association, linkage analysis was performed to identify rare variant signals.

Strong association and suggestive linkage signals were observed with two SNPs in *APOA5*: rs651821 ($P_{TG}=3.67x10^{-10}$, $LOD_{TG}=2.36$, MAF=14%) and rs2072560 ($P_{TG}=5.14x10^{-10}$, $LOD_{TG}=2.05$, MAF=13%). The two SNPs are highly correlated ($r^2=0.98$) and reciprocal conditional analyses were able to abolish the signal limiting the ability to statistically implicate a functional variant. Replication and meta-analysis inclusive of six independent Mexican American cohorts confirmed the association ($P_{rs2072560\_TG}=5.26x10^{-16}$; $P_{rs651821\_TG}=2.66x10^{-15}$) (**Table 3**). In addition, both SNPs were also nominally associated with HDL ($P_{rs651821}=2.63x10^{-3}$; $P_{rs2072560}=9.42x10^{-3}$), while no signal was detected for LDL (P>0.90) or TC (P>0.18).

APOA5 encodes an apoliporotein that plays an important role in regulating plasma triglyceride levels, which is a strong risk factor for CVD (Miller et al., 2011). This gene is located within the apolipoprotein gene cluster on chromosome 11q23.3, which contains multiple lipid-related genes including *APOA1*, *APOA3*, *APOA4*, *APOA5*, and *PCSK7*. Multiple strong association signals have been identified in the region with HDL, TG, and TC (Surakka, et al., 2015; Teslovich et al., 2010; Willer, et al., 2013). A recent Hispanic GWAS of lipid phenotypes identified SNP rs964184, 359 bases downstream of zinc finger protein 259 (*ZNF259*) and 11kb downstream of *APOA5*, as

significantly associated with TG (Below, et al., 2016). In IRASFS, SNP rs964184 was also associated with TG ($P=4.79x10^{-07}$). Including SNP rs964184 as a covariant failed to completely abolish the genetic signal of rs651821 ($P_{before}=3.67x10^{-10}$, $P_{after}=1.76x10^{-04}$), suggesting that the two signals were likely independent ($r^2=0.37$). In 2015, Do *et al.* (Do, et al., 2015) described a large exome sequencing study in 9,793 European and African Americans and suggested strong association with myocardial infraction (MI) and rare coding-sequence mutations in *APOA5*. However, despite numerous genetic study results, the causal variant for the *APOA5* association signal in the general population remains unclear.

In our study, two common variants (rs651821, rs2072560) within *APOA5* were identified with strong association signals with TG in a Mexican American family cohort. While SNP rs651821 and rs2072560 have been previously identified to be strongly associated with TG in multiple ethnicities (Europeans, East Asians, and North Africans) (Costanza, Beer-Borst, James, Gaspoz, & Morabia, 2012; Ken-Dror, Goldbourt, & Dankner, 2010; Kettunen et al., 2012; Tan et al., 2012; L. Zhou et al., 2013), this is the first reported evidence for their association in a Mexican ancestry population. SNP rs651821 is a 5'-UTR variant that is three bases upstream of the coding exon. Worth mentioning, rs651821 is also located in the binding site of transcription factor POLR2A as suggested in HepG2 cells by ENCODE ("An integrated encyclopedia of DNA elements in the human genome," 2012). Previous expression quantitative trait loci (eQTL) studies revealed associations between rs651821 and transgelin (*TAGLN*) gene expression level (Jeong, Chung, Park, Cho, & Hong, 2014) yet no *APOA5* expression regulation effect was found. SNP rs2072560 is located intronically between exons 3 and 4 of *APOA5*. Interestingly, rs2072560 is also a missense variant of an alternative transcript of *APOA5* (NM_001166598) which contains two exons. This variant marks a glutamic acid to glycine amino acid change (E66G) in exon 2. To further explore the potential function of the alternative transcript, four ENCODE primary

hepatocyte RNA sequencing experiments were analyzed and plotted using the UCSC genome browser (Bashliev, 1987; "An integrated encyclopedia of DNA elements in the human genome," 2012). However, no RNA sequencing evidence was found to support existence or function of the alternative transcript. Taken together, strong association, linkage, and replication signals were identified for the two *APOA5* SNPs with TG in Mexican Americans. While not enough biological evidence was found to support their causality, the results refined the scope of the *APOA5* association signals and provided information for future efforts to locate the causal variant in the region.

In addition to the *APOA5* signal, two correlated SNPs (r^2=1.0) rs189547099 (P_{TG}=6.31x10^{-08}, LOD_{TG}=3.12, MAF=0.5%) and chr4:157997598 (P_{TG}=6.31x10^{-08}, LOD_{TG}=3.12, MAF=0.5%) were detected with exome-wide significance. These are two rare SNPs with 12 heterozygous and no homozygous carriers. SNP rs189547099 is an intronic variant for both the chromosome 4 open reading frame 45 gene (*C4orf45*) and folliculin interacting protein 2 gene (*FNIP2*). *C4orf45* is an uncharacterized gene with unknown biological function. GTEx (Carithers & Moore, 2015) has detected that *C4orf45* is strongly expressed in testis, yet there was almost no expression in other tissues. *FNIP2* is a tumor suppressor gene that has been shown to be involved in regulating the apoptosis signaling pathway in tumors and is responsible for cellular metabolism and nutrient sensing (Hasumi et al., 2015; Linehan, Srinivasan, & Schmidt, 2010). SNP chr4:157997598 is an intronic variant in the glycine receptor beta gene (*GLRB*). This gene encodes the beta subunit of the glycine receptor and has been shown to function as a neurotransmitter-gated ion channel. Mutations in this gene have been shown to cause startle disease (Al-Owain et al., 2012; James et al., 2013). Interestingly, SNP chr4:157997598 is located in a CpG island and modifies the binding consensus sequence for a transcription factor zinc

finger protein 263 (*ZNF263*). Although no biological mechanism was found between GLRB and lipid metabolism, it is possible that SNP chr4:157997598 regulates TG levels through *ZNF263*.

Synonymous SNP rs1160983 in exon 5 of the translocase of outer mitochondrial membrane 40 gene (*TOMM40*) exhibited the strongest association signal after replication and meta-analysis with six independent cohorts (P_{LDL}=4.44x10^{-17}, MAF=2.7%). Before meta-analysis, this variant was nominally associated in IRASFS (P_{LDL}=8.61x10^{-06}). *TOMM40* has been shown to be the forming subunit of the translocase of the mitochondrial outer membrane complex and is essential for the import of protein precursors into mitochondria (Humphries et al., 2005). Genetic studies have identified two adjacent genes (*APOE* and *TOMM40*) in this region to be highly associated with circulating lipid levels (Salakhov et al., 2014; Willer, et al., 2013). After reviewing previously identified *APOE* variants, SNP rs7412 has the highest LD with rs1160983 (r^2=0.49, D'=0.94). In IRASFS, rs7412 was nominally associated with LDL (P_{LDL}=6.61x10^{-06}). Interestingly, after adjusting for the *APOE* variant (rs7412), rs1160983 remained nominally significant (P=9.10x10^{-03}) with LDL. This suggests that the known *APOE* signals do not fully explain the rs1160983 signal in *TOMM40*. It is possible that *TOMM40* may directly contribute to the regulation of LDL levels or SNP rs1160983 may influence *APOE* expression.

Another interesting observation in this study is the lack of overlapping between most linkage and association signals, even with exome sequencing data. One explanation is that association and linkage capture different mechanisms of phenotypic contributions. Association analysis detects signals that statistically associate with phenotypic variability either directly or through linkage disequilibrium (LD) and thus targets more proximal effects. Linkage detects the co-segregation of an allele with the phenotype in families and therefore can detect long-range effects due to limited recombination events across successive generations. Therefore, each approach has its advantages and limitations. For example, association analysis has gained much

success in common variant analysis while often suffering from reduced power to detect rare variants, e.g. statistical power is largely affected by inadequate sample size and limited LD with proximal common variants. In contrast, linkage analysis performance is largely dependent on family structures as well as the number of segregation events, e.g. when family structures are incomplete or allele segregation information is incomplete (only two generations or the parental generation allele information is missing), linkage analysis performance is largely dampened. On the other hand, rare variants (in populations) that are missed by association can be relatively common in a given family with segregation across multiple generations. In this scenario, linkage has increased power over association. In conclusion, both association and linkage analysis are valuable approaches for analyzing sequencing data in genetic studies, and they provide independent information.

Despite multiple strong signals identified, study limitations exist. First, the modest sample size in IRASFS (N=1,205) limits the power for the association analysis of rare variants, especially given the fact that 82% of the variants analyzed had a MAF <5%. As a complementary approach, linkage analysis in 90 pedigrees was performed and identified signals that were likely missed by association. Unfortunately, linkage analysis was unavailable in the replication cohorts, and therefore meta-analysis of linkage signals was not performed. Dyslipidemia was not required for participation in IRASFS, thus the majority of individuals were metabolically normal, providing no enrichment of risk alleles. Third, exome sequencing was not available in replication cohorts, and therefore replication was limited to GWAS imputed or proxy variants only. This approach modestly limited the number of SNPs for replication. Also, while all cohorts were of Mexican ancestry, different ascertainment criteria were used. For example, BetaGene recruited participants at high risk of gestational diabetes while HTN-IR recruited participants at high risk of

hypertension. This differs from IRASFS which is a population-based study recruited for large family size.

In summary, exome-wide association and linkage analyses were performed using exome sequencing data in 1,205 Mexican Americans from IRASFS. Multiple signals were detected with circulating lipid levels and top signals were analyzed in six additional independent cohorts for replication. Our results suggested multiple lipid genetic signals in *APOA5*, *TOMM40*, and *GLRB*/*C4orf45*, fine-mapped known lipid genes with exome sequencing data in IRASFS, and explored a combined approach of association and linkage analyzing sequencing data. These results confirm that exome sequencing is a powerful tool to screen for functional genetic variants in the population.

Compliance with Ethical Standards

The authors declare that they have no conflict of interest. Participants included in this study were recruited from clinical centers in San Antonio, TX and San Luis Valley, CO. The Institutional Review Board of each clinical (UT Health Science Center San Antonio Review Board and Colorado Multiple Institutional Review Board, respectively) and analysis (Wake Forest School of Medicine) site approved the study protocol and all participants provided their written informed consent.

Acknowledgements and funding information:

This research was jointly supported by HG007112 from the national Human Genome Research Institute (NHGRI) and DK097524 from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). Computational resources were provided, in part, by the Wake Forest School of Medicine Center for Public Health Genomics. The authors would like to acknowledge

the members of the GUARDIAN Consortium with research supported by DK085175 from NIDDK and from the following grants: IRAS Classic (HL047887, HL047889, HL047890, and HL47902), IRAS Family Study (HL060944 and HL061019), BetaGene (DK061628), MACAD (HL088457), HTN-IR (HL069794), and NIDDM (HL055798). The authors thank the other investigators, the staff, and the participants of the studies for their valuable contributions. The provision of genotyping data was supported in part by UL1TR000124 (CTSI), DK063491 (DRC), DK081350, HG007112 and DK087914.

Table 1. Demographic information of the study individuals.

	Discovery	Replications					
	IRASFS	IRAS	TRIPOD	BetaGene	HTN-IR	MACAD	NIDDM-Athero
n	1205	181	125	1218	763	749	244
Female (%)	58.5	41.1	0	27.5	41.05	42.6	38.87
Age (years)[b]	42.74±14.49	54.06±8.21	34.92±6.37	34.67±7.87	39.3±15.09	34.66±9.08	38.12±14.94
Body Mass Index (BMI; kg/m^2)[b]	28.92±6.16	28.21±5.11	30.76±5.56	29.57±6.07	29.27±5.74	28.96±5.15	29.08±6.27
High-density Lipoprotein (mg/dl)	43.61±12.86	43.18±14.63	37.39±9.52	46.89±11.08	48.13±13.24	46.14±12.12	47.5±11.7
HDL<40 mg/dl[1]	521 (43.45%)						
Low-density Lipoprotein (mg/dl)	109.41±31.04	140.04±36.75	109.01±27.30	102.80±28.65	106.03±31.38	108.23±31.3	106.63±32.01
LDL>160 mg/dl[1]	121 (10.18%)						
Total Cholesterol (mg/dl)	178.06±37.45	210.90±43.74	173.15±32.59	172.52±33.25	179.01±35.41	180.53±36.43	181.47±39.23
TC>200 mg/dl[1]	290 (24.17%)						
Triglycerides (mg/dl)	124.80±84.00	157.24±99.21	133.79±78.91	114.12±90.27	124.23±78.96	137.25±102.82	145.19±166
TG>200 mg/dl[1]	183 (15.25%)						
Lipid lowering medications	47 (3.90%)	NA[2]	NA[2]	NA[2]	NA[2]	NA[2]	NA[2]

[1]Desired cholesterol levels are defined based on National Cholesterol Education Program (NCEP) final report (2002); [2]NA represents not available.

Table 2. Top association ($P<9.11x10^{-08}$) and linkage signals (LOD>4).

SNP	Chr:pos (hg19)	Gene	Annotation	Alleles[a]	RAF[b]	P_{HDL}	LOD_{HDL}	P_{LDL}	LOG_{LDL}	P_{TC}	LOD_{TC}	P_{TG}	LOD_{TG}
rs1141070	1:3786189	*DFFB*	coding-synon	A/G	0.22	0.98	0.01	0.33	**4.30**	0.74	**3.93**	0.87	0.27
chr4:157997598	4:157997598	*GLRB*	intron	T/C	0.0050	2.88E-04	1.87	0.71	0	0.24	0	**6.31E-08**	**3.13**
rs189547099	4:159814881	*C4orf45*	intron	C/G	0.0050	2.88E-04	1.87	0.71	0	0.24	0	**6.31E-08**	**3.13**
rs2072560	11:116661826	*APOA5*	intron	T/C	0.13	2.03E-03	0.32	0.87	0	0.014	0.010	**5.14E-10**	2.06
rs651821	11:116662579	*APOA5*	intron	C/T	0.14	8.93E-04	0.30	0.99	0	0.018	0	**3.67E-10**	2.36
rs11648905	16:425298	*TMEM8A*	intron	T/G	0.41	0.74	0	0.78	**4.28**	0.84	**3.77**	0.18	0.82

[a]Reference/Other allele with minor allele on the left; [b]Reference allele frequency based on the entire population.

Table 3. Top association signals from meta-analysis.

SNP	Chr	Position (hg19)	RAF[a]	Trait	Gene	Annotation	Exome sequencing		Meta-analysis	
							N	P	N	P
rs2072560	11	116661826	0.13	TG	*APOA5*	intron	1205	5.14E-10	4241	5.67E-16
rs651821	11	116662579	0.14	TG	*APOA5*	intron	1205	3.67E-10	4241	2.66E-15
rs2070665	11	116707684	0.17	TG	*APOA1*	intron	1205	4.10E-05	4241	7.03E-09
rs1532625	16	57005301	0.38	HDL	*CETP*	intron	1205	2.46E-07	4235	7.72E-14
rs11076176	16	57007446	0.28	HDL	*CETP*	intron	1205	3.87E-06	4235	2.15E-08
rs1160983	19	45397229	0.027	LDL	*TOMM40*	coding-synon	1205	8.61E-06	4177	4.44E-17

[a]Reference (minor)/Other allele.

Figure 1. Opposed Manhattan plots for association and linkage analysis a. HDL, b. LDL, c. TC, d. TG.

Figure 1a.

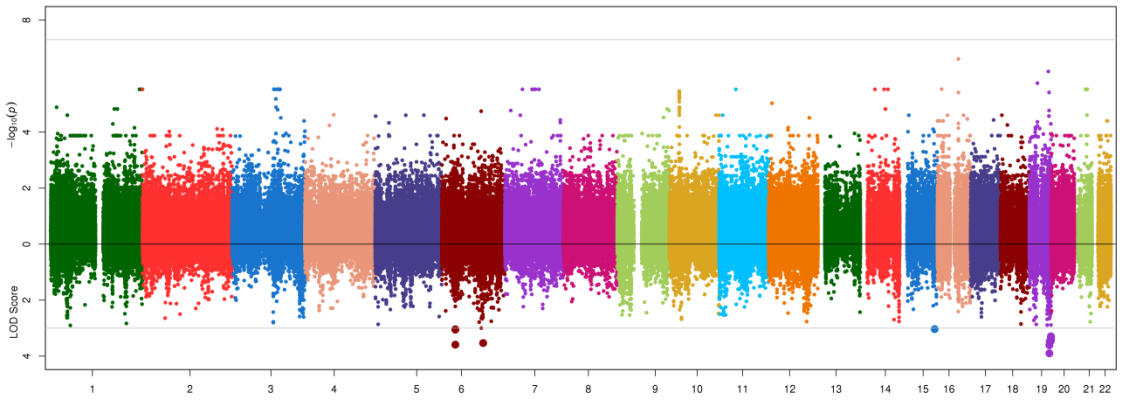

8
4
2
0
2
4
$-\log_{10}(p)$
LOD Score
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22

Figure 1b.

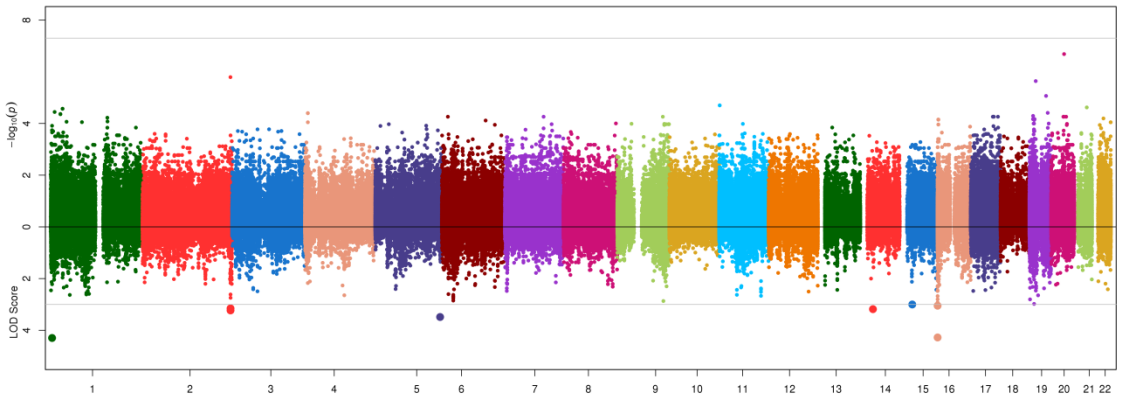

8
4
2
0
2
4
$-\log_{10}(p)$
LOD Score
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22

Figure 1c.

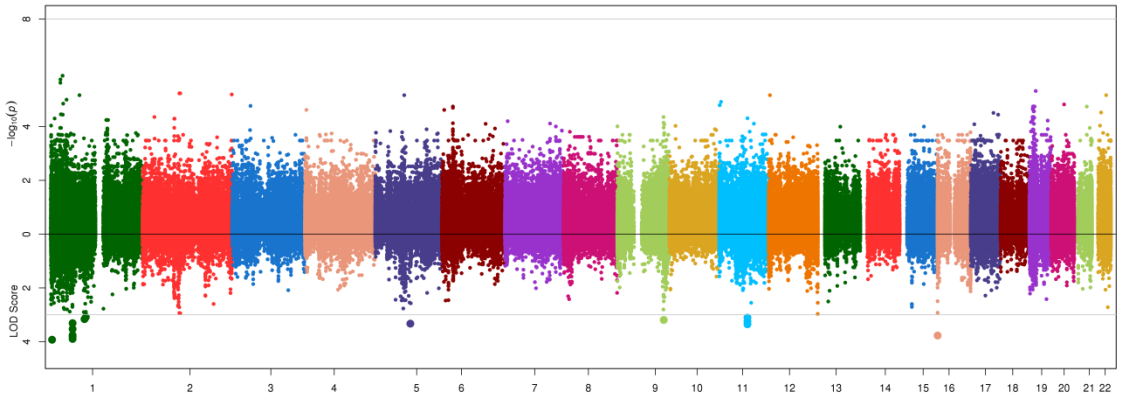
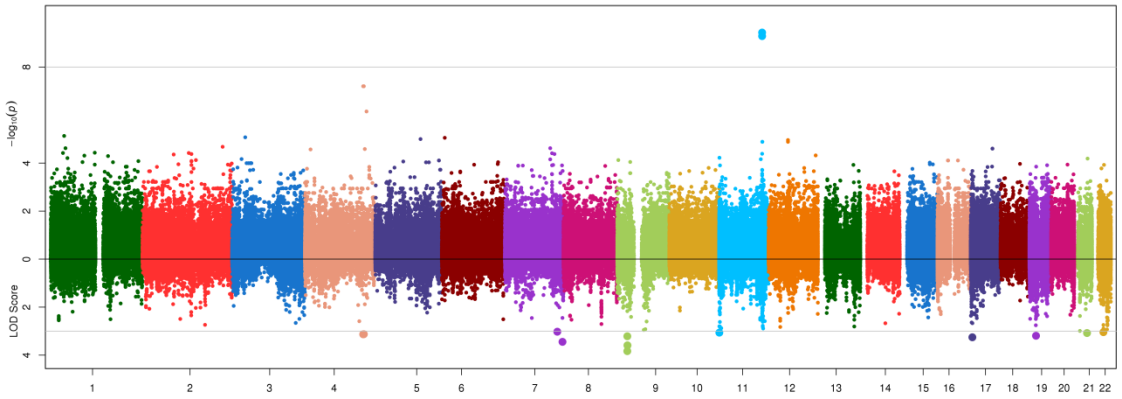

8
4
2
0
2
4
$-\log_{10}(p)$
LOD Score
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22

Figure 1d.



Chapter V

A Genome-wide Linkage and Association Analysis of Imputed Insertions and Deletions with Cardiometabolic Phenotypes in Mexican Americans: The Insulin Resistance Atherosclerosis Family Study

Chuan Gao, Fang-Chi Hsu, Latchezar M. Dimitrov, Hayrettin Okut, Yii-Der I. Chen, Kent D. Taylor, Jerome I. Rotter, Carl D. Langefeld, Donald W. Bowden, Nicholette D. Palmer

The following manuscript was submitted to Genetic Epidemiology. Stylistic variations are due to the requirements of the journal. Chuan Gao performed the Analyses and prepared the manuscript. Dr. Nicholette Allred acted in an advisory and editorial capacity.

Abstract

Insertions and deletions (INDELs) represent a significant fraction of inter-individual variation in the human genome yet their contribution to phenotypes is poorly understood. To confirm the quality of imputed INDELs and investigate their roles in mediating cardiometabolic phenotypes, genome-wide association and linkage analyses were performed for 15 phenotypes with 1,273,952 imputed INDELs in 1024 Mexican-origin Americans. Imputation quality was validated using whole exome sequencing with an average kappa of 0.93 in common INDELs (MAF≥5%). Association analysis revealed one genome-wide significant association signal for the Cholesterylester Transfer Protein gene (*CETP*) with high density lipoprotein levels (rs200373219, $P=3.06x10^{-12}$); linkage analysis identified two peaks with LOD>5 (rs60560566, LOD=5.36 with insulin sensitivity (S_I) and rs5825825, LOD=5.11 with adiponectin levels). Suggestive overlapping signals between linkage and association were observed: rs59849892 in the WSC Domain Containing 2 gene (*WSCD2*) was associated and nominally linked with S_I ($P=1.17x10^{-7}$, LOD=1.99). This gene has been implicated in glucose metabolism in human islet cell expression studies. In addition, rs201606363 was linked and nominally associated with low density lipoprotein ($P=4.73x10^{-4}$, LOD=3.67), apolipoprotein B ($P=1.39x10^{-3}$, LOD=4.64), and total cholesterol ($P=1.35x10^{-2}$, LOD=3.80) levels. rs201606363 is an intronic variant of the *UBE2F-SCLY* fusion gene which may regulate cholesterol through selenium metabolism. In conclusion, these results confirmed the feasibility of imputing INDELs from array-based SNP genotypes. Analysis of these variants using association and linkage replicated previously identified SNP signals and identified multiple novel INDEL signals. These results support the inclusion of INDELs into genetic studies to more fully interrogate the spectrum of genetic variation.

Key words: cardiometabolic disease, genome-wide association analysis, imputation, insertion/deletions, linkage analysis

Introduction:

Insertions and deletions (INDELs) of DNA sequence contribute a significant fraction of genetic variation in the human genome (Weischenfeldt, et al., 2013). Genetic studies have confirmed association signals between INDELs and multiple disease and non-disease related phenotypic variations including dietary starch consumption, autism, schizophrenia, Crohn's disease, rheumatoid arthritis, type 1 diabetes, and obesity (Cantsilieris & White, 2013; Craddock, et al., 2010; Jacquemont, et al., 2011; Malhotra & Sebat, 2012; Perry, et al., 2007; Pinto, et al., 2014). However, compared to single nucleotide polymorphism (SNP) association analyses, very few studies have included INDELs.

Traditional discovery and genotyping approaches for INDELs largely rely on polymerase chain reaction (PCR) and several other modified PCR-based techniques (Dhawan & Padh, 2009). These techniques are either expensive or inconvenient and therefore not ideal for large-scale studies (Almal & Padh, 2012). More recently, short-read DNA sequencing data from the 1000 Genomes Project (1000G) has enabled a better constructed set of INDELs across different ethnicities with enhanced size and breakpoint resolution (Mills, et al., 2011; Sudmant, et al., 2010). The release of the 1000G reference panel has enabled researchers, for the first time, to accurately impute large numbers of INDELs from array-based genotypes (Abecasis, et al., 2012). To further investigate the feasibility of INDEL imputation, Lu *et al.* characterized the linkage disequilibrium between INDELs and nearby tagging SNPs using Next Generation Sequencing (NGS) data. These results suggest a high concordance rate and correlation between INDELs and nearby SNPs of similar minor allele frequencies (MAF), which suggests ample opportunities for genome-wide association studies (GWAS) with imputed INDELs to capture additional variation across the genome.

To evaluate the quality of INDEL imputation as well as comprehensively investigate their role in cardiometabolic phenotypes, we performed genome-wide association and linkage analyses of 1,273,952 imputed INDELs in 1024 Mexican-origin Americans from the Insulin Resistance Atherosclerosis Family Study (IRASFS)(Henkin, et al., 2003). We hypothesize that INDELs can be imputed based on array-based data from GWAS and with appropriate association and linkage approaches, analysis of INDELs may identify additional novel signals and provide valuable biological insights.

Materials and Methods:

Insulin Resistance Atherosclerosis Family Study (IRASFS)

The study design, recruitment, and phenotyping for the IRASFS has been previously described (Henkin, et al., 2003). In brief, the IRASFS was designed to investigate the genetic and environmental basis of insulin resistance and adiposity. After removing individuals with diabetes, 1024 Mexican-origin Americans (recruited from clinical centers in San Antonio, TX and San Luis Valley, CO) from 88 pedigrees were included in this report. The study protocol was approved by the Institutional Review Board of each participating clinical and analysis site and all participants provided their written informed consent.

Genotyping and Imputation

GWAS genotyping was supported through the Genetics Underlying Diabetes in Hispanics (GUARDIAN) Consortium (Goodarzi, et al., 2014) using the Illumina OmniExpress and Omni1S arrays (Illumina Inc.; San Diego, CA, USA) in 1024 individuals as well as 13 duplicate controls. A detailed description of genotyping platforms and quality controls can be found in our previous publication (Palmer et al., 2015). Imputation was performed using IMPUTE2 (B. N. Howie, et al., 2009) and the 1000G phase l integrated reference panel. Variants were checked for confidence

score (>0.90), information score (>0.50), and Mendelian errors resulting in a total of 1,273,952 high quality INDELs. INDEL annotation was performed using ANNOVAR (K. Wang, et al., 2010). To validate the imputation quality, 9,370 INDELs that were captured by our whole exome sequencing were analyzed for sequence-imputation concordance using the kappa statistic.

Exome sequencing

Exome sequencing was performed at Texas Biomedical Research Institute using the Illumina Nextera Exome Enrichment System in conjunction with the Illumina HiSeq 2500 sequencer. All sequence reads were passed through the Illumina Data Analysis Pipeline. All sequence reads from samples passing QC criteria were mapped to the human genome reference sequence (hg19). A detailed description of the sequencing platforms and analysis pipeline can be found in supplemental material.

Statistical Analysis

Phenotype acquisition and variable calculations have been previously described (Henkin, et al., 2003; Wing, et al., 2011). Briefly, insulin sensitivity (S_I) and glucose effectiveness (S_G) were obtained using the frequently sampled intravenous glucose tolerance test with minimal model (MINMOD) analysis (Bergman, Finegood, & Ader, 1985; Pacini & Bergman, 1986). Acute insulin response (AIR) was measured at 8 minutes following glucose infusion as the mean insulin increment in plasma insulin concentration above the basal concentration. Disposition Index (DI) was calculated from the product of S_I and AIR. Metabolic clearance rate of insulin (MCRI) was calculated as the ratio of the insulin dose over the incremental area under the curve of insulin. Fasting plasma glucose (GFAST) and insulin (FINS), biomarkers, and lipid traits were measured from a fasting blood draw using standardized approaches. To fulfill the distributional assumptions of conditional normality, phenotypic transformations were performed. Specifically,

AIR and DI were sign-square root transformed; adiponectin (ADP), FINS, homeostatic model assessment of beta-cell function and insulin resistance ($HOMA_B$, and $HOMA_{IR}$, respectively), MCRI, S_I, high density lipoprotein (HDL), total cholesterol (CHOL), and triglycerides (TRIG) were log transformed; low density lipoprotein (LDL) and apolipoprotein B (APOB) were square root transformed; GFAST and S_G were normally distributed and therefore required no transformation. For ADP, as a missense variant G45R in the *ADIPOQ* gene has been previously identified to have a large impact on adiponectin levels in IRASFS, a separate analysis for ADP was performed with the adjustment for the G45R variant (rs200573126; ADP_G45R) (Bowden et al., 2010).

Tests of associations between individual variants and quantitative traits were computed using the Wald test from the variance component model implemented in Sequential Oligogenic Linkage Analysis Routines (SOLAR) (Almasy & Blangero, 1998). Each variant was coded to an additive model based on the minor allele as the reference allele. Association was calculated adjusting for age, gender, BMI, recruitment center (San Antonio, TX or San Luis Valley, CO), and admixture estimates. Admixture estimates were calculated as described previously using maximum likelihood estimation of individual ancestries as implemented in ADMIXTURE (Alexander, et al., 2009; Gao, et al., 2015). For family-based linkage analysis, variant specific identity-by-descent (IBD) probabilities for best-guess genotype (likelihood>0.9) were computed using the Monte Carlo method implemented in SOLAR as described previously (Hellwege, et al., 2014). Genotypes with likelihood estimates less than 0.9 were zeroed to avoid Mendelian inconsistencies. This resulted in a total of 1,273,952 INDELs for analysis (N=560,367 with MAF≥0.05) (O'Connell & Weeks, 1998). Two-point linkage was performed using the variance components method implemented in SOLAR, with adjustment of age, gender, recruitment center, and BMI. Estimation of the phenotypic variance explain by SNPs and INDELs was performed using GCTA (Yang et al., 2010; Yang, Lee, Goddard, & Visscher, 2011). Statistical

analysis of kappa value variations across categories was performed with non-parametric GLM and the enrichment of MAF of linked INDELs was performed with an exact chi-square test using SAS version 9.4 (SAS Institute, Inc, Cary, North Carolina).

Results:

Characteristics of the study individuals are shown in **Table 1**. Overall, individuals were predominantly female (41% male) and were overweight with an average BMI of 28.3 kg/m^2. A total of 1024 individuals were included for the analysis of 15 cardiometabolic phenotypes. Overall, 1,273,952 INDELs were successfully analyzed for association and two-point linkage. Among them, 12.4% (N=157,954) were rare variants with only single or double minor allele observations and 55% (N=713,582) were low frequency variants as defined by a MAF<0.05 (**Figure 1a**). The INDELs were then stratified by the size of the insertion/deletion, which was defined as the difference of the length of the two alleles. As shown in **Figure 1b**, 53% (681,840) of the variants were single base insertions/deletions and those greater or equal to 8bp constituted less than 3% (53,245). Functional annotation suggested the majority of the INDELs were intergenic and intronic, i.e. 54% and 37%, respectively. There were only 4,582 exonic variants (less than 1%, including alternative splicing and ncRNA exonic variants) (**Figure 1c**). The comparison of exonic and non-exonic variants suggested no difference in MAF and a trend of enrichment in multiple bases insertion/deletions (P=0.23) in exonic variants.

Imputation quality

Imputation quality was evaluated through comparison of genotypes from 9,370 imputed and sequenced INDELs using Cohen's kappa statistic, which is a conservative measure of agreement over concordance rate. The results are summarized in **Figure 2a**. Overall, imputed genotypes were highly concordant with sequenced genotypes with an average kappa of 0.87. Most of the

poorly imputed variants were of low frequency and the average kappa for common variants (MAF≥0.05) was 0.93 (**Figure 2b**). Further analysis revealed kappa values varied significantly based on different INDEL sizes (P=0.0006, lower kappa values for large INDEL variants) and functional annotations (P<0.001).

Top signals from association and linkage analysis

The strongest evidence of association was observed with insertion rs36229491 with HDL levels ($P=3.06x10^{-12}$, LOD=1.52, **Table 2**). This is a common variant with a MAF of 27% located 1,591 bp upstream of the cholesterylester transfer protein gene (*CETP*). On average, minor allele carriers have a 25% increase in plasma HDL per allele. In addition, a *CETP* intronic insertion rs35874588 (MAF=0.45) was also detected with a strong LOD score of 4.33 and a P value of $5.48x10^{-4}$. The two variants are poorly correlated (r^2=0.05) and analysis conditioned on rs36229491 failed to abolish the rs35874588 signal, suggesting at least two independent signals exist. These results are consistent with previous findings and rs36229491 and rs35874588 are in strong LD with previously identified SNPs rs3764261 and rs5882 (r^2=0.98 and 0.95, respectively) (Hellwege, et al., 2014; Willer, et al., 2008).

For linkage analysis, two strong linkage peaks were detected with LOD scores greater than 5: rs60560566 (7q11.22, MAF=0.30) was strongly linked with S_I with a LOD of 5.36 and rs5825825 (18q22.1, MAF=0.33) was strongly linked with ADP with a LOD of 5.11 after adjustment for the *ADIPOQ* G45R variant (rs200573126). rs60560566 marks a single base pair deletion located in a gene dessert region, i.e. no genes were identified ±400kb. rs5825825 is a single base pair insertion located intergenically between dermatan sulfate epimerase-like gene (*DSEL*) and thioredoxin related transmembrane protein 3 gene (*TMX3*). Another variant rs201639667 (MAF=0.39) was strongly linked to APOB and CHOL with LOD scores of 4.69 and 3.71,

respectively. This is a single base pair insertion located between the macrophage scavenger receptor 1 gene (*MSR1*) and fibroblast growth factor 20 gene (*FGF20*). *MSR1* is closely involved in macrophage-associated physiological and pathological processes and overexpression of the gene results in increased clearance of modified lipoproteins by Kupffer cells, resulting reduced VLDL levels (Herijgers et al., 2000). rs59849892, an intronic variant in the WSC domain containing 2 gene (*WSCD2*), was both associated and linked with S_I ($P=1.17x10^{-7}$, LOD=1.99). *WSCD2* has been shown to be involved in glucose metabolism consistent with expression patterns observed in human islets (Taneera et al., 2015). Variant rs201606363 was strongly linked and associated with LDL ($P=4.73x10^{-4}$, LOD=3.67), APOB ($P=1.39x10^{-3}$, LOD=4.64), and TC ($P=1.35x10^{-2}$, LOD=3.80). This common variant (MAF=0.43) is located between the ubiquitin-conjugating enzyme E2F gene (*UBE2F*) and the selenocysteine lyase gene (*SCLY*). *SCLY* has been shown to be involved in selenium metabolism and selenoproteins synthesis (Mihara, Kurihara, Watanabe, Yoshimura, & Esaki, 2000). High serum selenium concentrations have been shown to be associated with increased serum concentrations of total and LDL cholesterol in epidemiological studies (Laclaustra, Stranges, Navas-Acien, Ordovas, & Guallar, 2010). Therefore, *SCLY* may play an important role in this association.

Characteristics of the nominally associated and linked INDELs

To investigate whether an enrichment of a certain characteristic exists for nominally linked and associated INDELs, the dataset was parsed based on nominal association ($P<0.05$), association ($P<0.005$), nominal linkage ($LOD>0.5$), and linkage ($LOD>1$) as exemplified by TRIG and HDL (**Figure 3**). As shown in the figures, no enrichment was observed for INDEL size with association and linkage signals (P=0.91). For functional annotation, a slightly higher proportion of exonic variants was observed for associated and linked variants. However, no statistically significant conclusions could be drawn due to the limited number of exonic variants in this study (n=4,582)

(P=0.79). In contrast, a significant enrichment of common variants (MAF≥0.05) for linked INDELs was observed (P<0.0001) but not for association (P=0.82). This observation is likely attributed to the greater power of common variants in linkage analysis.

Discussion:

In this study, we successfully imputed INDELs using array-based SNP genotypes in a Mexican-origin American cohort. Genome-wide association and linkage analysis were performed with cardiometabolic phenotypes using 1.27 million high quality imputed INDELs.

One key question we wanted to ask during the beginning of the study was whether INDELs could be imputed using array-based genotypes and be analyzed in complement to SNPs. Thus, a comparison was performed with genotypes derived from imputation and exome sequencing data using the kappa statistic. Despite some variations in rare variants (MAF<5%), common variants reached an average kappa value of 0.93, suggesting INDELs can be reliably imputed from array-based SNP genotypes. In addition, genome-wide association and linkage analyses were successfully performed for imputed INDELs using 15 cardiometabolic phenotypes. After excluding rare variants (MAF<0.01), all models were well behaved with inflation factors ranging from 0.96-1.08. Furthermore, examples of novel INDEL signals were identified and previously identified signals were successfully replicated from both association and linkage analysis as exemplified by *CETP* and *ADIPOQ*, respectively. In addition, the phenotypic variance explained by genome-wide common SNPs (N=5,650,462, MAF≥5%) increased 0.04% and 0.03% after including common INDELs (N=596,746, MAF≥5%) for HDL and TG, respectively. To further explore the coverage of imputed INDELs, a comparison was performed between imputed INDELs and exome sequencing captured INDELs. Overall, 38,122 INDELs were captured by exome sequencing and among them 9,370 were also successfully imputed based on GWAS genotypes.

The variants missed by imputation were predominantly (92%) rare variants with MAF<0.05, which indicates good imputation coverage of common INDEL variants derived from array-based SNP genotypes. The imputation of rare INDELs suffers similarly to that of rare SNPs (B. Howie, Fuchsberger, Stephens, Marchini, & Abecasis, 2012).

Consistent with our previous findings, linkage and association tend to provide independent genetic information (Hellwege, et al., 2014), i.e. association and linkage signals have limited overlap. This observation is not beyond expectation as association screens for variants having high LD with the causal variant (short range) while linkage detects variants co-segregating with the causal variants (long range). However, when the causal variant is directly tagged, both linkage and association are robustly able to detect the variant (An et al., 2013; Bowden, et al., 2010). Interestingly, when the causal variant is not well tagged, linkage tends to be more sensitive than association. For example, SNP rs200573126 (G45R) was previously reported as the causal variant for low adiponectin levels in IRASFS (An, et al., 2013; Bowden, et al., 2010). This variant was not well tagged by nearby INDELs and thus, no association signals were observed in the region. In contrast, linkage analysis detected a LOD peak of 7.67 (rs112871276; spanning Chr3:181332042-186759808), suggesting a better sensitivity based on limited LD with the causal variant. In conclusion, linkage analysis is a valuable approach for genetic studies which provides information independent from association analysis.

Despite the encouraging results study limitations exist. First, the modest sample size in IRASFS (n=1024) limits the power for the association analysis of rare variants, especially given the fact that 55% of the INDELs in the study had a MAF<0.05. In addition, although we observed good correlation and concordance rates between imputed and sequenced variants, it is still possible that imputation errors impede signal discoveries, especially for rare variants which were observed to have poorer imputation quality. Third, since SOLAR identity-by-descent (IBD)

calculations do not accept dosage genotypes, analysis requires a conversion of imputed dosage genotypes into best-guess. Also, to remove all Mendelian inconsistencies, all genotypes with potential errors were removed, which further dampened statistical power. Despite the rapidly developing technologies in the field, the genotyping and calling methods for short repeats in regions of complex genomic structure are still not robust (Treangen & Salzberg, 2012). Therefore, the INDELs included in this study represent an incomplete coverage of the genome. Whole genome sequencing or using a reference panel with a better genome coverage for imputation is able to address this issue.

In summary, a total of 1,273,952 INDELs were successfully imputed based on array-based SNP genotypes and a genome-wide association and linkage analysis were performed. Imputation quality was partially confirmed using exome sequencing data. We identified a genome-wide significant signal for association (*CETP*) with HDL and two signals (four variants) reached a linkage LOD score of 5 for S_I and ADP. Overall, INDEL analysis successfully confirmed previously identified signals from exome chip and GWAS (Hellwege, et al., 2014) and evidence of suggestive novel loci was observed. Our results support that INDELs can be confidently imputed based on array-based SNP genotypes and association and linkage are complimentary approaches that can be used for the analysis of imputed INDELs.

Acknowledgements and funding information:

This research was supported by DK097524 from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) and HG007112 from the national Human Genome Research Institute (NHGRI). Computational resources were provided, in part, by the Wake Forest School of Medicine Center for Public Health Genomics. The authors thank the other investigators, the

staff, and the participants of the studies for their valuable contributions. The provision of genotyping data was supported in part by UL1TR000124 (CTSI) and DK063491 (DRC).

Conflict of interest:

None

Ethical statement:

Participants included in this study were recruited from clinical centers in San Antonio, TX and San Luis Valley, CO. The Institutional Review Board of each clinical (UT Health Science Center San Antonio Review Board and Colorado Multiple Institutional Review Board, respectively) and analysis (Wake Forest School of Medicine) site approved the study protocol and all participants provided their written informed consent.

Table 1. Demographic characteristics of the study population.

	N	Mean ± SD	Median
Age (years)	1024	40.63±13.68	39.3
Male (%)	1024	0.41%	-
BMI (kg/m^2)	1024	28.28±5.77	27.52
S_I	1016	2.14±1.86	1.7
S_G	1016	0.021±0.0089	0.02
AIR ($pmol^{-1}$)	1016	760.33±649.12	586.95
DI (AIR x S_I)	1016	1315.91±1235.30	1004.79
GFAST (mg/dl)	1014	93.41±9.48	92
FINS (µU/ml)	1015	14.90±11.03	12
$HOMA_{IR}$	974	1.67±1.04	1.4
$HOMA_B$	974	120.76±45.62	113.85
MCRI	948	5.47±2.46	5.15
APOB	947	88.23±21.95	86
TRIG (mg/dl)	1012	118.30±82.02	97
LDL (mg/dl)	1004	109.04±30.11	106
HDL (mg/dl)	1012	43.58±12.27	42
CHOL (mg/dl)	1013	176.12±35.64	174
ADP (µg/ml)	940	13.36±6.36	12.49

Table 2. A selection of top signals from association and linkage analysis.

SNP	Chr:pos(hg19)	Alleles[a]	RAF[b]	function	Gene	P-value	Beta±SE	LOD	Trait
rs3832063	2:238944441	A/ACT	0.39	ncRNA_intronic	*UBE2F-SCLY*	3.50E-05	-0.27±0.066	2.70	LDL
rs201606363	2:238962655	A/ATG	0.43	ncRNA_intronic	*UBE2F-SCLY*	4.73E-04	-0.23±0.065	3.67	LDL
rs201606363	2:238962655	A/ATG	0.43	ncRNA_intronic	*UBE2F-SCLY*	1.39E-03	-0.16±0.051	4.60	APOB
rs201606363	2:238962655	A/ATG	0.43	ncRNA_intronic	*UBE2F-SCLY*	1.35E-02	-0.022±0.0090	3.80	CHOL
rs200229736	6:77298727	C/CTTA	0.22	intergenic	*IMPG1-HTR1B*	1.60E-06	-0.11±0.024	2.80	SI
rs201447751	7:67151523	ACC/A	0.29	intergenic	*LINC01372-LOC102723427*	8.99E-02	0.037±0.022	5.21	SI
rs149336631	7:67152121	C/CTG	0.30	intergenic	*LINC01372-LOC102723427*	1.16E-01	0.034±0.022	5.30	SI
rs60560566	7:67153704	T/TG	0.30	intergenic	*LINC01372-LOC102723427*	1.19E-01	0.034±0.022	5.36	SI
rs200122735	8:16588076	TA/T	0.39	intergenic	*MSR1-FGF20*	4.01E-01	0.047±0.056	4.50	APOB
rs201639667	8:16588078	AC/A	0.39	intergenic	*MSR1-FGF20*	4.34E-01	0.044±0.056	4.69	APOB
rs201639667	8:16588078	AC/A	0.39	intergenic	*MSR1-FGF20*	8.04E-01	-0.0025±0.010	3.70	CHOL
rs138536353	9:9513612	C/CAT	0.13	intronic	*PTPRD*	8.59E-06	-0.36±0.080	3.00	APOB
rs5896378	9:10194679	TTACA/T	0.16	intronic	*PTPRD*	6.38E-01	-0.033±0.070	3.90	APOB
rs201315549	12:108590892	ACCC/A	0.02	intronic	*WSCD2*	1.75E-06	0.39±0.081	1.79	SI
rs59849892	12:108593020	AT/A	0.02	intronic	*WSCD2*	1.20E-07	0.41±0.077	2.00	SI
rs36229491	16:56994244	TA/T	0.27	intergenic	*HERPUD1-CETP*	3.06E-12	0.092±0.013	1.52	HDL
rs66512242	16:56996645	G/GCC	0.46	intronic	*CETP*	3.00E-05	0.050±0.012	2.00	HDL
rs12720908	16:57001254	TCACA/T	0.18	intronic	*CETP*	2.70E-05	-0.065±0.016	2.01	HDL
rs35874588	16:57009651	TC/T	0.45	intronic	*CETP*	5.50E-04	0.042±0.012	4.30	HDL
rs5825825	18:65832839	A/AT	0.33	intergenic	*LOC643542-TMX3*	6.13E-01	0.013±0.025	5.11	ADP_G45R

[a]Reference/Other allele with minor allele on the left; [b]Reference allele frequency based on the entire population.

Figure 1. Distribution of MAF (a), INDEL size (b), and function annotations (c) for the entire INDELs dataset (N=1,273,952);

Figure 1a.

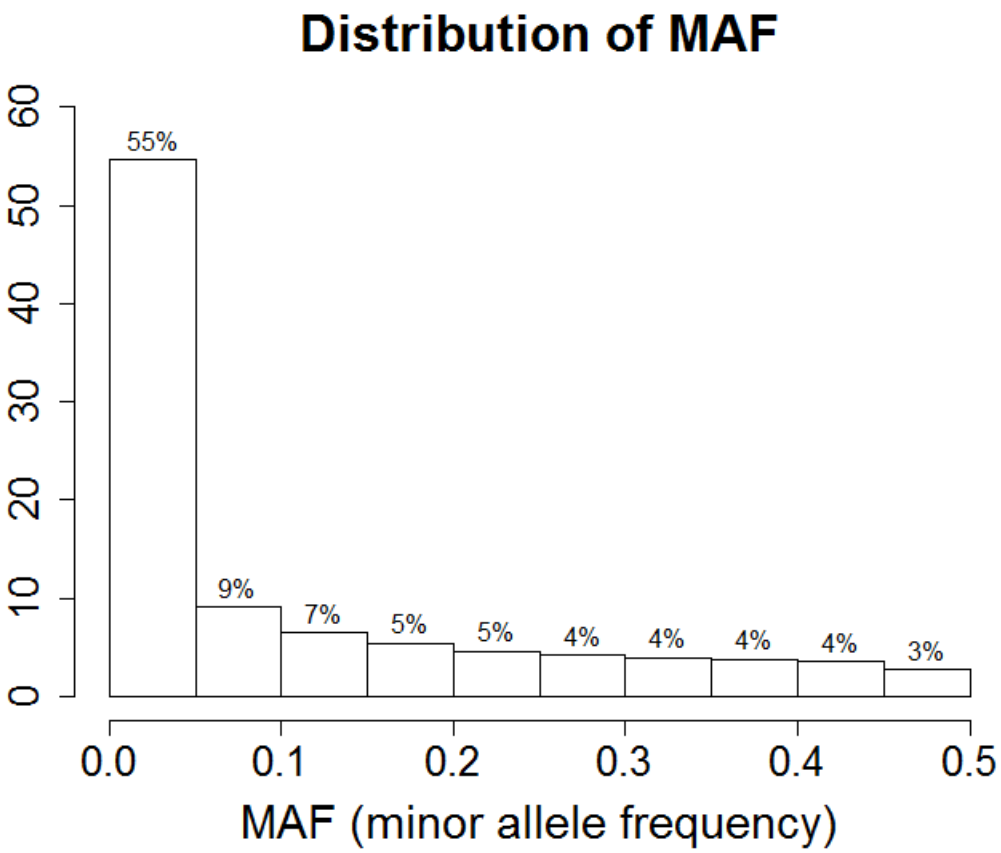

Distribution of MAF
0
10
20
30
40
50
60
55%
9%
7%
5%
5%
4%
4%
4%
4%
3%
0.0
0.1
0.2
0.3
0.4
0.5
MAF (minor allele frequency)

Figure 1b.

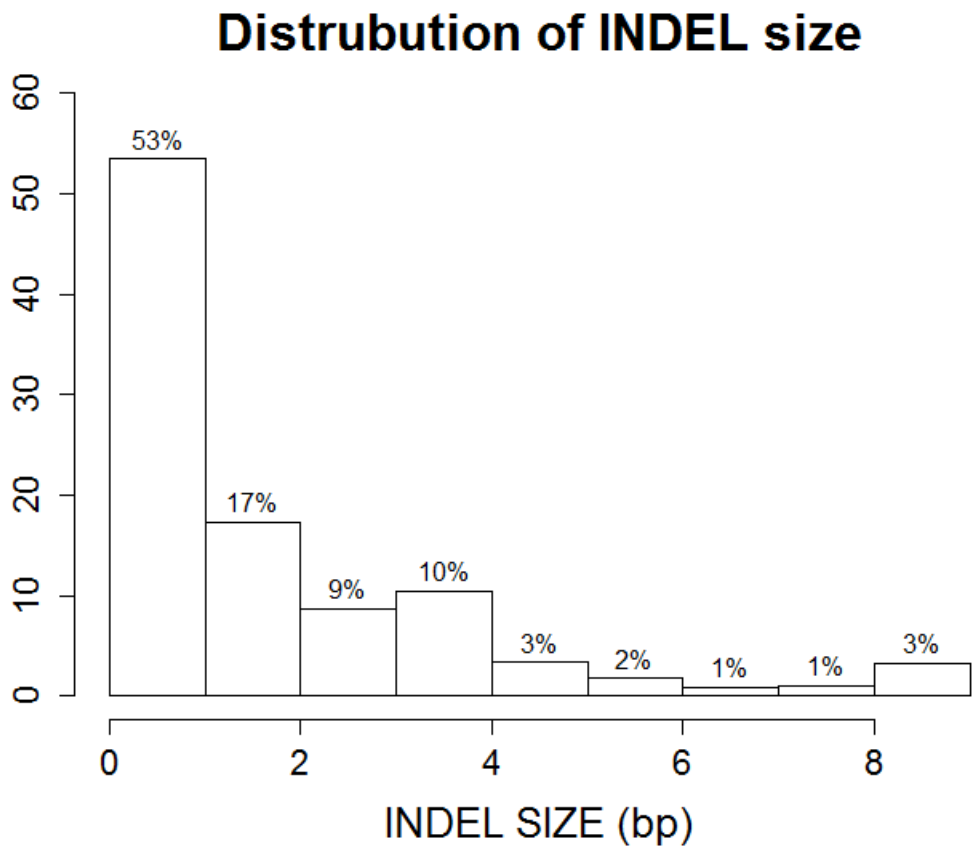

Distrubution of INDEL size
0
10
20
30
40
50
60
53%
17%
9%
10%
3%
2%
1%
1%
3%
0
2
4
6
8
INDEL SIZE (bp)

Figure 1c.

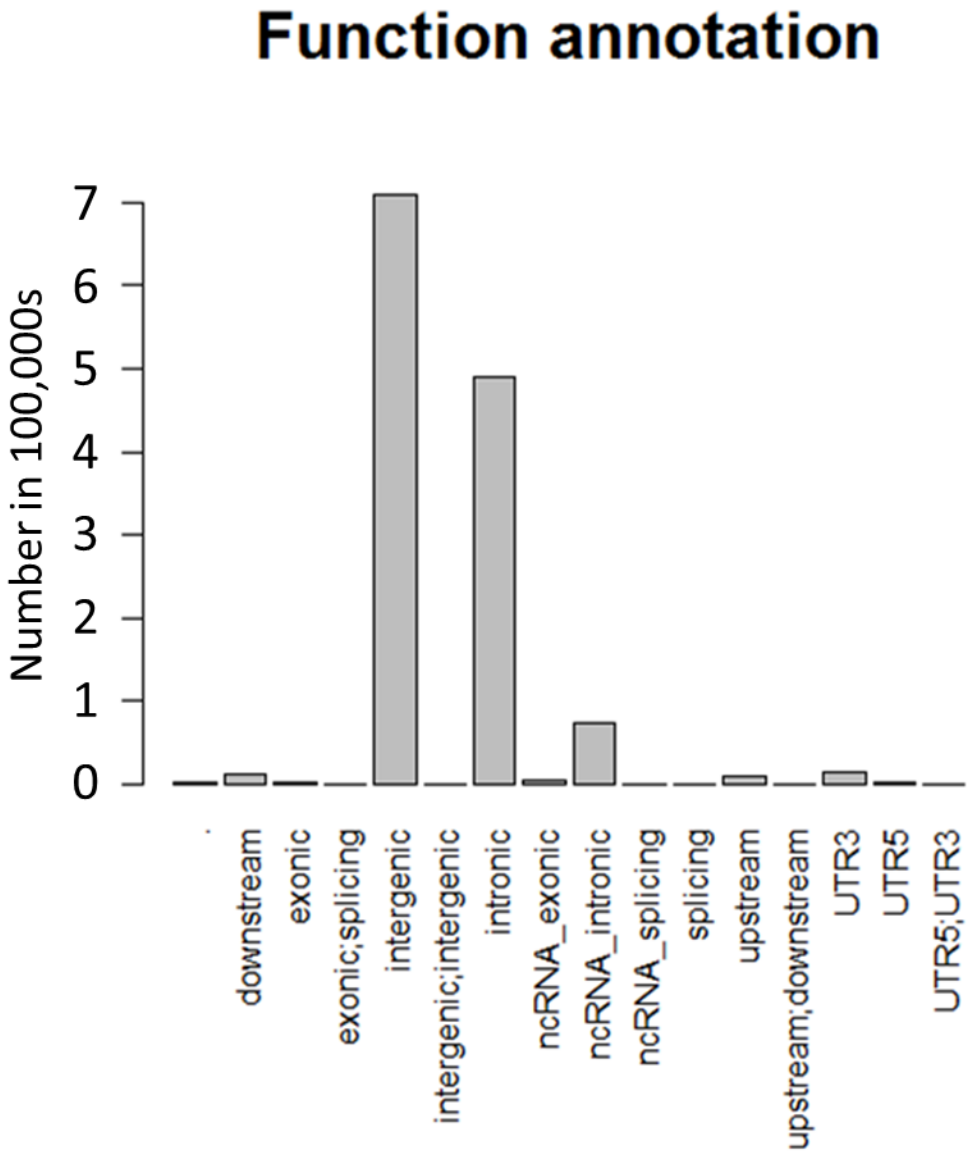

Function annotation
Number in 100,000s
0
1
2
3
4
5
6
7
downstream
exonic
exonic;splicing
intergenic
intergenic;intergenic
intronic
ncRNA_exonic
ncRNA_intronic
ncRNA_splicing
splicing
upstream
upstream;downstream
UTR3
UTR5
UTR5;UTR3

Figure 2. The evaluation of INDELs imputation quality using exome sequencing data. a. the distribution of kappa for all tested variants (N=9,370); b. the distribution of kappa for common (MAF≥0.05) variants (N=3,280).

Figure 2a.

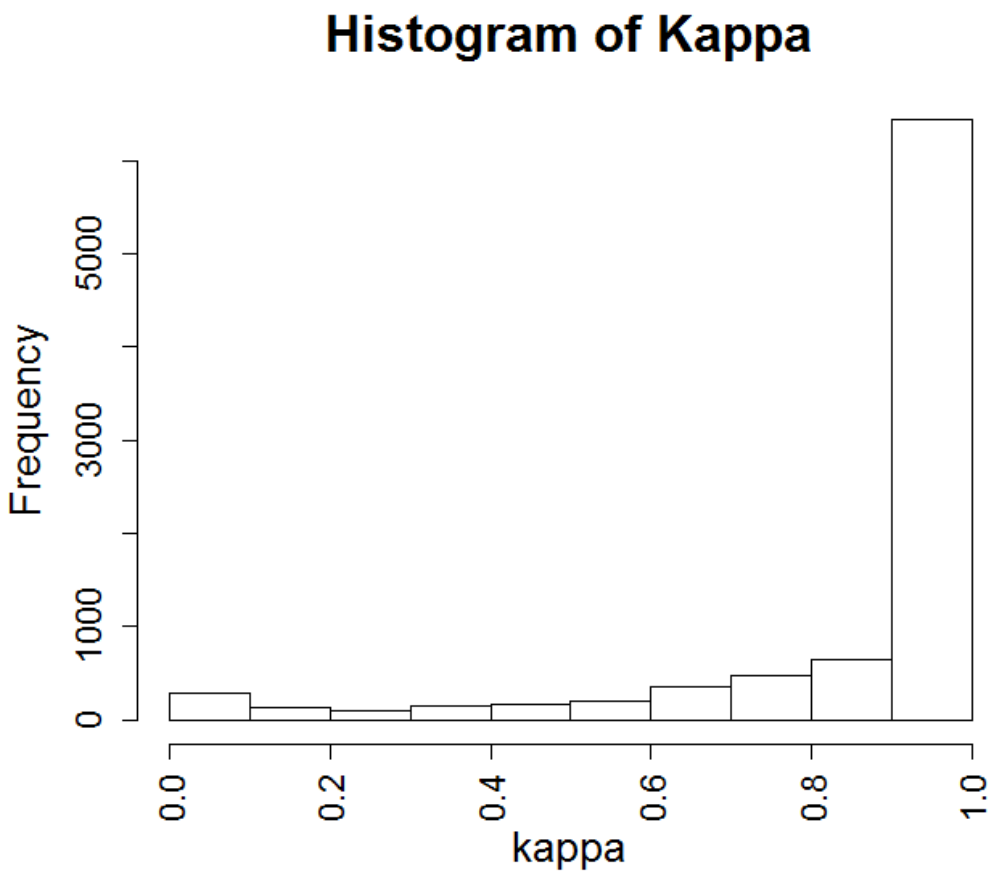

Histogram of Kappa
Frequency
0
1000
3000
5000
0.0
0.2
0.4
0.6
0.8
1.0
kappa

Figure 2b.

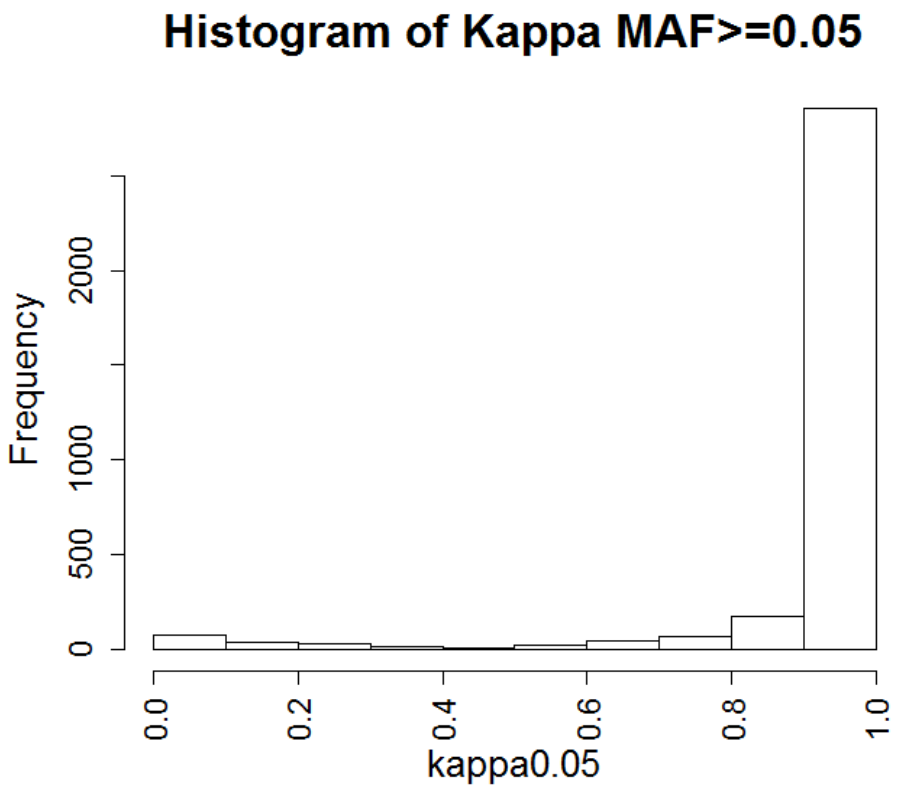

Histogram of Kappa MAF>=0.05
Frequency
0
500
1000
2000
0.0
0.2
0.4
0.6
0.8
1.0
kappa0.05

Figure 3. The distribution of INDELs nominally associated (P<0.05) and linked (LOD>0.5) with HDL a. INDEL size distributions, b. function annotations, c. MAF.

Figure 3a.

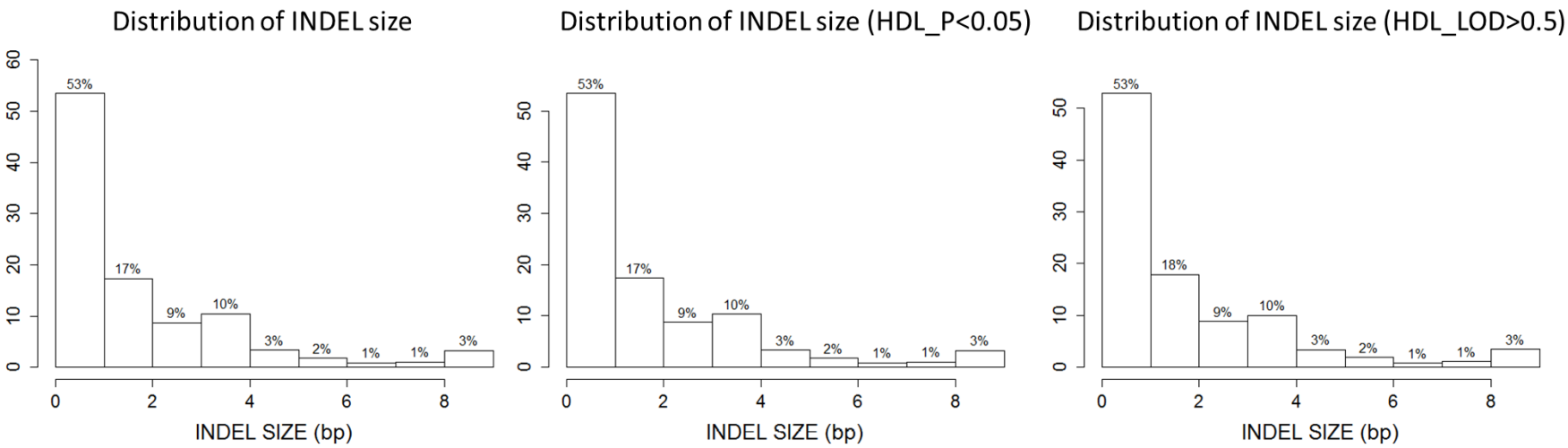

Distribution of INDEL size
53%
17%
9%
10%
3%
2%
1%
1%
3%
0
10
20
30
40
50
60
0
2
4
6
8
INDEL SIZE (bp)
Distribution of INDEL size (HDL_P<0.05)
53%
17%
9%
10%
3%
2%
1%
1%
3%
0
10
20
30
40
50
0
2
4
6
8
INDEL SIZE (bp)
Distribution of INDEL size (HDL_LOD>0.5)
53%
18%
9%
10%
3%
2%
1%
1%
3%
0
10
20
30
40
50
0
2
4
6
8
INDEL SIZE (bp)

Figure 3b.

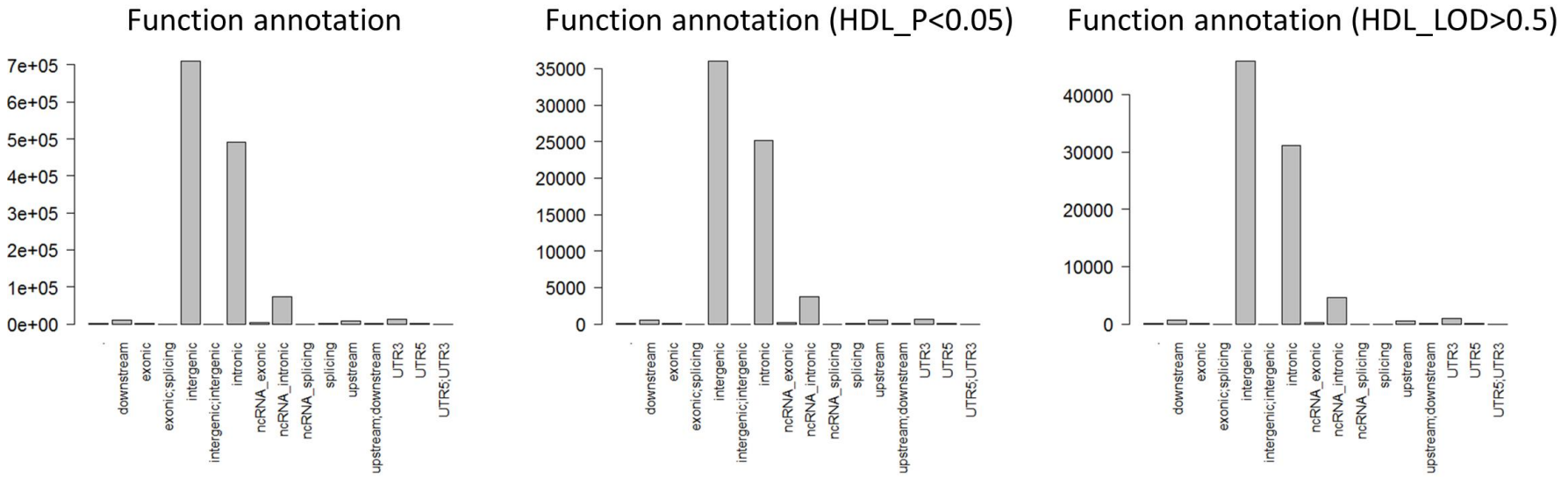
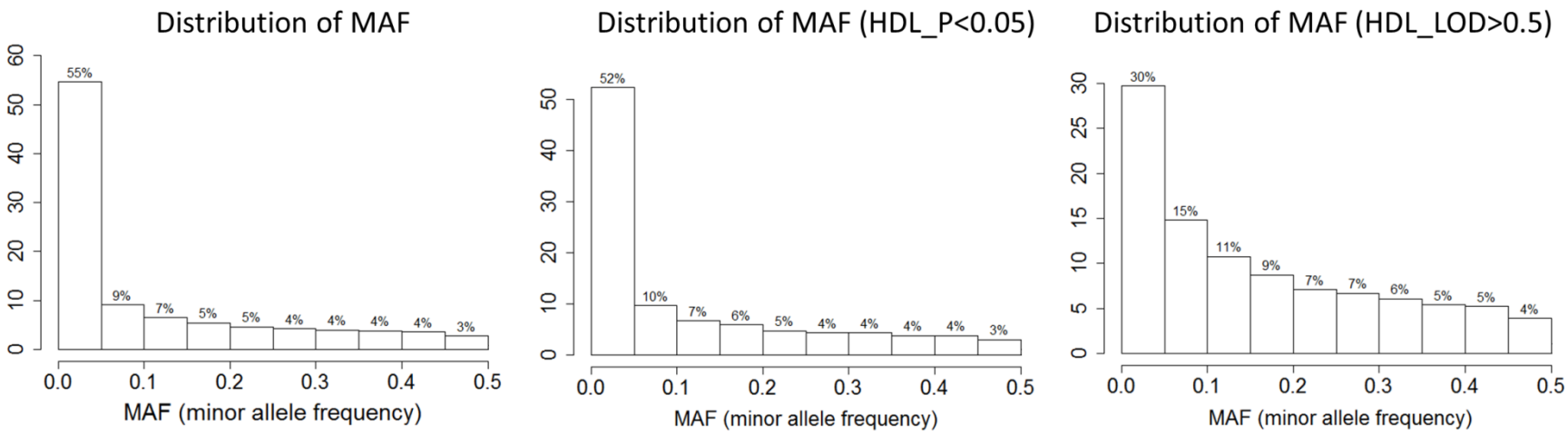

Function annotation
7e+05
6e+05
5e+05
4e+05
3e+05
2e+05
1e+05
0e+00
downstream
exonic
exonic;splicing
intergenic
intergenic;intergenic
intronic
ncRNA_exonic
ncRNA_intronic
ncRNA_splicing
splicing
upstream
upstream;downstream
UTR3
UTR5
UTR5;UTR3
Function annotation (HDL_P<0.05)
35000
30000
25000
20000
15000
10000
5000
0
Function annotation (HDL_LOD>0.5)
40000
30000
20000
10000
0

Figure 3c.


Distribution of MAF
55%
9%
7%
5%
5%
4%
4%
4%
4%
3%
0 10 20 30 40 50 60
0.0 0.1 0.2 0.3 0.4 0.5
MAF (minor allele frequency)
Distribution of MAF (HDL_P<0.05)
52%
10%
7%
6%
5%
4%
4%
4%
4%
3%
0 10 20 30 40 50
0.0 0.1 0.2 0.3 0.4 0.5
MAF (minor allele frequency)
Distribution of MAF (HDL_LOD>0.5)
30%
15%
11%
9%
7%
7%
6%
5%
5%
4%
0 5 10 15 20 25 30
0.0 0.1 0.2 0.3 0.4 0.5
MAF (minor allele frequency)

Chapter VI

DISCUSSION

Chuan Gao prepared this chapter. Dr. Nicholette Allred acted in an advisory and editorial capacity.

Cardiometabolic syndrome (CMS) is a clustering of interrelated risk factors that lead to substantial increase in CVD morbidity and mortality (Govindarajan, Whaley-Connell, Mugo, Stump, & Sowers, 2005). It is estimated that around 25% of the world's adults are suffering from CMS (Srivastava, 2012). Large numbers of genetic association studies have been performed yet the majority of the disease heritability still remains unexplained (Locke, et al., 2015; Willer, et al., 2013). This thesis aims to evaluate the genetic architecture of CMS risk factors and search for novel genetic signals that are responsible for the phenotypic variance. Through novel phenotype examination and evaluation of multiple categories of genetic variation using complimentary statistical approaches, we have evaluated the source of the missing heritability.

Genotyping Platforms

In order to investigate the full spectrum of allele frequencies, we employed multiple genotyping platforms, e.g. GWAS bead array, exome chip, GWAS imputation, and exome sequencing. By design, each of the platforms has a unique spectrum of the overage of the genome: GWAS provide an unbiased scan of common variants with regional linkage disequilibrium (LD); exome chip is a cost effective strategy to selectively focus on coding variants; imputation aims to increase the coverage of un-typed variants based on the observations of allele haplotypes; exome sequencing seeks to detect all variants within the exome.

In Chapter II and III, two GWAS bead arrays (Illumina, San Diego, CA) were used (OmniExpress and Omni1S). The OmniExpress array was designed as a powerful tool to provide a thorough coverage of common variants (MAF > 5%). As a complement, Omni1S array extended the coverage to low frequency variants (MAF > 2.5%). These arrays were optimized based on the

HapMap project and the SNPs were strategically selected to capture the greatest amount of common genetic variations across the entire genome. However, limitations exist for the application of those arrays in this thesis. For example, those arrays were designed based on the LD structures in Europeans and the selected SNPs may not tag the LD structure as well in Mexican Americans. Also, the majority of the variants on these arrays are common and nonfunctional, providing very limited information for disease risk prediction (Choquet & Meyre, 2011; Locke, et al., 2015). For example in Chapter II, *ZGRF1* was identified with significant associations with BMI and WAIST. However, the top SNPs identified in the region were mostly intronic and carry very low effect sizes, making it difficult for biological inference.

To complement the GWAS approach, the exome bead array was also included in this thesis. Different from GWAS, exome chip delivers decent coverage of putative functional exonic regions in a cost effective manner. The variants selected are mostly coding with potential larger effects, making it easier for interpretation. In Chapter II, SNP rs34218846 was identified by exome chip and marks a valine to isoleucine amino acid change to *IDH1*. Compared to the *ZGRF1* GWAS signal (MAF=48%), the *IDH1* signal has a lower allele frequency (MAF=6.8%). In addition, the biological function of *IDH1* also aligned with the association signals we observed: knockdown of IDH1 expression by RNA interference (RNAi) was able to inhibit adipocyte differentiation and lipogenesis in 3T3-L1 preadipocytes; in diet-induced obese mice transduced with IDH1 short-hairpin RNA, a loss of body weight and reduction of triglyceride levels were observed (Nam, et al., 2012). In summary, these results suggested that exome chip is an efficient approach to investigate the function of rare coding variants in a cost-effective manner.

Compared to the entire human genome, the exome region constitutes a very small proportion (1% of the total genome). To increase the coverage of un-typed low frequency variants (MAF between 1% and 5%), statistical imputation was performed based on GWAS genotypes, i.e.

statistical inference of unobserved genotypes based on 1000G Phase I reference panels (Abecasis, et al., 2012; Marchini & Howie, 2010; Scheet & Stephens, 2006). For example in IRASFS, the number of variants increased from around 1.6 million to 27 million after imputation. The additional genotypes significantly increased the coverage and fine-mapped association signals. For example, In Chapter III, most of the index SNPs came from imputation; in Chapter IV, all of the INDELs analyzed also came from imputation. Another advantage of the imputation approach is it does not incur the additional cost of lab materials, i.e. it was entirely calculated using statistical models. However, since all imputed variants were estimated based on statistical algorithms and the limited number of observations for rare allele haplotypes, imputation quality can vary for rare variants. For example, in Chapter III, only a subset of the imputed variants (~8 million) was analyzed for the associations with CT measures; in Chapter V, we noticed the imputation quality started to vary when allele frequency drops below 5%. In addition, imputation can only calculate the variants that are present in the reference panel, i.e. private variants specific to a population or a family cannot be imputed. For example, many of the rare variants identified by exome sequencing in Chapter IV were unsuccessfully or poorly imputed in IRASFS.

In Chapter IV, exome sequencing was introduced to investigate rare variant associations with cardiometabolic phenotypes. In contrast to exome chips, exome sequencing is able to sequence all expressed genes and detect every genetic polymorphism within the captured retion regardless of allele frequency, size, and position, e.g. in IRASFS, 81,599 and 555,651 variants were successfully identified and analyzed in exome chip and exome sequencing, respectively. We hypothesized that exome sequencing has the potential to identify novel non-synonymous or loss-of-function variants with significant associations to disease phenotypes. These associations can provide valuable information and help researchers understand the gene-disease interaction

and predict disease susceptibility. Overall, more than 555,651 variants were identified (40% were exonic coding variants) yet very few functional variant associations were detected. One possible explanation is the lack of statistical power, i.e. these functional variants have large effects on protein function and therefore are more likely to be excluded during evolution, resulting in extremely low allele frequencies. Another interesting observation was, despite the focus of the exome coding regions, many associated signals identified were non-coding, e.g. intronic variants. For example, among the six top signals identified in Chapter IV, five were intronic and one was a synonymous coding variant. This is a very meaningful observation because it suggests that the protein coding variants may be very rare in the general population and less likely to be the major contributors to disease susceptibility. In other words, transcription and translation regulators could be essential to disease risk in the general population, e.g. transcription factors, enhancers, micro RNAs, etc.

In conclusion, genotyping arrays are able to examine common variants across the genome. On the other hand, exome sequencing has a much higher resolution yet only detects the exonic regions. Therefore whole genome sequencing (WGS), despite the high cost and large data volume, has been proposed as the ultimate approach to shed light on the darkness of missing heritability (Manolio, et al., 2009).

INDEL Variant Analysis

In Chapter V, 1,273,952 INDELs were successfully imputed (confidence score>0.9, info score>0.4) based on IRASFS GWAS genotypes using the 1000G reference panel Phase I Version 3 (Abecasis, et al., 2012). To further validate the imputation quality, genotypes derived from imputation were compared with exome sequencing data using kappa statistic. Overall, common imputed

INDELs correlated well with exome sequencing data (kappa=0.93). However, the imputation quality starts to vary for less frequent INDELs (MAF<5%). This observation was not beyond expectation as imputation is known to have lower confidence for rare variants (Huang et al., 2015). Secondarily, we evaluated whether INDELs can be analyzed using traditional single variant analysis approaches. Genome-wide association and linkage analyses were performed for imputed INDELs using 15 cardiometabolic phenotypes. After excluding rare variants (MAF<1%), statistical models were well behaved with inflation factors ranging from 0.96-1.08 and examples of novel INDEL signals and previously signals were identified.

Despite the exciting results from the analysis of imputed INDELs, no novel INDEL signals were found, i.e. the phenotypic variance explained by genome-wide common SNPs (N=5,650,462, MAF≥5%) only increased 0.04% and 0.03% after including common INDELs (N=596,746, MAF≥5%) for HDL and TG, respectively. The major reason for that was all of the INDELs came from imputation, which was calculated based on the LD with proximal SNPs. Similar to rare SNP imputation, the imputation quality for rare INDELs is largely affected by limited proximal common variant haplotypes. Therefore, only INDELs present in the imputation panel and tagged by nearby SNPs were successfully imputed. For example, 38,122 INDELs were captured by exome sequencing yet only 9,370 were also successfully imputed based on GWAS genotypes, i.e. about 75% of the INDELs remained undetected after imputation. This is further complicated by the lack of efficient algorithms for INDEL detection from sequencing data, i.e. current calling algorithms often have trouble detecting INDELs (Ghoneim, Myers, Tuttle, & Paciorkowski, 2014). In conclusion, although there is no doubt INDELs have the potential to explain additional missing heritability in complex disease, INDEL research is very challenging based on current techniques. To fully interrogate the function of INDELs in a genome-wide scale, longer sequencing reads and better calling algorithms are needed.

Advanced phenotypes of obesity

In this thesis, a total of six obesity phenotypes were examined. In addition to the three anthropometric traits (BMI, WAIST, WHR), CT and DEXA were used for more accurate and comprehensive measures of obesity. Limitations exist for anthropometric measures, i.e. BMI is not a direct measure of fat deposition, which is closely linked to health outcomes; WHR and WAIST have been well-recognized as complementary approaches to estimate fat deposition, yet they are often skewed by age and skeletal structure (Taylor, et al., 2000). Also, BMI and WAIST are known to be very sensitive to environmental factors, e.g. dietary habits, life style. However, these environment effects are hard and expensive to capture and therefore are not included in most genetics studies.

Fat distribution between SAT and VAT has been shown to be responsible for the increased mortality and risk for metabolic disorders and is less sensitive to environmental factors, i.e. the volume of VAT has been suggested to be a strong risk factor for metabolic syndromes while SAT is benign (P. Cohen, et al., 2014; Lapidus, et al., 1984; Larsson, et al., 1984; Wajchenberg, 2000). Therefore, we hypothesized that CT yields more reliable and robust measures of SAT and VAT and therefore have increased power to detect novel signals with clear biological mechanisms.

However, as shown in the results in Chapter II, most of the significant signals came from traditional anthropometric measures. One reason was that 20-30% did not consent for CT and DEXA scans and therefore were excluded from the analysis. Also, compared to classical anthropometric measures, CT and DEXA measures were observed to have a much higher variance (Chapter II, **Table 1**), resulting in a decreased power for association analysis. In addition, due to economic reasons, not many cohorts have these sophisticated phenotypes and it was

difficult to find replication resources. Another notable limitation for CT in this study was that SAT and VAT were calculated based on a single scout of the abdomen followed by a 10-mm thick axial image at L4-L5 disc space. Therefore, it may not be able to capture the whole body fat distribution well.

Gene-sex Stratification and Interaction

It is a common observation that males and females have different patterns of adipose tissue distribution, i.e. females have a higher proportion of gluteal-femoral body fat whereas males have more in the abdominal (visceral) region (Blaak, 2001; Zillikens, et al., 2008). To investigate the genetic mechanisms of sex specificity, we performed sex stratified analyses paralleled with formal gene-sex interaction analyses (Chapter III). By design, the two approaches have different power to detect signals with differential origins: the sex stratified approach is able to detect sex-specific signals; the interaction approach is able to take advantage of the entire sample and search for signals with different effect sizes between the two groups. Overall, the two approaches yielded slightly different yet very consistent results (Chapter III **Table 2**). However, both approaches suffer a similar power issue, both approaches require further stratification of the sample and therefore result in a reduced power compared main effect tests. For future analysis, a joint analysis approach has the potential to overcome the power limitation (Dick, 2011). Different from either of the two approaches, the joint test is a two degree-of-freedom test that is able to combine both the genetic effect and the interaction effect. Simulation studies have suggested that the joint test approach has increased power compared to the interaction approach and has the potential to identify novel genetic signals (Dick, 2011).

Association and Linkage Analysis

Association and linkage are both powerful approaches to detect significant genetic signals regulating disease-related phenotypes. In this thesis, the two approaches were utilized and compared across multiple genotyping platforms. Overall compared to linkage, association has multiple advantages. First, association analysis is computationally efficient. For example, linkage analysis requires IBD estimation and that is extremely computationally expensive. On average, linkage analysis takes about ten times the amount of resources (time and storage space) as compared to association. Second, association aligns well with GWAS imputation analysis pipelines, i.e. dosage genotypes can be used directly for association. On the other hand, imputed posterior probabilities have to be converted to best guess genotypes in order to be analyzed for linkage. Third, association has been shown to be reliable and sensitive to small-effect signals and multiple meta-analysis approaches with association summary statistics are available. However, linkage signals are very sensitive to incomplete penetrance and therefore they are only ideal for large effect signals. Also, there is still not yet a well-accepted approach for linkage meta-analysis.

Despite the limitations of linkage analysis, we also observed promising results for applying linkage in combination with rare variant analysis, e.g. variants from exome sequencing (Chapter IV). Different from common variants evaluated in the GWAS setting, functional exome sequencing variants are hypothesized to have very low allele frequencies, which makes them difficult to detect with association. For linkage, the power is largely dependent on family structures as well as the number of segregation events, e.g. a rare variant in the population can be very common in a given family, making linkage a more powerful approach compared to association.

Conclusion

In conclusion, this thesis explored the hypotheses that naturally occurring DNA variations are responsible for the disease heritability of CMS in Mexican Americans. Genetic association and linkage analyses were performed for common and rare SNPs with plasma lipid levels and obesity related phenotypes in Mexican Americans from IRASFS. In addition, INDEL variants were imputed based on GWAS genotypes and the imputation quality was validated using exome sequencing. CT and DEXA based comprehensive measures of adipose tissue deposition and distribution were evaluated and multiple novel signals were identified. These results emphasized the complexity of the genetics mechanisms, confirmed the importance of genetic factors in CMS susceptibility, and provided directions for future complex disease genetics research.

Reference

Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., . . . McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature, 491*(7422), 56-65. doi: 10.1038/nature11632
nature11632 [pii]

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., . . . Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nat Methods, 7*(4), 248-249. doi: 10.1038/nmeth0410-248
nmeth0410-248 [pii]

Al-Owain, M., Colak, D., Al-Bakheet, A., Al-Hashmi, N., Shuaib, T., Al-Hemidan, A., . . . Kaya, N. (2012). Novel mutation in GLRB in a large family with hereditary hyperekplexia. *Clin Genet, 81*(5), 479-484. doi: 10.1111/j.1399-0004.2011.01661.x

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res, 19*(9), 1655-1664. doi: 10.1101/gr.094052.109
gr.094052.109 [pii]

Almal, S. H., & Padh, H. (2012). Implications of gene copy-number variation in health and diseases. *J Hum Genet, 57*(1), 6-13. doi: 10.1038/jhg.2011.108
jhg2011108 [pii]

Almasy, L., & Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet, 62*(5), 1198-1211. doi: S0002-9297(07)61542-0 [pii]
10.1086/301844

An, S. S., Palmer, N. D., Hanley, A. J., Ziegler, J. T., Brown, W. M., Haffner, S. M., . . . Bowden, D. W. (2013). Estimating the contributions of rare and common genetic variations and clinical measures to a model trait: adiponectin. *Genet Epidemiol, 37*(1), 13-24. doi: 10.1002/gepi.21685

Awata, T., Yamashita, H., Kurihara, S., Morita-Ohkubo, T., Miyashita, Y., Katayama, S., . . . Mawatari, Y. (2013). A low-frequency GLIS3 variant associated with resistance to Japanese type 1 diabetes. *Biochem Biophys Res Commun, 437*(4), 521-525. doi: 10.1016/j.bbrc.2013.06.102
S0006-291X(13)01112-1 [pii]

Bahram, S., & Inoko, H. (2007). Microsatellite markers for genome-wide association studies. *Nature Reviews Genetics, 8*(2).

Bashliev, I. (1987). [Temporary loss of work capacity in myocardial infarct patients who underwent rehabilitation]. *Vutr Boles, 26*(6), 45-50.

Below, J. E., Parra, E. J., Gamazon, E. R., Torres, J., Krithika, S., Candille, S., . . . Valladares-Salgado, A. (2016). Meta-analysis of lipid-traits in Hispanics identifies novel loci, population-specific effects, and tissue-specific enrichment of eQTLs. *Sci Rep, 6*, 19429. doi: 10.1038/srep19429
srep19429 [pii]

Benjamin, E. J., Blaha, M. J., Chiuve, S. E., Cushman, M., Das, S. R., Deo, R., . . . Muntner, P. (2017). Heart Disease and Stroke Statistics-2017 Update: A Report From the American Heart Association. *Circulation*. doi: CIR.0000000000000485 [pii]
10.1161/CIR.0000000000000485

Bergman, R. N., Finegood, D. T., & Ader, M. (1985). Assessment of insulin sensitivity in vivo. *Endocr Rev, 6*(1), 45-86. doi: 10.1210/edrv-6-1-45

Berndt, S. I., Gustafsson, S., Magi, R., Ganna, A., Wheeler, E., Feitosa, M. F., . . . Ingelsson, E. (2013). Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nat Genet, 45*(5), 501-512. doi: 10.1038/ng.2606
ng.2606 [pii]
Blaak, E. (2001). Gender differences in fat metabolism. *Curr Opin Clin Nutr Metab Care, 4*(6), 499-502.
Black, M. H., Fingerlin, T. E., Allayee, H., Zhang, W., Xiang, A. H., Trigo, E., . . . Watanabe, R. M. (2008). Evidence of interaction between PPARG2 and HNF4A contributing to variation in insulin sensitivity in Mexican Americans. *Diabetes, 57*(4), 1048-1056. doi: db07-0848 [pii]
10.2337/db07-0848
Blake JA, B. C., Eppig JT, Kadin JA, Richardson JE; The Mouse Genome Database Group. . (2014). The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. *Nucleic Acids Res, 2014 42(D1):D810-D817*.
Botstein, D., & Risch, N. (2003). Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet, 33 Suppl*, 228-237. doi: 10.1038/ng1090
ng1090 [pii]
Bowden, D. W., An, S. S., Palmer, N. D., Brown, W. M., Norris, J. M., Haffner, S. M., . . . Langefeld, C. D. (2010). Molecular basis of a linkage peak: exome sequencing and family-based analysis identify a rare genetic variant in the ADIPOQ gene in the IRAS Family Study. *Hum Mol Genet, 19*(20), 4112-4120. doi: 10.1093/hmg/ddq327
ddq327 [pii]
Bradfield, J. P., Taal, H. R., Timpson, N. J., Scherag, A., Lecoeur, C., Warrington, N. M., . . . Grant, S. F. (2012). A genome-wide association meta-analysis identifies new childhood obesity loci. *Nat Genet, 44*(5), 526-531. doi: 10.1038/ng.2247
ng.2247 [pii]
Browning, B. L., Brashear, D. L., Butler, A. A., Cyr, D. D., Harris, E. C., Nelsen, A. J., . . . Wagner, M. J. (2004). Linkage analysis using single nucleotide polymorphisms. *Hum Hered, 57*(4), 220-227. doi: 81449 [pii]
10.1159/000081449
Buchanan, T. A., Xiang, A. H., Peters, R. K., Kjos, S. L., Berkowitz, K., Marroquin, A., . . . Azen, S. P. (2000). Response of pancreatic beta-cells to improved insulin sensitivity in women at high risk for type 2 diabetes. *Diabetes, 49*(5), 782-788.
Buchanan, T. A., Xiang, A. H., Peters, R. K., Kjos, S. L., Marroquin, A., Goico, J., . . . Azen, S. P. (2002). Preservation of pancreatic beta-cell function and prevention of type 2 diabetes by pharmacological treatment of insulin resistance in high-risk hispanic women. *Diabetes, 51*(9), 2796-2803.
Buimer, M., Keijser, R., Jebbink, J. M., Wehkamp, D., van Kampen, A. H., Boer, K., . . . Ris-Stalpers, C. (2008). Seven placental transcripts characterize HELLP-syndrome. *Placenta, 29*(5), 444-453. doi: 10.1016/j.placenta.2008.02.007
S0143-4004(08)00072-6 [pii]
Cannon, C. P. (2007). Cardiovascular disease and modifiable cardiometabolic risk factors. *Clin Cornerstone, 8*(3), 11-28. doi: S1098-3597(07)80025-1 [pii]
Cantsilieris, S., & White, S. J. (2013). Correlating multiallelic copy number polymorphisms with disease susceptibility. *Hum Mutat, 34*(1), 1-13. doi: 10.1002/humu.22172

Carithers, L. J., & Moore, H. M. (2015). The Genotype-Tissue Expression (GTEx) Project. *Biopreserv Biobank, 13*(5), 307-308. doi: 10.1089/bio.2015.29031.hmm
Castro, J. P., El-Atat, F. A., McFarlane, S. I., Aneja, A., & Sowers, J. R. (2003). Cardiometabolic syndrome: pathophysiology and treatment. *Curr Hypertens Rep, 5*(5), 393-401.
Cawley, J., & Meyerhoefer, C. (2012). The medical care costs of obesity: An instrumental variables approach. *Journal of Health Economics, 31*(1), 219-230. doi: 10.1016/j.jhealeco.2011.10.003
CDC. (2013). Deaths, percent of total deaths, and death rates for the 15 leading causes of death in 10-year age groups, by race and sex: United States, 2013.
CDC. (2017). https://www.cdc.gov/heartdisease/facts.htm. Accessed March 14, 2017.
Chelbi, S. T., Doridot, L., Mondon, F., Dussour, C., Rebourcet, R., Busato, F., . . . Vaiman, D. (2011). Combination of promoter hypomethylation and PDX1 overexpression leads to TBX15 decrease in vascular IUGR placentas. *Epigenetics, 6*(2), 247-255. doi: 13791 [pii]
Cheng, L. S., Davis, R. C., Raffel, L. J., Xiang, A. H., Wang, N., Quinones, M., . . . Rotter, J. I. (2001). Coincident linkage of fasting plasma insulin and blood pressure to chromosome 7q in hypertensive hispanic families. *Circulation, 104*(11), 1255-1260.
Choquet, H., & Meyre, D. (2011). Genetics of Obesity: What have we Learned? *Curr Genomics, 12*(3), 169-179. doi: 10.2174/138920211795677895
CG-12-169 [pii]
Christians, J. K., de Zwaan, D. R., & Fung, S. H. (2013). Pregnancy associated plasma protein A2 (PAPP-A2) affects bone size and shape and contributes to natural variation in postnatal growth in mice. *PLoS One, 8*(2), e56260. doi: 10.1371/journal.pone.0056260
PONE-D-12-32294 [pii]
Cohen, J. C., Boerwinkle, E., Mosley, T. H., Jr., & Hobbs, H. H. (2006). Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N Engl J Med, 354*(12), 1264-1272. doi: 354/12/1264 [pii]
10.1056/NEJMoa054013
Cohen, P., Levy, J. D., Zhang, Y., Frontini, A., Kolodin, D. P., Svensson, K. J., . . . Spiegelman, B. M. (2014). Ablation of PRDM16 and beige adipose causes metabolic dysfunction and a subcutaneous to visceral fat switch. *Cell, 156*(1-2), 304-316. doi: 10.1016/j.cell.2013.12.021
S0092-8674(13)01590-0 [pii]
Comuzzie, A. G., Cole, S. A., Laston, S. L., Voruganti, V. S., Haack, K., Gibbs, R. A., & Butte, N. F. (2012). Novel genetic loci identified for the pathophysiology of childhood obesity in the Hispanic population. *PLoS One, 7*(12), e51954. doi: 10.1371/journal.pone.0051954
PONE-D-12-24984 [pii]
Conover, C. A., Boldt, H. B., Bale, L. K., Clifton, K. B., Grell, J. A., Mader, J. R., . . . Powell, D. R. (2011). Pregnancy-associated plasma protein-A2 (PAPP-A2): tissue expression and biological consequences of gene knockout in mice. *Endocrinology, 152*(7), 2837-2844. doi: 10.1210/en.2011-0036
en.2011-0036 [pii]
Costanza, M. C., Beer-Borst, S., James, R. W., Gaspoz, J. M., & Morabia, A. (2012). Consistency between cross-sectional and longitudinal SNP: blood lipid associations. *Eur J Epidemiol, 27*(2), 131-138. doi: 10.1007/s10654-012-9670-1
Craddock, N., Hurles, M. E., Cardin, N., Pearson, R. D., Plagnol, V., Robson, S., . . . Donnelly, P. (2010). Genome-wide association study of CNVs in 16,000 cases of

eight common diseases and 3,000 shared controls. *Nature, 464*(7289), 713-720. doi: 10.1038/nature08979
nature08979 [pii]
Crosley, E. J., Durland, U., Seethram, K., MacRae, S., Gruslin, A., & Christians, J. K. (2014). First-trimester levels of pregnancy-associated plasma protein A2 (PAPP-A2) in the maternal circulation are elevated in pregnancies that subsequently develop preeclampsia. *Reprod Sci, 21*(6), 754-760. doi: 10.1177/1933719113512532
1933719113512532 [pii]
Dewey, F. E., Murray, M. F., Overton, J. D., Habegger, L., Leader, J. B., Fetterolf, S. N., . . . Carey, D. J. (2016). Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science, 354*(6319). doi: aaf6814 [pii]
10.1126/science.aaf6814
354/6319/aaf6814 [pii]
Dhawan, D., & Padh, H. (2009). Pharmacogenetics: technologies to detect copy number variations. *Curr Opin Mol Ther, 11*(6), 670-680.
Dichgans, M., Malik, R., Konig, I. R., Rosand, J., Clarke, R., Gretarsdottir, S., . . . Schunkert, H. (2014). Shared genetic susceptibility to ischemic stroke and coronary artery disease: a genome-wide analysis of common variants. *Stroke, 45*(1), 24-36. doi: 10.1161/STROKEAHA.113.002707
STROKEAHA.113.002707 [pii]
Dick, D. M. (2011). Gene-environment interaction in psychological traits and disorders. *Annu Rev Clin Psychol, 7*, 383-409. doi: 10.1146/annurev-clinpsy-032210-104518
Dikoglu, E., Simsek-Kiper, P. O., Utine, G. E., Campos-Xavier, B., Boduroglu, K., Bonafe, L., . . . Unger, S. (2013). Homozygosity for a novel truncating mutation confirms TBX15 deficiency as the cause of Cousin syndrome. *Am J Med Genet A, 161A*(12), 3161-3165. doi: 10.1002/ajmg.a.36173
Do, R., Stitziel, N. O., Won, H. H., Jorgensen, A. B., Duga, S., Angelica Merlini, P., . . . Kathiresan, S. (2015). Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. *Nature, 518*(7537), 102-106. doi: 10.1038/nature13917
nature13917 [pii]
Flegal, K. M., Carroll, M. D., Kit, B. K., & Ogden, C. L. (2012). Prevalence of obesity and trends in the distribution of body mass index among US adults, 1999-2010. *JAMA, 307*(5), 491-497. doi: 10.1001/jama.2012.39
jama.2012.39 [pii]
Fox, C. S., Liu, Y., White, C. C., Feitosa, M., Smith, A. V., Heard-Costa, N., . . . Borecki, I. B. (2012). Genome-wide association for abdominal subcutaneous and visceral adipose reveals a novel locus for visceral fat in women. *PLoS Genet, 8*(5), e1002695. doi: 10.1371/journal.pgen.1002695
PGENETICS-D-11-02103 [pii]
Fox, C. S., White, C. C., Lohman, K., Heard-Costa, N., Cohen, P., Zhang, Y., . . . Liu, Y. (2012). Genome-wide association of pericardial fat identifies a unique locus for ectopic fat. *PLoS Genet, 8*(5), e1002705. doi: 10.1371/journal.pgen.1002705
PGENETICS-D-11-02554 [pii]
Frayling, T. M., Timpson, N. J., Weedon, M. N., Zeggini, E., Freathy, R. M., Lindgren, C. M., . . . McCarthy, M. I. (2007). A Common Variant in the FTO Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity. *Science, 316*(5826), 889-894. doi: 10.1126/science.1141634

Gao, C., Wang, N., Guo, X., Ziegler, J. T., Taylor, K. D., Xiang, A. H., . . . Palmer, N. D. (2015). A Comprehensive Analysis of Common and Rare Variants to Identify Adiposity Loci in Hispanic Americans: The IRAS Family Study (IRASFS). *PLoS One, 10*(11), e0134649. doi: 10.1371/journal.pone.0134649
PONE-D-15-17356 [pii]
Ghoneim, D. H., Myers, J. R., Tuttle, E., & Paciorkowski, A. R. (2014). Comparison of insertion/deletion calling algorithms on human next-generation sequencing data. *BMC Res Notes, 7*, 864. doi: 10.1186/1756-0500-7-864
1756-0500-7-864 [pii]
Gibson, G. (2011). Rare and common variants: twenty arguments. *Nat Rev Genet, 13*(2), 135-145. doi: 10.1038/nrg3118
nrg3118 [pii]
Go, A. S., Mozaffarian, D., Roger, V. L., Benjamin, E. J., Berry, J. D., Borden, W. B., . . . Turner, M. B. (2013). Heart disease and stroke statistics--2013 update: a report from the American Heart Association. *Circulation, 127*(1), e6-e245. doi: 10.1161/CIR.0b013e31828124ad
CIR.0b013e31828124ad [pii]
Going, S. B., Lohman, T. G., Cussler, E. C., Williams, D. P., Morrison, J. A., & Horn, P. S. (2011). Percent body fat and chronic disease risk factors in U.S. children and youth. *Am J Prev Med, 41*(4 Suppl 2), S77-86. doi: 10.1016/j.amepre.2011.07.006
S0749-3797(11)00492-2 [pii]
Goodarzi, M. O., Guo, X., Taylor, K. D., Quinones, M. J., Saad, M. F., Yang, H., . . . Rotter, J. I. (2004). Lipoprotein lipase is a gene for insulin resistance in Mexican Americans. *Diabetes, 53*(1), 214-220.
Goodarzi, M. O., Guo, X., Taylor, K. D., Quinones, M. J., Samayoa, C., Yang, H., . . . Rotter, J. I. (2003). Determination and use of haplotypes: ethnic comparison and association of the lipoprotein lipase gene and coronary artery disease in Mexican-Americans. *Genet Med, 5*(4), 322-327. doi: 10.1097/01.GIM.0000076971.55421.AD
Goodarzi, M. O., Langefeld, C. D., Xiang, A. H., Chen, Y. D., Guo, X., Hanley, A. J., . . . Wagenknecht, L. E. (2014). Insulin sensitivity and insulin clearance are heritable and have strong genetic correlation in Mexican Americans. *Obesity (Silver Spring), 22*(4), 1157-1164. doi: 10.1002/oby.20639
Goodarzi, M. O., Taylor, K. D., Guo, X., Quinones, M. J., Cui, J., Li, X., . . . Rotter, J. I. (2005). Variation in the gene for muscle-specific AMP deaminase is associated with insulin clearance, a highly heritable trait. *Diabetes, 54*(4), 1222-1227. doi: 54/4/1222 [pii]
Govindarajan, G., Whaley-Connell, A., Mugo, M., Stump, C., & Sowers, J. R. (2005). The cardiometabolic syndrome as a cardiovascular risk factor. *Am J Med Sci, 330*(6), 311-318. doi: S0002-9629(15)32882-2 [pii]
Grant, S. F. A., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Manolescu, A., Sainz, J., . . . Stefansson, K. (2006). Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nature Genetics, 38*(3), 320-323. doi: 10.1038/ng1732
Han, B., & Eskin, E. (2011). Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am J Hum Genet, 88*(5), 586-598. doi: 10.1016/j.ajhg.2011.04.014
S0002-9297(11)00155-8 [pii]
Haslam, D. W., & James, W. P. T. (2005). Obesity. *The Lancet, 366*(9492), 1197-1209. doi: 10.1016/s0140-6736(05)67483-1

Hasumi, H., Baba, M., Hasumi, Y., Lang, M., Huang, Y., Oh, H. F., . . . Schmidt, L. S. (2015). Folliculin-interacting proteins Fnip1 and Fnip2 play critical roles in kidney tumor suppression in cooperation with Flcn. *Proc Natl Acad Sci U S A, 112*(13), E1624-1631. doi: 10.1073/pnas.1419502112
1419502112 [pii]
Heggeness, M. H., Esses, S. I., & Kostuik, J. P. (1991). Acquisition of lytic spondylolisthesis in the adult. *J Spinal Disord, 4*(4), 486-489.
Heid, I. M., Jackson, A. U., Randall, J. C., Winkler, T. W., Qi, L., Steinthorsdottir, V., . . . Lindgren, C. M. (2010). Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nat Genet, 42*(11), 949-960. doi: 10.1038/ng.685
ng.685 [pii]
Hellwege, J. N., Palmer, N. D., Raffield, L. M., Ng, M. C., Hawkins, G. A., Long, J., . . . Bowden, D. W. (2014). Genome-wide family-based linkage analysis of exome chip variants and cardiometabolic risk. *Genet Epidemiol, 38*(4), 345-352. doi: 10.1002/gepi.21801
Heng, J. I., Nguyen, L., Castro, D. S., Zimmer, C., Wildner, H., Armant, O., . . . Guillemot, F. (2008). Neurogenin 2 controls cortical neuron migration through regulation of Rnd2. *Nature, 455*(7209), 114-118. doi: 10.1038/nature07198
nature07198 [pii]
Henkin, L., Bergman, R. N., Bowden, D. W., Ellsworth, D. L., Haffner, S. M., Langefeld, C. D., . . . Rich, S. S. (2003). Genetic epidemiology of insulin resistance and visceral adiposity. The IRAS Family Study design and methods. *Ann Epidemiol, 13*(4), 211-217. doi: S104727970200412X [pii]
Herijgers, N., de Winther, M. P., Van Eck, M., Havekes, L. M., Hofker, M. H., Hoogerbrugge, P. M., & Van Berkel, T. J. (2000). Effect of human scavenger receptor class A overexpression in bone marrow-derived cells on lipoprotein metabolism and atherosclerosis in low density lipoprotein receptor knockout mice. *J Lipid Res, 41*(9), 1402-1409.
Hindorff LA, M. J. E. B. I., Morales J (European Bioinformatics Institute), Junkins HA, Hall PN, Klemm AK, and Manolio TA. A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies. Accessed 11/25/2014.
Hirschhorn, J. N., & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics, 6*(2), 95-108. doi: 10.1038/nrg1521
Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., & Abecasis, G. R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet, 44*(8), 955-959. doi: 10.1038/ng.2354
ng.2354 [pii]
Howie, B. N., Donnelly, P., & Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet, 5*(6), e1000529. doi: 10.1371/journal.pgen.1000529
Hu, Z., Shi, Y., Mo, X., Xu, J., Zhao, B., Lin, Y., . . . Shen, H. (2013). A genome-wide association study identifies two risk loci for congenital heart malformations in Han Chinese populations. *Nat Genet, 45*(7), 818-821. doi: 10.1038/ng.2636
ng.2636 [pii]
Huang, J., Howie, B., McCarthy, S., Memari, Y., Walter, K., Min, J. L., . . . Consortium, U. K. (2015). Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nature Communications, 6*. doi: Artn 8111
10.1038/Ncomms9111

Humphries, A. D., Streimann, I. C., Stojanovski, D., Johnston, A. J., Yano, M., Hoogenraad, N. J., & Ryan, M. T. (2005). Dissection of the mitochondrial import and assembly pathway for human Tom40. *J Biol Chem, 280*(12), 11535-11543. doi: M413816200 [pii]
10.1074/jbc.M413816200
An integrated encyclopedia of DNA elements in the human genome. (2012). *Nature, 489*(7414), 57-74. doi: 10.1038/nature11247
nature11247 [pii]
Jacquemont, S., Reymond, A., Zufferey, F., Harewood, L., Walters, R. G., Kutalik, Z., . . . Froguel, P. (2011). Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. *Nature, 478*(7367), 97-102. doi: 10.1038/nature10406
nature10406 [pii]
James, V. M., Bode, A., Chung, S. K., Gill, J. L., Nielsen, M., Cowan, F. M., . . . Harvey, R. J. (2013). Novel missense mutations in the glycine receptor beta subunit gene (GLRB) in startle disease. *Neurobiol Dis, 52*, 137-149. doi: 10.1016/j.nbd.2012.12.001
S0969-9961(12)00384-1 [pii]
Jensen, M. D., Ryan, D. H., Apovian, C. M., Ard, J. D., Comuzzie, A. G., Donato, K. A., . . . Tomaselli, G. F. (2014). 2013 AHA/ACC/TOS guideline for the management of overweight and obesity in adults: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and The Obesity Society. *Circulation, 129*(25 Suppl 2), S102-138. doi: 10.1161/01.cir.0000437739.71477.ee
01.cir.0000437739.71477.ee [pii]
Jeong, S. W., Chung, M., Park, S. J., Cho, S. B., & Hong, K. W. (2014). Genome-wide association study of metabolic syndrome in koreans. *Genomics Inform, 12*(4), 187-194. doi: 10.5808/GI.2014.12.4.187
Jia, J., Li, L., Zhao, Q., Zhang, L., Ru, J., Liu, X., . . . Shi, L. (2012). Association of a single nucleotide polymorphism in pregnancy-associated plasma protein-A2 with developmental dysplasia of the hip: a case-control study. *Osteoarthritis Cartilage, 20*(1), 60-63. doi: 10.1016/j.joca.2011.10.004
S1063-4584(11)00287-1 [pii]
Johnson, A. D., Handsaker, R. E., Pulit, S. L., Nizzari, M. M., O'Donnell, C. J., & de Bakker, P. I. (2008). SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics, 24*(24), 2938-2939. doi: 10.1093/bioinformatics/btn564
btn564 [pii]
Kapushesky, M., Emam, I., Holloway, E., Kurnosov, P., Zorin, A., Malone, J., . . . Brazma, A. (2010). Gene Expression Atlas at the European Bioinformatics Institute. *Nucleic Acids Research, 38*, D690-D698. doi: Doi 10.1093/Nar/Gkp936
Keith, S. W., Redden, D. T., Katzmarzyk, P. T., Boggiano, M. M., Hanlon, E. C., Benca, R. M., . . . Allison, D. B. (2006). Putative contributors to the secular increase in obesity: exploring the roads less traveled. *Int J Obes (Lond), 30*(11), 1585-1594. doi: 0803326 [pii]
10.1038/sj.ijo.0803326
Ken-Dror, G., Goldbourt, U., & Dankner, R. (2010). Different effects of apolipoprotein A5 SNPs and haplotypes on triglyceride concentration in three ethnic origins. *J Hum Genet, 55*(5), 300-307. doi: 10.1038/jhg.2010.27
jhg201027 [pii]

Kettunen, J., Tukiainen, T., Sarin, A. P., Ortega-Alonso, A., Tikkanen, E., Lyytikainen, L. P., . . . Ripatti, S. (2012). Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nat Genet, 44*(3), 269-276. doi: 10.1038/ng.1073
ng.1073 [pii]
Kilpelainen, T. O., Zillikens, M. C., Stancakova, A., Finucane, F. M., Ried, J. S., Langenberg, C., . . . Loos, R. J. (2011). Genetic variation near IRS1 associates with reduced adiposity and an impaired metabolic profile. *Nat Genet, 43*(8), 753-760. doi: 10.1038/ng.866
ng.866 [pii]
Kitahara, C. M., Flint, A. J., Berrington de Gonzalez, A., Bernstein, L., Brotzman, M., MacInnis, R. J., . . . Hartge, P. (2014). Association between class III obesity (BMI of 40-59 kg/m2) and mortality: a pooled analysis of 20 prospective studies. *PLoS Med, 11*(7), e1001673. doi: 10.1371/journal.pmed.1001673
PMEDICINE-D-13-04080 [pii]
Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., . . . Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science, 308*(5720), 385-389. doi: 1109557 [pii]
10.1126/science.1109557
Knoblauch, H., Busjahn, A., Munter, S., Nagy, Z., Faulhaber, H. D., Schuster, H., & Luft, F. C. (1997). Heritability analysis of lipids and three gene loci in twins link the macrophage scavenger receptor to HDL cholesterol concentrations. *Arterioscler Thromb Vasc Biol, 17*(10), 2054-2060.
Koh, H. J., Lee, S. M., Son, B. G., Lee, S. H., Ryoo, Z. Y., Chang, K. T., . . . Huh, T. L. (2004). Cytosolic NADP+-dependent isocitrate dehydrogenase plays a key role in lipid metabolism. *J Biol Chem, 279*(38), 39968-39974. doi: 10.1074/jbc.M402260200
M402260200 [pii]
Laclaustra, M., Stranges, S., Navas-Acien, A., Ordovas, J. M., & Guallar, E. (2010). Serum selenium and serum lipids in US adults: National Health and Nutrition Examination Survey (NHANES) 2003-2004. *Atherosclerosis, 210*(2), 643-648. doi: 10.1016/j.atherosclerosis.2010.01.005
S0021-9150(10)00008-0 [pii]
Lafontan, M., & Berlan, M. (2003). Do regional differences in adipocyte biology provide new pathophysiological insights? *Trends Pharmacol Sci, 24*(6), 276-283. doi: S0165-6147(03)00132-9 [pii]
10.1016/S0165-6147(03)00132-9
Lango Allen, H., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. N., Rivadeneira, F., . . . Hirschhorn, J. N. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature, 467*(7317), 832-838. doi: 10.1038/nature09410
nature09410 [pii]
Lapidus, L., Bengtsson, C., Larsson, B., Pennert, K., Rybo, E., & Sjostrom, L. (1984). Distribution of adipose tissue and risk of cardiovascular disease and death: a 12 year follow up of participants in the population study of women in Gothenburg, Sweden. *Br Med J (Clin Res Ed), 289*(6454), 1257-1261.
Larsson, B., Svardsudd, K., Welin, L., Wilhelmsen, L., Bjorntorp, P., & Tibblin, G. (1984). Abdominal adipose tissue distribution, obesity, and risk of cardiovascular disease and death: 13 year follow up of participants in the study of men born in 1913. *Br Med J (Clin Res Ed), 288*(6428), 1401-1404.

Lettre, G., Jackson, A. U., Gieger, C., Schumacher, F. R., Berndt, S. I., Sanna, S., . . . Hirschhorn, J. N. (2008). Identification of ten loci associated with height highlights new biological pathways in human growth. *Nat Genet, 40*(5), 584-591. doi: 10.1038/ng.125
ng.125 [pii]
Li, X., Allayee, H., Xiang, A. H., Trigo, E., Hartiala, J., Lawrence, J. M., . . . Watanabe, R. M. (2009). Variation in IGF2BP2 interacts with adiposity to alter insulin sensitivity in Mexican Americans. *Obesity (Silver Spring), 17*(4), 729-736. doi: 10.1038/oby.2008.593
oby2008593 [pii]
Libby, P. (2002). Inflammation in atherosclerosis. *Nature, 420*(6917), 868-874. doi: 10.1038/nature01323
nature01323 [pii]
Linehan, W. M., Srinivasan, R., & Schmidt, L. S. (2010). The genetic basis of kidney cancer: a metabolic disease. *Nat Rev Urol, 7*(5), 277-285. doi: 10.1038/nrurol.2010.47
nrurol.2010.47 [pii]
Lisinski, I., Matsumoto, H., Yver, D. R., Schurmann, A., Cushman, S. W., & Al-Hasani, H. (2006). Identification and characterization of p49/STRAP as a novel GLUT4-binding protein. *Biochem Biophys Res Commun, 344*(4), 1179-1185. doi: S0006-291X(06)00808-4 [pii]
10.1016/j.bbrc.2006.04.017
Liu, C. T., Monda, K. L., Taylor, K. C., Lange, L., Demerath, E. W., Palmas, W., . . . Fox, C. S. (2013). Genome-wide association of body fat distribution in African ancestry populations suggests new loci. *PLoS Genet, 9*(8), e1003681. doi: 10.1371/journal.pgen.1003681
PGENETICS-D-12-02200 [pii]
Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., . . . Speliotes, E. K. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature, 518*(7538), 197-206. doi: 10.1038/nature14177
nature14177 [pii]
Malhotra, D., & Sebat, J. (2012). CNVs: harbingers of a rare variant revolution in psychiatric genetics. *Cell, 148*(6), 1223-1241. doi: 10.1016/j.cell.2012.02.039
S0092-8674(12)00277-2 [pii]
Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., . . . Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature, 461*(7265), 747-753. doi: 10.1038/nature08494
nature08494 [pii]
Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nat Rev Genet, 11*(7), 499-511. doi: 10.1038/nrg2796
nrg2796 [pii]
Mihara, H., Kurihara, T., Watanabe, T., Yoshimura, T., & Esaki, N. (2000). cDNA cloning, purification, and characterization of mouse liver selenocysteine lyase. Candidate for selenium delivery protein in selenoprotein synthesis. *J Biol Chem, 275*(9), 6195-6200.
Miller, M., Stone, N. J., Ballantyne, C., Bittner, V., Criqui, M. H., Ginsberg, H. N., . . . Pennathur, S. (2011). Triglycerides and cardiovascular disease: a scientific statement from the American Heart Association. *Circulation, 123*(20), 2292-2333. doi: 10.1161/CIR.0b013e3182160726
CIR.0b013e3182160726 [pii]

Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C., . . . Korbel, J. O. (2011). Mapping copy number variation by population-scale genome sequencing. *Nature, 470*(7332), 59-65. doi: 10.1038/nature09708
nature09708 [pii]
Monda, K. L., Chen, G. K., Taylor, K. C., Palmer, C., Edwards, T. L., Lange, L. A., . . . Haiman, C. A. (2013). A meta-analysis identifies new loci associated with body mass index in individuals of African ancestry. *Nat Genet, 45*(6), 690-696. doi: 10.1038/ng.2608
ng.2608 [pii]
Montague, C. T., & O'Rahilly, S. (2000). The perils of portliness: causes and consequences of visceral adiposity. *Diabetes, 49*(6), 883-888.
Moore, K. J., & Tabas, I. (2011). Macrophages in the pathogenesis of atherosclerosis. *Cell, 145*(3), 341-355. doi: 10.1016/j.cell.2011.04.005
S0092-8674(11)00422-3 [pii]
Mozaffarian, D., Benjamin, E. J., Go, A. S., Arnett, D. K., Blaha, M. J., Cushman, M., . . . Turner, M. B. (2016). Heart Disease and Stroke Statistics-2016 Update: A Report From the American Heart Association. *Circulation, 133*(4), e38-60. doi: 10.1161/CIR.0000000000000350
CIR.0000000000000350 [pii]
Mullaney, J. M., Mills, R. E., Pittard, W. S., & Devine, S. E. (2010). Small insertions and deletions (INDELs) in human genomes. *Human Molecular Genetics, 19*, R131-R136. doi: 10.1093/hmg/ddq400
Nam, W. S., Park, K. M., & Park, J. W. (2012). RNA interference targeting cytosolic NADP(+)-dependent isocitrate dehydrogenase exerts anti-obesity effect in vitro and in vivo. *Biochim Biophys Acta, 1822*(8), 1181-1188. doi: 10.1016/j.bbadis.2012.04.003
S0925-4439(12)00090-7 [pii]
Narkiewicz, K. (2005). Obesity and hypertension--the issue is more complex than we thought. *Nephrology Dialysis Transplantation, 21*(2), 264-267. doi: 10.1093/ndt/gfi290
Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., . . . Bamshad, M. J. (2010). Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet, 42*(1), 30-35. doi: 10.1038/ng.499
ng.499 [pii]
Nogueira, T. C., Paula, F. M., Villate, O., Colli, M. L., Moura, R. F., Cunha, D. A., . . . Eizirik, D. L. (2013). GLIS3, a susceptibility gene for type 1 and type 2 diabetes, modulates pancreatic beta cell apoptosis via regulation of a splice variant of the BH3-only protein Bim. *PLoS Genet, 9*(5), e1003532. doi: 10.1371/journal.pgen.1003532
PGENETICS-D-12-02709 [pii]
Norris, J. M., Langefeld, C. D., Scherzinger, A. L., Rich, S. S., Bookman, E., Beck, S. R., . . . Wagenknecht, L. E. (2005). Quantitative trait loci for abdominal fat and BMI in Hispanic-Americans and African-Americans: the IRAS Family study. *Int J Obes (Lond), 29*(1), 67-77. doi: 0802793 [pii]
10.1038/sj.ijo.0802793
O'Connell, J. R., & Weeks, D. E. (1998). PedCheck: a program for identification of genotype incompatibilities in linkage analysis. *Am J Hum Genet, 63*(1), 259-266.
Ogden, C. L., Carroll, M. D., Kit, B. K., & Flegal, K. M. (2014). Prevalence of childhood and adult obesity in the United States, 2011-2012. *JAMA, 311*(8), 806-814. doi: 10.1001/jama.2014.732
1832542 [pii]

Overgaard, M. T., Boldt, H. B., Laursen, L. S., Sottrup-Jensen, L., Conover, C. A., & Oxvig, C. (2001). Pregnancy-associated plasma protein-A2 (PAPP-A2), a novel insulin-like growth factor-binding protein-5 proteinase. *J Biol Chem, 276*(24), 21849-21853. doi: 10.1074/jbc.M102191200
M102191200 [pii]
Pacini, G., & Bergman, R. N. (1986). MINMOD: a computer program to calculate insulin sensitivity and pancreatic responsivity from the frequently sampled intravenous glucose tolerance test. *Comput Methods Programs Biomed, 23*(2), 113-122.
Palmer, N. D., Goodarzi, M. O., Langefeld, C. D., Wang, N., Guo, X., Taylor, K. D., . . . Wagenknecht, L. E. (2015). Genetic Variants Associated With Quantitative Glucose Homeostasis Traits Translate to Type 2 Diabetes in Mexican Americans: The GUARDIAN (Genetics Underlying Diabetes in Hispanics) Consortium. *Diabetes, 64*(5), 1853-1866. doi: 10.2337/db14-0732
db14-0732 [pii]
Perry, G. H., Dominy, N. J., Claw, K. G., Lee, A. S., Fiegler, H., Redon, R., . . . Stone, A. C. (2007). Diet and the evolution of human amylase gene copy number variation. *Nat Genet, 39*(10), 1256-1260. doi: ng2123 [pii]
10.1038/ng2123
Pinto, D., Delaby, E., Merico, D., Barbosa, M., Merikangas, A., Klei, L., . . . Scherer, S. W. (2014). Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *Am J Hum Genet, 94*(5), 677-694. doi: 10.1016/j.ajhg.2014.03.018
S0002-9297(14)00150-5 [pii]
Randall, J. C., Winkler, T. W., Kutalik, Z., Berndt, S. I., Jackson, A. U., Monda, K. L., . . . Heid, I. M. (2013). Sex-stratified Genome-wide Association Studies Including 270,000 Individuals Show Sexual Dimorphism in Genetic Loci for Anthropometric Traits. *PLoS Genet, 9*(6), e1003500. doi: 10.1371/journal.pgen.1003500
PGENETICS-D-12-01098 [pii]
Rees, S. D., Hydrie, M. Z., O'Hare, J. P., Kumar, S., Shera, A. S., Basit, A., . . . Kelly, M. A. (2011). Effects of 16 genetic variants on fasting glucose and type 2 diabetes in South Asians: ADCY5 and GLIS3 variants may predispose to type 2 diabetes. *PLoS One, 6*(9), e24710. doi: 10.1371/journal.pone.0024710
PONE-D-11-07898 [pii]
Reitman, Z. J., & Yan, H. (2010). Isocitrate dehydrogenase 1 and 2 mutations in cancer: alterations at a crossroads of cellular metabolism. *J Natl Cancer Inst, 102*(13), 932-941. doi: 10.1093/jnci/djq187
djq187 [pii]
Salakhov, R. R., Goncharova, I. A., Makeeva, O. A., Golubenko, M. V., Kulish, E. V., Kashtalap, V. V., . . . Puzyrev, V. P. (2014). [TOMM40 gene polymorphism association with lipid profile]. *Genetika, 50*(2), 222-229.
Scheet, P., & Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet, 78*(4), 629-644. doi: S0002-9297(07)63701-X [pii]
10.1086/502802
Scully, T. (2014). Public health: Society at large. *Nature, 508*(7496), S50-51. doi: 10.1038/508S50a
508S50a [pii]
Senee, V., Chelala, C., Duchatelet, S., Feng, D., Blanc, H., Cossec, J. C., . . . Julier, C. (2006). Mutations in GLIS3 are responsible for a rare syndrome with neonatal

diabetes mellitus and congenital hypothyroidism. *Nat Genet, 38*(6), 682-687. doi: ng1802 [pii]
10.1038/ng1802
Shu, Y. H., Hartiala, J., Xiang, A. H., Trigo, E., Lawrence, J. M., Allayee, H., . . . Watanabe, R. M. (2009). Evidence for sex-specific associations between variation in acid phosphatase locus 1 (ACP1) and insulin sensitivity in Mexican-Americans. *J Clin Endocrinol Metab, 94*(10), 4094-4102. doi: 10.1210/jc.2008-2751
jc.2008-2751 [pii]
Shungin, D., Winkler, T. W., Croteau-Chonka, D. C., Ferreira, T., Locke, A. E., Magi, R., . . . Mohlke, K. L. (2015). New genetic loci link adipose and insulin biology to body fat distribution. *Nature, 518*(7538), 187-196. doi: 10.1038/nature14132
nature14132 [pii]
Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., . . . Loos, R. J. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet, 42*(11), 937-948. doi: 10.1038/ng.686
ng.686 [pii]
Srivastava, A. K. (2012). Challenges in the treatment of cardiometabolic syndrome. *Indian Journal of Pharmacology, 44*(2), 155-156. doi: 10.4103/0253-7613.93579
Sudmant, P. H., Kitzman, J. O., Antonacci, F., Alkan, C., Malig, M., Tsalenko, A., . . . Eichler, E. E. (2010). Diversity of human copy number variation and multicopy genes. *Science, 330*(6004), 641-646. doi: 10.1126/science.1197005
330/6004/641 [pii]
Surakka, I., Horikoshi, M., Magi, R., Sarin, A. P., Mahajan, A., Lagou, V., . . . Ripatti, S. (2015). The impact of low-frequency and rare variants on lipid levels. *Nat Genet, 47*(6), 589-597. doi: 10.1038/ng.3300
ng.3300 [pii]
Tabb, K. L., Hellwege, J. N., Palmer, N. D., Dimitrov, L., Sajuthi, S., Taylor, K. D., . . . Bowden, D. W. (2016). Analysis of whole exome sequencing with cardiometabolic traits using family-based linkage and association in the IRAS Family Study. *Annals of Human Genetics*. doi: 10.1111/ahg.12184
Tan, A., Sun, J., Xia, N., Qin, X., Hu, Y., Zhang, S., . . . Xu, J. (2012). A genome-wide association and gene-environment interaction study for serum triglycerides levels in a healthy Chinese male population. *Hum Mol Genet, 21*(7), 1658-1664. doi: 10.1093/hmg/ddr587
ddr587 [pii]
Taneera, J., Fadista, J., Ahlqvist, E., Atac, D., Ottosson-Laakso, E., Wollheim, C. B., & Groop, L. (2015). Identification of novel genes for glucose metabolism based upon expression pattern in human islets and effect on insulin secretion and glycemia. *Hum Mol Genet, 24*(7), 1945-1955. doi: 10.1093/hmg/ddu610
ddu610 [pii]
Tatusova, T., Ciufo, S., Fedorov, B., O'Neill, K., & Tolstoy, I. (2015). RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res, 43*(7), 3872. doi: 10.1093/nar/gkv278
gkv278 [pii]
Taylor, R. W., Jones, I. E., Williams, S. M., & Goulding, A. (2000). Evaluation of waist circumference, waist-to-hip ratio, and the conicity index as screening tools for high trunk fat mass, as measured by dual-energy X-ray absorptiometry, in children aged 3-19 y. *Am J Clin Nutr, 72*(2), 490-495.

Teasdale, R. D., Loci, D., Houghton, F., Karlsson, L., & Gleeson, P. A. (2001). A large family of endosome-localized proteins related to sorting nexin 1. *Biochem J, 358*(Pt 1), 7-16.

Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., . . . Kathiresan, S. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature, 466*(7307), 707-713. doi: 10.1038/nature09270
nature09270 [pii]

Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. (2002). *Circulation, 106*(25), 3143-3421.

Treangen, T. J., & Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet, 13*(1), 36-46. doi: 10.1038/nrg3117
nrg3117 [pii]

Van Oosterhout, C., Hutchinson, W. F., Wills, D. P. M., & Shipley, P. (2004). MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. *Molecular Ecology Notes, 4*(3), 535-538. doi: 10.1111/j.1471-8286.2004.00684.x

Wagenknecht, L. E., Langefeld, C. D., Scherzinger, A. L., Norris, J. M., Haffner, S. M., Saad, M. F., & Bergman, R. N. (2003). Insulin sensitivity, insulin secretion, and abdominal fat: the Insulin Resistance Atherosclerosis Study (IRAS) Family Study. *Diabetes, 52*(10), 2490-2496.

Wagenknecht, L. E., Mayer, E. J., Rewers, M., Haffner, S., Selby, J., Borok, G. M., . . . et al. (1995). The insulin resistance atherosclerosis study (IRAS) objectives, design, and recruitment results. *Ann Epidemiol, 5*(6), 464-472. doi: 1047279795000623 [pii]

Wajchenberg, B. L. (2000). Subcutaneous and visceral adipose tissue: their relation to the metabolic syndrome. *Endocr Rev, 21*(6), 697-738.

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res, 38*(16), e164. doi: 10.1093/nar/gkq603
gkq603 [pii]

Wang, Y.-P., Kandeel, F., Taylor, K. D., Hernandez, D., Saad, M. F., Nadler, J. L., & Raffel, L. J. (2000). Insulin and blood pressure are linked to the LDL receptor-related protein locus on chromosome 12q (Abstract). *Diabetes, 49(Supp 1)*, A204.

Watanabe, R. M., Allayee, H., Xiang, A. H., Trigo, E., Hartiala, J., Lawrence, J. M., & Buchanan, T. A. (2007). Transcription factor 7-like 2 (TCF7L2) is associated with gestational diabetes mellitus and interacts with adiposity to alter insulin secretion in Mexican Americans. *Diabetes, 56*(5), 1481-1485. doi: db06-1682 [pii]
10.2337/db06-1682

Weischenfeldt, J., Symmons, O., Spitz, F., & Korbel, J. O. (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat Rev Genet, 14*(2), 125-138. doi: 10.1038/nrg3373
nrg3373 [pii]

Willer, C. J., Sanna, S., Jackson, A. U., Scuteri, A., Bonnycastle, L. L., Clarke, R., . . . Abecasis, G. R. (2008). Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet, 40*(2), 161-169.

Willer, C. J., Schmidt, E. M., Sengupta, S., Peloso, G. M., Gustafsson, S., Kanoni, S., . . . Abecasis, G. R. (2013). Discovery and refinement of loci associated with lipid levels. *Nat Genet, 45*(11), 1274-1283. doi: 10.1038/ng.2797
ng.2797 [pii]

Willyard, C. (2014). Heritability: The family roots of obesity. *Nature, 508*(7496), S58-60. doi: 10.1038/508S58a
508S58a [pii]
Wiltshire, S., Morris, A. P., McCarthy, M. I., & Cardon, L. R. (2005). How useful is the fine-scale mapping of complex trait linkage peaks? Evaluating the impact of additional microsatellite genotyping on the posterior probability of linkage. *Genet Epidemiol, 28*(1), 1-10. doi: 10.1002/gepi.20023
Wing, M. R., Ziegler, J. M., Langefeld, C. D., Roh, B. H., Palmer, N. D., Mayer-Davis, E. J., . . . Bowden, D. W. (2011). Analysis of FTO gene variants with obesity and glucose homeostasis measures in the multiethnic Insulin Resistance Atherosclerosis Study cohort. *Int J Obes (Lond), 35*(9), 1173-1182. doi: 10.1038/ijo.2010.244
ijo2010244 [pii]
Wolf, A. M., & Colditz, G. A. (1998). Current estimates of the economic cost of obesity in the United States. *Obesity Research, 6*(2), 97-106.
Xiang, A. H., Azen, S. P., Raffel, L. J., Tan, S., Cheng, L. S., Diaz, J., . . . Buchanan, T. A. (2001). Evidence for joint genetic control of insulin sensitivity and systolic blood pressure in hispanic families with a hypertensive proband. *Circulation, 103*(1), 78-83.
Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., . . . Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat Genet, 42*(7), 565-569. doi: 10.1038/ng.608
ng.608 [pii]
Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet, 88*(1), 76-82. doi: 10.1016/j.ajhg.2010.11.011
S0002-9297(10)00598-7 [pii]
Yang, J., Loos, R. J., Powell, J. E., Medland, S. E., Speliotes, E. K., Chasman, D. I., . . . Visscher, P. M. (2012). FTO genotype is associated with phenotypic variability of body mass index. *Nature, 490*(7419), 267-272. doi: 10.1038/nature11401
nature11401 [pii]
Zheng, B., Ma, Y. C., Ostrom, R. S., Lavoie, C., Gill, G. N., Insel, P. A., . . . Farquhar, M. G. (2001). RGS-PX1, a GAP for GalphaS and sorting nexin in vesicular trafficking. *Science, 294*(5548), 1939-1942. doi: 10.1126/science.1064757
294/5548/1939 [pii]
Zhou, L., He, M., Mo, Z., Wu, C., Yang, H., Yu, D., . . . Wu, T. (2013). A genome wide association study identifies common variants associated with lipid levels in the Chinese population. *PLoS One, 8*(12), e82420. doi: 10.1371/journal.pone.0082420
PONE-D-12-11699 [pii]
Zhou, X., Marks, P. A., Rifkind, R. A., & Richon, V. M. (2001). Cloning and characterization of a histone deacetylase, HDAC9. *Proc Natl Acad Sci U S A, 98*(19), 10572-10577. doi: 10.1073/pnas.191375098
191375098 [pii]
Zillikens, M. C., Yazdanpanah, M., Pardo, L. M., Rivadeneira, F., Aulchenko, Y. S., Oostra, B. A., . . . van Duijn, C. M. (2008). Sex-specific genetic effects influence variation in body composition. *Diabetologia, 51*(12), 2233-2241. doi: 10.1007/s00125-008-1163-0

CURRICULUM VITAE

EDUCATION:

Ph.D. Candidate Molecular Genetics and Genomics

2012-2017 (anticipated): Wake Forest School of Medicine, Winston-Salem, NC

M.B.A. Business Administration

2010-2012: Winthrop University, Rock Hill, SC

B.S. Pharmaceutical Sciences

2006-2010: School of Pharmacy, Shandong University of Traditional Chinese Medicine, China

TRAININGS AND SHORT COURSES:

- Short Course on Nutrigenetics, Nutrigenomics and Precision Nutrition, University of North Carolina, NC (2016)
- Clinical Informatics Short Course, Wake Forest School of Medicine, NC (2016)
- Short Course on Statistical Genetics and Genomics, University of Alabama at Birmingham, AL (2014)

TECHNIQUES:

- **Statistical Genetics**: multiple linear regression models, mixed effects models, GWAS, gene burden test, family linkage analysis, next-generation sequencing data analysis, family heritability estimation, gene-gene/environment interaction analysis.
    - Software: SOLAR, PLINK, IMPUTE2, METAL, GWAMA, GCTA, SKAT, GWAF, etc.
- **Computer Skills**: data organization (shell script) and statistical analysis (SAS, R), high-performance computing cluster.
- **Molecular and Cellular Biology**: DNA/RNA isolation, PCR, western blot, molecular cloning, cell culture, flow cytometry, plasma and tissue lipids analysis, atherosclerotic lesion analysis.
- **Other:** image analysis with ImageJ (Java script) and fundamentals of Intellectual Property (prior patent search and business case analysis).

EXPERIENCE:

- **Feb 2013–Present, Ph.D. Dissertation Research**

  Dr. Nicholette Allred, Biochemistry, Wake Forest School of Medicine

  Dissertation: Investigation of the genetic architecture of cardiometabolic disease using statistical genetics approaches.

    - Conducted data cleaning and quality control for population genetics and clinical records

- Identified, modified, and implemented statistical packages
- Integrated common genetics databases, resources and tools
- Performed association and linkage for complex disease phenotypes in family cohorts

- **March 2016–May 2016, Intern**

  Licensing Office, Wake Forest Innovations, Winston-Salem, NC

  - Prior art analysis (patent search)
  - Licensing term sheet drafting
  - Business and market analysis

- **June 2015–August 2015, Summer Intern**

Verrix LLC, Pasadena, CA

  - Built OMERO image database on AMAZON cloud
  - Developed ImageJ script for fluorescent image signal quantification
  - Investigated bacteria spore germination conditions using *subtilis* and *geobacillus*

- **November 2012–February 2013.2, Rotation Student**

  Dr. Richard Loeser, Molecular Medicine, Wake Forest School of Medicine

  - Studied inflammatory signaling pathways of primary human chondrocytes triggered by fibronectin fragments of different domains.

- **August 2012–November 2012, Rotation Student**

  Dr. John Parks, Molecular Medicine, Wake Forest School of Medicine

  - Established methods for quantification of atherosclerotic lesion immune cells and explored mechanisms underlying atherogenesis using a mouse model.

- **October 2011–July 2012, Research Assistant**

  Dr. Eric Birgbauer, Biology Department, Winthrop University

  - Constructed siRNA containing retrovirus vector against chicken LPA receptors

AWARDS:

- Alumni Student Travel Award, Wake Forest University, NC (2016)
- Travel Award, Short Course on Nutrigenetics, Nutrigenomics and Precision Nutrition, University of North Carolina, NC (2016)
- Scholarship, Clinical Informatics Short Course, Wake Forest School of Medicine, NC (2016)
- Travel Award, Short Course on Statistical Genetics and Genomics, University of Alabama at Birmingham, AL (2014)
- Graduate Fellowship, Wake Forest School of Medicine, NC (2012-2013)

ABSTRACTS:

Gao C, Keaton J, Hsu F, Okut H, Rotter J, Chen Y, Taylor K, Langefeld C, Wagenknecht L, Bowden D, Palmer N. Genome-wide gene-environment interaction analysis identifies genetic signals associated with cardiometabolic phenotypes and physical activity in Hispanic Americans. *American Society of Human Genetics annual meeting, 2016. Vancouver, Canada*

Gao C, Keaton J, Hsu F, Okut H, Rotter J, Chen Y, Taylor K, Langefeld C, Wagenknecht L, Bowden D, Palmer N. Genome-wide gene-environment interaction analysis identifies genetic signals associated with cardiometabolic phenotypes and physical activity in Hispanic Americans. *Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Investigator Meeting, 2016. Charlottesville, VA*

Tabb K, **Gao C**, Hawkins G, Rotter J, Chen Y, Norris J, Lorenzo C, Freedman B, Bowden D, Palmer N. Adiponectin isoform distribution in carriers of ethnic-specific ADIPOQ mutations: the insulin resistance atherosclerosis family study (IRASFS). *Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Investigator Meeting*, *2016. Charlottesville, VA*

Keaton J, **Gao C**, Guan M, Hellwege J, Palmer N, Pankow J, Fornage M, Wilson J, Correa A, Rasmussen-Torvik L, Rotter J, Chen Y, Taylor K, Rich S, Wagenknecht L, Freedman B, Ng M, Bowden D. Genome-wide interaction with fasting glucose risk score reveals novel type 2 diabetes susceptibility genes in African Americans. *Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Investigator Meeting*, *2016. Charlottesville, VA*

Gao C, Ziegler J, Taylor K, Norris J, Chen Y, Hellwege J, Speliotes E, Rotter J, Watanabe R, Wagenknecht L, Langefeld C, Palmer N. Genome-wide and exome chip study of subcutaneous and visceral adipose tissue reveals novel gender-specific adiposity loci in Hispanic Americans: The Insulin Resistance Atherosclerosis Family Study. *American Society of Human Genetics annual meeting, 2014. San Diego, CA*

Palmer N, **Gao C**, Norris J, Rotter J, Wagenknecht L, Bowden D, Speliotes E, Langefeld C. Exome Chip-based Association Analysis Identifies Novel Coding Variants Associated with Adiposity Traits in Hispanic Americans: The IRAS Family Stud. *American Society of Human Genetics annual meeting, 2013. Boston, MA*

Bi X, Zhu X, **Gao C**, Cao Q, Liu M, Shewale S, Boudyguina E, Wilson M, Parks J. Myeloid Cell Specific ABCA1 Deletion Does Not Significantly Accelerate Atherogenesis in LDL Receptor Knockout Mice. *Arteriosclerosis, Thrombosis and Vascular Biology Scientific Sessions, 2013. Lake Buena Vista, FL*

PUBLICATIONS:

Gao C, Tabb K, Dimitrov L, Taylor K, Wang N, Guo X, Long J, Rotter J, Watanabe R, Langefeld C, Bowden D, Palmer N. Exome Sequencing Identifies APOA5 Variants Associated with Circulating Lipid Levels in Mexican Americans: The Insulin Resistance Atherosclerosis Family Study (IRASFS). *[manuscript in preparation]*

Gao C, Dimitrov L, Ziegler J, Taylor K, Norris J, Chen Y, Hellwege J, Speliotes E, Rotter J, Watanabe R, Wagenknecht L, Langefeld C, Palmer N. Genome-wide and exome chip study of subcutaneous and visceral adipose tissue reveals novel gender-specific adiposity loci in Mexican-origin Americans: The Insulin Resistance Atherosclerosis Family Study (IRASFS)*. [manuscript in preparation]*

Tabb K, **Gao C**, Hicks P, Hawkins G, Rotter J, Chen Y, Norris J, Lorenzo C, Freedman B, Bowden D, Palmer N. Adiponectin isoform patterns in ethnic-specific ADIPOQ mutation carriers: The Insulin Resistance Atherosclerosis Family Study. *[manuscript submitted to Journal of Clinical Endocrinology & Metabolism]*

Gao C, Hsu F, Dimitrov L, Okut H, Chen Y, Taylor K, Rotter J, Langefeld C, Bowden D, Palmer N. A Genome-wide Linkage and Association Analysis of Imputed Insertions and Deletions with Cardiometabolic Phenotypes in Mexican Americans: The Insulin Resistance Atherosclerosis Family Study. *[manuscript accepted by Genetic Epidemiology]*

Ma J, Guan M, Bowden D, Ng C.Y., Hicks P, Lea J, Ma L, **Gao C**, Palmer N, Freedman B. Association Analysis Of The Cubilin (CUBN) And Megalin (LRP2) Genes With End-Stage Kidney Disease In African Americans. *Clin J Am Soc Nephrol*. *2016 Jun 6;11(6):1034-43.*

Gao C, Wang N, Guo X, Ziegler J, Taylor K, Xiang A, Raffel L, Chen Y, Norris J, Rotter J, Watanabe R, Wagenknecht L, Bowden D, Speliotes E, Goodarzi M, Langefeld C, Palmer N. A Comprehensive Analysis of Common and Rare Variants to Identify Novel Adiposity Loci in Hispanic Americans: The Insulin Resistance Atherosclerosis Family Study (IRASFS). *PLoS One. 2015 Nov 24;10(11):e0134649.*

Bi X, Zhu X, **Gao C**, Shewale S, Cao Q, Liu M, Boudyguina E, Gebre A, Wilson M, Brown A, and Parks JS. Myeloid Cell-Specific ABCA1 Deletion Has Minimal Impact on Atherogenesis in Atherogenic Diet-Fed LDL Receptor Knockout Mice. *Arteriosclerosis, Thrombosis and Vascular Biology. 2014 Sep;34(9):1888-99.*