The genetics of obesity-related traits and lipoproteins in Filipino women

Amanda Faith Marvelle

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Curriculum of Genetics and Molecular Biology.

Chapel Hill
2010

Approved by:

Advisor: Karen L. Mohlke, PhD

Reader: Linda S. Adair, PhD

Reader: Leslie A. Lange, PhD

Reader: Daniel Pomp, PhD

Reader: Patrick F. Sullivan, MD, FRANZCP

ABSTRACT

Amanda Faith Marvelle: The genetics of obesity-related traits and lipoproteins in Filipino women
(Under the direction of Dr. Karen L. Mohlke)

The underlying genetic component of risk factors for cardiovascular disease (CVD) is not well understood. Recently, advances in high-throughput genotyping, single nucleotide polymorphism (SNP) discovery, and the development of databases such as the International Haplotype Map (HapMap) have provided scientists with tools to complete a genetic analysis of complex diseases such as CVD. Research presented in this dissertation aims to further understand the genetics of obesity-related traits and lipoprotein levels and identify variants that are associated with these traits in a cohort of adult women from metro Cebu, Philippines, who participated in the Cebu Longitudinal Health and Nutrition Survey (CLHNS). Initially I assess the transferability of tag SNPs chosen from HapMap panels to the CLHNS. I show that the Asian HapMap samples are an effective resource for studies in the CLHNS. I then investigate the association between 19 candidate variants in 10 genes previously reported to be associated with obesity-related traits with similar traits in the CLHNS. We observe evidence for association with the A-allele of rs9939609 of *FTO* and *ADRB3* Trp64-allele with obesity traits. I perform a genome-wide association study for HDL-C, triglycerides, LDL-C, and total cholesterol. Among ~2 million SNPs analyzed, we observe evidence of association for 11 loci previously described. We observe suggestive evidence of trait association ($P < 10^{-5}$) for Tankyrase (*TNKS*) with LDL-C and Collecting-12 (*COLEC12*) with total

cholesterol. In a separate study, I investigate an HDL-C associated locus, *GALNT2*, to identify functional variants responsible for the association signal. I identify variants in moderate linkage disequilibrium (r^2 >.5) with HDL-C associated SNPs, clone regions that have suggestive regulatory function into a luciferase reporter vector, and measure transcriptional activity in HepG2 cells. The results suggest that a 21 bp deletion, rs4849913, and/or rs2144300 may act to increase the transcriptional activity of *GALNT2* or an unknown novel intronic transcript to increase HDL-C. These studies present the first genetic study of CVD traits in the CLHNS and a molecular study of a gene that is associated with HDL-C. Together this research provides a solid foundation for one day identifying the molecular mechanism underlying complex diseases.

To my loving husband and daughter.

ACKNOWLEDGEMENTS
(Specific credits are listed before each chapter)

In completing this dissertation and the graduate school process, I am indebted to all the people who have given me guidance and support.

First and foremost, I would like to gratefully and sincerely thank my wonderful advisor, Karen Mohlke for her supervision, understanding, and most importantly, patience during my graduate studies. I cannot imagine being in a better lab with a better mentor. No matter what, she was always looking out for my best interests. She believed in me and pushed me to work to my potential. Her mentorship was paramount in providing a well-rounded experience during graduate school. Karen has encouraged me to not only grow as an experimentalist and a geneticist, but also as an independent thinker. For everything she has done, I thank her.

I would like to thank members of my dissertation committee: Drs. Leslie Lange, Linda Adair, Patrick Sullivan, and Daniel Pomp. Their positive support, accessibility, and research advice was essential to completing this thesis. In particular, I would like to thank Linda Adair and Leslie Lange for their hard work, expertise and patience on guiding me through my projects. Linda helped me with learning all about the CLHNS and Leslie taught me statistics and lent a sympathetic ear when I needed to get the stresses of life off my chest. Leslie was not only a committee member but also a great friend.

Of course, I would like to thank all the members of the Mohlke lab. There are two

members of the lab, Li and Kyle, that have been there since I started graduate school. Li Qin, our lab manager, has been so much more than someone who keeps everything organized and running smoothly, she has been a friend and confidant and someone I turn to when I want to talk about anything. Kyle Gaulton, a fellow graduate student in the genetics curriculum joined Karen's lab at the same time as I did, and we were Karen's first graduate students. Over the past five years I have sat no further than a foot and a half from Kyle, which has allowed us to become very close both figuratively and literally. We have had many long afternoon chats about life, and he has given me some great advice on my projects. I am very grateful for his friendship. In addition, I would like to thank Damien Croteau-Chonka for his input on Aim 3 which would have been impossible without him; Arlene Gonzalez Martinez, a summer undergraduate student who was a tremendous help during the experimental work of Aim 4; as well as the other lab members both past and present who have supported me: Marie Fogarty, Tami Panhuis, Ying Wu, Ghenadie Curocichin, Martin Buchkovich, Jennifer Kulzer, and Tamara Roman.

I would also like to thank my family and friends because without them I would not be the person I am today. I thank my parents, Deborah and David Nave, for their faith in me and for allowing me to be as ambitious as I wanted. It was under their watchful eye that I gained so much drive and an ability to tackle challenges head on. They instilled in me a can-do spirit and without them I would not have never thought of going to graduate school. I would also like to thank my sister, Michelle, and nephew, Jesse, for their support. I would like to thank my daughter, Elizabeth Cora, for a very easy pregnancy while I was trying to finish lab work and write this dissertation and for

waiting to be born until after I defended! She is the most beautiful baby, and I love her dearly. I would like to thank my closest friends Michelle and Mark Leslie. They were the first people I met when I moved to Chapel Hill and have become some of the most important people in my life. We have shared in many good and bad (not so many of these, thankfully) experiences and I can't imagine the last six years without them.

Last but certainly not least, I thank my husband, Nathan, for all he has done for me. He has been by my side from the applications to the completion of graduate school. Before we were married he was willing to move with me wherever I decided to go to school (with a few exceptions), which really showed how much he loved me. He was my motivator, editor, and confidant. He even helped label tubes and keep me company until 2am during my last experiment. His support, encouragement, quiet patience and unwavering love were undeniably the bedrock upon which the past eight years of my life have been built.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AFA	arm fat area
AMA	arm muscle area
ANCOVA	analysis of covariance
BMI	body mass index
CAD	coronary artery disease
CEPH	Centre d'Etude du Polymorphisme Humain
CEU	Caucasians from Utah, USA with northern and western European ancestry
CHB	Han Chinese from Beijing, China
ChIP	chromatin immunoprecipitation
CLHNS	Cebu Longitudinal Health and Nutrition Survey
CVD	cardiovascular diseases
DHS	DNase I hypersensitivity
ENCODE	ENCyclopedia Of DNA Elements
FAIRE	formaldehyde-assisted identification of regulatory elements
FUSION	Finland-United States Investigation of NIDDM Genetics
GWA	genome-wide association
HapMap	The International Haplotype Map
HDL-C	high density lipoprotein cholesterol
IBD	identity-by-descent
IBD	identity-by-state
JPT	Japanese from Tokyo, Japan
kb	kilobase

LD	linkage disequilibrium
LDL-C	low density lipoprotein cholesterol
MAF	minor allele frequency
Mb	megabase
nt	nucleotide
OPS	Office of Population Studies
PC	principal components
Q-Q	quantile-quantile
SiSF	suprailiac skinfold thickness
SNP	single nucleotide polymorphism
TSF	triceps skinfold thickness
YRI	Yoruban of Ibadan, Nigeria

CHAPTER I. INTRODUCTION

Obesity and unhealthy lipid profiles, along with insulin resistance and high blood pressure, increase risk for cardiovascular diseases (CVD) [1]. CVDs are common, complex traits with a substantial public health burden. It is estimated that CVDs are the leading cause of morbidity and mortality globally. Disease prevalence varies with age, gender, and population [2].

The genetics of obesity-related traits and lipoproteins are not clearly understood. Recently, advances in high throughput genotyping, single nucleotide polymorphism (SNP) discovery, and the development of databases such as the International Haplotype Map (HapMap) [3] have provided scientists with tools to complete a genetic analysis of complex diseases. This dissertation has taken advantage of these exciting technical advances to identify variants that are associated with traits influencing CVD, specifically, obesity and lipoprotein levels, in a cohort of women from the Philippines.

CARDIOVASCULAR DISEASES (CVDs)

Cardiovascular diseases are a group of disorders of the heart and blood vessels. The most common reason for these diseases is a build-up of fatty deposits on the walls of the blood vessels known as atherosclerosis [2]. Atherosclerotic cardiovascular diseases include coronary heart disease, peripheral vascular disease, and stroke. There are several risk factors for CVDs. Environmental risk factors include high fat diet, physical inactivity, alcohol and tobacco use. The effects of these may include increased blood pressure, blood glucose, blood lipid levels, and body weight [4]. Other determinants of CVDs are poverty and stress [2]. Additionally there is a strong genetic component to

CVDs; a recent study found that on average, genetic effects explain 25% of the variance for 20 measures of cardiovascular function [5].

By 2030, an estimated 23.6 million people worldwide will die from CVDs, with the largest increase in the number of deaths occurring in the southeast Asia region [6]. Increasingly, low and middle income countries are becoming disproportionally affected by CVDs, with ~80% of all CVD deaths taking place in these countries. The populations of these countries are more exposed to risk factors leading to CVDs and have less access to prevention efforts and effective health care services such as early detection [2].

OBESITY

Obesity is a worldwide epidemic. The World Health Organization estimates that over 1.6 billion adults worldwide are considered overweight and 400 million of these are considered obese. Once a problem in only high-income countries, obesity prevalence is now on the rise in low and middle income countries as well [7].

Obesity is a condition in which excess fat has accumulated in the body. Obesity is a risk factor for CVD, type 2 diabetes, metabolic syndrome, hypertension, stroke, musculoskeletal disorders, and some forms of cancer [8]. Overweight and obesity are usually evaluated using body mass index (BMI), defined as weight (in kilograms) divided by height (in meters) squared. For both genders and across populations, the World Health Organization defines overweight as BMI $\geq$ 25 kg/m^2 and obesity as BMI $\geq$ 30 [7]. These cutoffs provide benchmarks for the relative risk of mortality. However, in Asian subjects, a substantial proportion of individuals with BMI lower than 25 kg/m^2 have an increased risk of type 2 diabetes and CVD [9]. Therefore, for public health messages

lower cutoffs have been proposed for Asian populations (overweight $\geq$ 23 kg/m^2 and obese $\geq$ 25 kg/m^2), but not implemented widely. BMI is correlated with percent body fat (r = 0.7-0.8) [10]. Other obesity measures include waist circumference, a measure of central adiposity, and skinfold thinknesses, measures of subcutaneous adiposity.

Many factors influence obesity. The obesity epidemic may reflect the relatively new environmental obesogenic factors in society: overall reduced exercise caused by a shift from labor intensive jobs to more sedentary work; invention of automated systems and motorized vehicles; and increased accessibility and consumption of food with high calorie and fat content and low vitamins and nutrients. Age, poverty, increased number of pregnancies to a point, and menopausal status may influence a person's risk of obesity. There is also evidence of a strong genetic component to obesity. Obesity shows familial correlation and aggregation. Family, twin, and adoption studies have established heritability estimates for BMI ranging from 15 to 85 percent [8].

Worldwide, obesity is a tremendous public health problem, which needs to be addressed immediately. In order to solve this epidemic, it is necessary to understand the molecular mechanisms of obesity that will lead to the development medical strategies that affect energy balance to help prevent or treat obesity.

LIPOPROTEINS

Plasma lipoprotein levels are associated with risk of coronary artery disease (CAD). One of the underlying pathologies of CAD is atherosclerosis. Atherosclerosis takes place as low density lipoprotein cholesterol (LDL-C) deposits on the inner lining of arteries causing a thickening and hardening of the walls. Eventually the arteries cannot

supply blood to the heart or brain leading to myocardial infarction and/or stroke [11]. There has also been evidence supporting the association of lipoprotein concentration and CVD incidence [12,13]. Specifically, higher levels of LDL-C are associated with increased CVD with each 1% decrease resulting in a ~1% reduced risk of coronary heart disease [14]. Independently, clinical data shows that low levels of high-density lipoprotein cholesterol (HDL-C) are associated with increased CAD. Every 1% increase in HDL-C can decrease cardiovascular risk by ~2-3% [15]. Additionally, there is evidence that higher levels of triglycerides can lead to an increased risk of CVD [16].

While diet, exercise, weight, socioeconomic status, age, alcohol consumption, and smoking affect levels of lipoproteins, heritability estimates for lipoprotein concentrations are as high as 75% [17]. However, the influence of particular genetic variants on lipoprotein levels may differ across populations. For these reasons, there has been considerable interest in understanding these genetic factors that contribute to inter-individual variation.

GENETIC FACTORS IN CARDIOVASCULAR DISEASE

The identification of genes that contribute to the etiology of CVD and other complex diseases and traits has evolved remarkably over the last decade [18]. For instance, the identification of variants contributing to the inter-individual variation of lipoprotein levels initially proved difficult. Mutations in genes such as *LDLR*, *APOB*, *ABC1*, and *APOA1* were identified in families with severe hypercholesterolemia, defective apoB100, familial combined hyperlidemia, and extremely low HDL-C [19]. Nonetheless, mutations causing these extreme phenotypes are relatively rare. Most of the

phenotypic variation in the general population may be due to multiple common genetic variants all with a small contribution. In the 1980's and 90's numerous loci were identified in sib-pair and family linkage studies [20-24] and in association studies of variation in known lipoprotein metabolism candidate genes (apopliporproteins, lipolytic enzymes, receptors, and transporters) [25]. However, replication was inconsistent between studies [25], but the majority of these studies were statistically underpowered. More recently, significantly more statistically powered genome wide association (GWA) studies have consistently replicated genes originally identified through other approaches such as *CETP, LIPC, ABCA1, LIPG, APOE-APOC1-APOC4-APOC2* cluster, *APOB, LDLR, APOA5-APOA4-APOC3-APOA1* cluster, *GCKR,* and *LPL* and have identified loci not previously implicated in lipid metabolism [26].

As of 2005, the Human Obesity Gene Map reported 426 statistically significant associations between an obesity-related phenotype and a DNA variant in 127 candidate genes. Of these genes, 22 were supported by positive replication in at least five studies [27]. However, additional reports exist for these loci that are inconsistent with preliminary data.

In the spring of 2003, the Human Genome Project mapped more than a million SNPs, locations in the genome where a single nucleotide (A, T, C, or G) varies between individuals [28]. By observing which common SNPs are inherited alongside diseases, researchers have been able to locate genes and regions of the genome that increase an individual's risk to disease. These studies are known as GWA studies and have been made possible by the continuing advances in SNP discovery, genotyping technologies, and statistical tools to analyze large data sets. This technique allows researchers to study

the variation across the entire genome rather than focusing on a single gene or prior hypothesis. These studies have been remarkably successful [29].

Recent GWA studies in populations of primarily European descent have identified many loci that are associated with blood lipoprotein levels [30-42] and obesity [43-51]. Many variants are being identified in genes and regions of the genome that have not been previously selected as candidates for obesity and lipoproteins. Additionally, many loci are being identified as associated with multiple traits and diseases. Despite the large number of loci identified in these studies, the associated SNPs only explain a small proportion of the heritability within these traits, suggesting that many genetic factors remain to be identified. There are still other types of genetic variants not yet evaluated including but not limited to copy number variants and less common SNPs.

Furthermore, examining understudied ethnic groups may identify additional genes that are associated with disease. Asian populations are undergoing socio-economic development and lifestyle changes, which are resulting in an increase of CVD burden [52]. The environmental differences between European and Asian populations may modify the effect of genetic variants on obesity and lipoprotein levels. Additionally, because of the potential differences in genetic architecture (patterns of linkage disequilibrium and allele frequencies) between Asian and European individuals, genetic variants identified in populations of European descent may differ in Asian populations. For example, variants in KCNQ1 were first identified to be associated with type 2 diabetes in an Asian population [53,54], but it was not until later that the association was verified in Caucasians because of ancestry-related differences in the allele frequencies increasing the power to detect an association in an Asian population compared to a

European population (frequency 5% versus 40%) [55]. This example demonstrates the value of using diverse populations to identify complex-trait associations.

THE CEBU LONGITUDINAL HEALTH AND NUTRITION SURVEY (CLHNS)

The Cebu Longitudinal Health and Nutrition Survey (CLHNS) is an ongoing population study that began in 1983 (www.cpc.unc.edu/projects/cebu). The CLHNS is a birth cohort of Filipino women and their children spanning one year (May 1, 1983 to April 30, 1984), providing a representative sample of births over all seasons. This dissertation focuses on the mothers of this cohort. Baseline interviews were conducted on 3,327 women during their 6^{th} to 7^{th} month of pregnancy. Surveys took place immediately after birth, then every two months for 24 months. Follow-up surveys were conducted in 1991, 1994-5, 1998-99, 2002, 2005, and 2007. The 2005 survey collected blood for DNA and biomarkers, clinical, and phenotypic data on 1,889 un-related mothers.

The CLHNS data was collected during in-home surveys using structured questionnaires by qualified and trained interviewers. All surveys collected socioeconomic, demographic, community, environmental, diet, physical activity, and anthropometric data. Height, weight, mid-upper arm circumference, and triceps skinfold thickness were measured in all surveys; waist and hip circumferences were measured in the 1998 to 2005 surveys, and suprailiac skinfold thickness was measured in the 2005 survey. Data on menopause status was collected beginning in 1991. Complete reproductive history including the total number of pregnancies, months pregnant, and months lactating was updated at each survey. Plasma was collected in 2005 from which

biomarkers such as HDL-C, triglycerides, LDL-C, and total cholesterol have been measured.

The CLHNS was initially designed to investigate the determinants and consequences of infant feeding and maternal health; however, the wealth of data collected by this study over the past 22 years has allowed the focus of the work to expand. Most recently, researchers have been focusing on CVD risk factors including weight, fat patterning, lipoprotein levels, hypertension, and diabetes. However, I am the first to look at the genetic associations with obesity-related traits and lipoprotein levels within the women of CLHNS.

THE INTERNATIONAL HAPLOTYPE MAP: LINKAGE DISEQUILIBIRUM AND TAG SNPs

For efficient genotyping, GWA studies make use of patterns of linkage disequilibrium (LD) between genetic polymorphisms [46,47]. SNPs located near each other tend to be correlated. While a region may contain many SNPs, often only a few are needed to uniquely identify the common haplotypes across this region. When alleles of two SNPs are inherited together more often than expected based on the product of their allele frequencies, the SNPs are said to be in LD with each other. One measure of LD is the squared correlation coefficient, r^2. For example, when two SNPs are perfectly redundant, $r^2=1$; when they are completely independent, $r^2=0$. LD patterns reflect the ancestry of a population and vary considerably throughout the genome [48-50]. The International Haplotype Map Consortium has studied and cataloged DNA sequence variation and characterized the patterns of LD across the genome [3], allowing correlated

SNPs to be excluded from genotyping, but considered in analysis. The first phases of HapMap were made up of four panels of dense haplotype maps for individuals characterized as Han Chinese from Beijing, China (CHB), Japanese from Tokyo, Japan (JPT), Caucasians from Utah, USA with northern and western European ancestry (CEU), and Yoruban of Ibadan, Nigeria (YRI) and reported more than 3.1 million SNPs across the genome with an average SNP density of 1 SNP/ 760bp [3]. Most recently in 2009, HapMap phase III was released and contains genotype data from 1,115 individuals from 11 populations (http://hapmap.ncbi.nlm.nih.gov/).

Measures of LD, gene density, and haplotype blocks vary across the genome [51,52], and the HapMap ENCyclopedia Of DNA Elements (ENCODE) reference regions represent a range of these and other genomic characteristics [53]. These regions were re-sequenced in 48 unrelated individuals (8 CHB, 8 JPT, 16 CEU, and 16 YRI) for SNP discovery and reflect the actual density of SNPs in the genome more accurately than other regions in HapMap [54].

The HapMap is an important resource for choosing tag SNPs in disease association and population studies [3]. Through extensive SNP discovery and simulations, de Bakker *et al.* 2006 [55] showed that the power to detect disequilibrium-based association is only modestly compromised when an appropriate selection of tag SNPs are chosen from HapMap samples and applied to other case-control samples. For populations similar to those genotyped in the HapMap project, HapMap data may be used to impute or directly predict genotypes of non-tag SNPs for analysis in association studies [56,57]. Utilizing the patterns of haplotype variation has been very useful in inferring missing genotype *in silico* using imputation to deduce genotypes of more SNPs

across the genome. However, prior to this dissertation very little genetic data was available for Filipinos, and data was unavailable to assess the effectiveness of using the existing HapMap data to guide SNP selection and interpretation for samples from the CLHNS samples.

IDENTIFYING FUNCTIONAL VARIANTS WITHIN LOCI IMPLICATED IN DISEASE BY GWA STUDIES

In the past three years, GWA studies have successfully identified dozens of loci confirmed to be strongly associated with CVD traits [17,20]. These results are only the beginning of understanding how each locus may contribute to disease. A GWA study may report only the most associated SNP, but in some cases hundreds of variants may account for the association signal. Additionally, for some loci, the associated variants span a large number of genes or are located in regions of the genome where there are no known protein-coding genes. One of the next steps towards understanding the molecular mechanisms of obesity-related traits and lipoproteins is to identify the exact underlying variants and genes contributing to an association signal.

There are many potential molecular mechanisms for a functional SNP. In the simplest scenario, a SNP allele can cause a nonsense, frameshift, or missense substitution altering a key amino acid residue or altogether destroying the function or activity of a protein [58-60]. However, many of the recent association signals identified in GWA studies are entirely located in non-coding sequence, suggesting that non-coding variation plays an important role in common disease. Non-coding SNPs may influence biological processes by reducing transcription factor binding affinity; modifying RNA splicing or

polyadenylation; or influencing mRNA stability, microRNAs, or microRNA target sites. One loci identified in recent GWA studies, *GALNT2*, contains at least eight SNPs within the non-coding sequence of intron one that are associated with HDL-C (all $P < 1 \times 10^{-7}$) [29,33].

Most DNA is wound around a histone core, forming nucleosomes. When DNA is in this state, it is inaccessible to transcription factors, RNA polymerase, or trans-regulatory factors. DNA segments that are actively regulating transcription are characterized by the depletion of nucleosomes from the chromatin [61]. Several experimental approaches can help identify open chromatin. These include DNase I hypersensitivity (DHS) [62], formaldehyde-assisted identification of regulatory elements (FAIRE) [63], markers of histone modification [64], and chromatin immunoprecipitation (ChIP) with regulatory proteins [65]. Open chromatin can differ between tissues, and data from these techniques is becoming available for many medically relevant cell types. By combining predictions of regulatory variants with SNP data, we may be able to predict and experimentally test functional regulatory variants that contribute to CVD traits such as enhancers or promoters.

RESEARCH PRESENTED IN THIS DISSERTATION

As described in the subsequent chapters, the overall goals of this dissertation are as follows:

Chapter II.

Assess the similarity of genetic variants in the CLHNS samples with the variants in four HapMap samples. Using 627 SNPs from 10 HapMap ENCODE reference regions, I employ allele frequency estimates, pairwise LD (r^2), and haplotype frequency estimates as measures of similarity to compare four HapMap populations to the individuals of the CLHNS. Furthermore, I study the efficiency of tag SNPs selected from two HapMap East Asian panels for capturing genetic variation in CLHNS samples.

Hypothesis: The CLHNS samples are more genetically similar to the Asian HapMap samples, CHB and JPT, because they are geographically related.

Chapter III.

Examine whether selected SNPs reported to be associated with obesity-related traits in other populations are also associated with similar traits in the CLHNS samples. I perform tests of association in the CLHNS using SNPs selected from the literature based on previous evidence of association with an obesity-phenotype in other populations. I test association with BMI, waist circumference, and percent body fat for all SNPs. For SNPs with evidence for association with at least one trait, I further test for association with additional phenotypes: baseline BMI (measured in 1983-84), weight, fat mass, Triceps skinfold thickness, suprailiac skinfold thickness, arm fat area, arm muscle area, and height. I also examine whether these SNPs modify the effect of established environmental risk factors of diet and physical activity level by performing tests of interaction between genotype and total caloric intake, estimated percent dietary fat and carbohydrates, and activity level. Additionally I perform a longitudinal analysis

incorporating all available BMI measurements (for the up to eight measurements) spanning 22 years.

Hypothesis: Some but not all of the previously reported SNPs are associated with obesity-phenotypes in the CLHNS samples reflecting environmental and genetic differences between the CLHNS cohort and previously studied populations, statistical power, and false positive results in the literature.

Chapter IV

Test evidence of association between SNPs and plasma lipoprotein levels in the CLHNS samples in a GWA study. Using the Affymetrix Genome-Wide Human SNP Array 5.0, I genotype 1,889 CLHNS women on ~500K SNPs and perform association analysis on 2.3 million genotyped or imputed SNPs for HDL-C, LDL-C, triglycerides, and total cholesterol. I perform conditional analyses to evaluate whether any additional SNPs were independently associated with an outcome after accounting for the most associated SNP at the association signal.

Hypothesis: Some of the previously reported associated SNPs show evidence of association with lipoprotein levels in the CLHNS samples and additional putative novel loci are identified to be associated in the CLHNS population.

Chapter V.

Evaluate allele-specific effects on transcriptional regulation for HDL-C associated SNPs located in intron 1 of the *GALNT2* gene. After developing a more comprehensive list of potential functional variants in the *GALNT2* association region, I

prioritize the variants based on sequence annotation and maps of open chromatin from hepatocytes and test high priority variants for allele-specific effects on promoter and enhancer activity.

Hypothesis: A functional SNP(s) within the region of association may act as an enhancer or promoter.

Chapter VI.

Synthesize the findings of the research chapters and describe the future of genetics of obesity-related traits and lipoproteins. Together these findings represent important contributions to the field of human genetics and will serve as building blocks for future studies.

REFERENCES

1. National Cholesterol Education Program Expert Panel on Detection Evaluation and Treatment of High Blood Cholesterol in Adults (2002) 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. Circulation 106: 3143-3421.

2. World Health Organization: Fact sheet N°317 - Cardiovascular diseases (CVDs). 2009, http://www.who.int/mediacentre/factsheets/fs317/en/index.html.

3. International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851-861.

4. Grundy SM, Pasternak R, Greenland P, Smith S, Jr., Fuster V (1999) Assessment of cardiovascular risk by use of multiple-risk-factor assessment equations: a statement for healthcare professionals from the American Heart Association and the American College of Cardiology. Circulation 100: 1481-1492.

5. Pilia G, Chen WM, Scuteri A, Orru M, Albai G, et al. (2006) Heritability of cardiovascular and personality traits in 6,148 Sardinians. PLoS Genet 2: e132.

6. Ueshima H, Sekikawa A, Miura K, Turin TC, Takashima N, et al. (2008) Cardiovascular disease and risk factors in Asia: a selected review. Circulation 118: 2702-2709.

7. World Health Organization: Fact sheet N°311– Obesity and overweight. 2006, http://www.who.int/mediacentre/factsheets/fs311/en/index.html.

8. Yang W, Kelly T, He J (2007) Genetic epidemiology of obesity. Epidemiol Rev 29: 49-61.

9. World Health Organization Expert Consultation (2004) Appropriate body-mass index for Asian populations and its implications for policy and intervention strategies. Lancet 363: 157-163.

10. Willett WC (1998) Anthropometric measures and body composition. Nutritional Epidemiology. New York: Oxford University Press. pp. 244-272.

11. Macaky, J. and Mensah, G.A. The atlas of heart disease and stroke (World Health Organization, Geneva, 2004).

12. Law MR, Wald NJ, Rudnicka AR (2003) Quantifying effect of statins on low density lipoprotein cholesterol, ischaemic heart disease, and stroke: systematic review and meta-analysis. BMJ 326: 1423.

13. Kuulasmaa K, Tunstall-Pedoe H, Dobson A, Fortmann S, Sans S, et al. (2000) Estimation of contribution of changes in classic risk factors to trends in coronary-event rates across the WHO MONICA Project populations. Lancet 355: 675-687.

14. Grundy SM, Cleeman JI, Merz CN, Brewer HB, Jr., Clark LT, et al. (2004) Implications of recent clinical trials for the National Cholesterol Education Program Adult Treatment Panel III guidelines. Circulation 110: 227-239.

15. Gotto AM, Jr., Brinton EA (2004) Assessing low levels of high-density lipoprotein cholesterol as a risk factor in coronary heart disease: a working group report and update. J Am Coll Cardiol 43: 717-724.

16. Nordestgaard BG, Benn M, Schnohr P, Tybjaerg-Hansen A (2007) Nonfasting triglycerides and risk of myocardial infarction, ischemic heart disease, and death in men and women. JAMA 298: 299-308.

17. O'Connell DL, Heller RF, Roberts DC, Allen JR, Knapp JC, et al. (1988) Twin study of genetic and environmental effects on lipid levels. Genet Epidemiol 5: 323-341.

18. Mohlke KL, Boehnke M, Abecasis GR (2008) Metabolic and cardiovascular traits: an abundance of recently identified common genetic variants. Hum Mol Genet 17: R102-108.

19. Breslow JL (2000) Genetics of lipoprotein abnormalities associated with coronary artery disease susceptibility. Annu Rev Genet 34: 233-254.

20. Ciccarese M, Pacifico A, Tonolo G, Pintus P, Nikoshkov A, et al. (2000) A new locus for autosomal recessive hypercholesterolemia maps to human chromosome 15q25-q26. Am J Hum Genet 66: 453-460.

21. Cohen JC, Vega GL, Grundy SM (1999) Hepatic lipase: new insights from genetic and metabolic studies. Curr Opin Lipidol 10: 259-267.

22. Haddad L, Day IN, Hunt S, Williams RR, Humphries SE, et al. (1999) Evidence for a third genetic locus causing familial hypercholesterolemia. A non-LDLR, non-APOB kindred. J Lipid Res 40: 1113-1122.

23. Knoblauch H, Muller-Myhsok B, Busjahn A, Ben Avi L, Bahring S, et al. (2000) A cholesterol-lowering gene maps to chromosome 13q. Am J Hum Genet 66: 157-166.

24. Varret M, Rabes JP, Saint-Jore B, Cenarro A, Marinoni JC, et al. (1999) A third major locus for autosomal dominant hypercholesterolemia maps to 1p34.1-p32. Am J Hum Genet 64: 1378-1387.

25. Ordovas JM (2002) HDL genetics: candidate genes, genome wide scans and gene-environment interactions. Cardiovasc Drugs Ther 16: 273-281.

26. Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, et al. (2008) Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet 40: 161-169.

27. Rankinen T, Zuberi A, Chagnon YC, Weisnagel SJ, Argyropoulos G, et al. (2006) The human obesity gene map: the 2005 update. Obesity (Silver Spring) 14: 529-644.

28. Collins FS, Morgan M, Patrinos A (2003) The Human Genome Project: lessons from large-scale biology. Science 300: 286-290.

29. Hindorff LA, Junkins HA, Mehta JP, and Manolio TA. A catalog of published genome-wide association studies. Available at: www.genome.gov/gwastudies.

30. Chasman DI, Pare G, Zee RYL, Parker AN, Cook NR, et al. (2008) Genetic loci associated with plasma concentration of low-density lipoprotein cholesterol, high-density lipoprotein cholesterol, triglycerides, apolipoprotein a1, and apolipoprotein b among 6382 white women in genome-wide analysis with replication. Circ Cardiovasc Genet 1: 21-30.

31. Aulchenko YS, Ripatti S, Lindqvist I, Boomsma D, Heid IM, et al. (2009) Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. Nat Genet 41: 47-55.

32. Burkhardt R, Kenny EE, Lowe JK, Birkeland A, Josowitz R, et al. (2008) Common SNPs in HMGCR in micronesians and whites associated with LDL-cholesterol levels affect alternative splicing of exon13. Arterioscler Thromb Vasc Biol 28: 2078-2084.

33. Heid IM, Boes E, Muller M, Kollerits B, Lamina C, et al. (2008) Genome-wide association analysis of high-density lipoprotein cholesterol in the population-based KORA study sheds new light on intergenic regions. Circ Cardiovasc Genet 1: 10-20.

34. Kathiresan S, Willer CJ, Peloso GM, Demissie S, Musunuru K, et al. (2009) Common variants at 30 loci contribute to polygenic dyslipidemia. Nat Genet 41: 56-65.

35. Kooner JS, Chambers JC, Aguilar-Salinas CA, Hinds DA, Hyde CL, et al. (2008) Genome-wide scan identifies variation in MLXIPL associated with plasma triglycerides. Nat Genet 40: 149-151.

36. Pollin TI, Damcott CM, Shen H, Ott SH, Shelton J, et al. (2008) A null mutation in human APOC3 confers a favorable plasma lipid profile and apparent cardioprotection. Science 322: 1702-1705.

37. Sabatti C, Service SK, Hartikainen AL, Pouta A, Ripatti S, et al. (2009) Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. Nat Genet 41: 35-46.

38. Kathiresan S, Melander O, Guiducci C, Surti A, Burtt NP, et al. (2008) Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. Nat Genet 40: 189-197.

39. Sandhu MS, Waterworth DM, Debenham SL, Wheeler E, Papadakis K, et al. (2008) LDL-cholesterol concentrations: a genome-wide association study. Lancet 371: 483-491.

40. Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, et al. (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 316: 1331-1336.

41. Wallace C, Newhouse SJ, Braund P, Zhang F, Tobin M, et al. (2008) Genome-wide association study identifies genes for biomarkers of cardiovascular disease: serum urate and dyslipidemia. Am J Hum Genet 82: 139-149.

42. Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, et al. (2008) Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet 40: 161-169.

43. Chambers JC, Elliott P, Zabaneh D, Zhang W, Li Y, et al. (2008) Common genetic variation near MC4R is associated with waist circumference and insulin resistance. Nat Genet 40: 716-718.

44. Loos RJ, Lindgren CM, Li S, Wheeler E, Zhao JH, et al. (2008) Common variants near MC4R are associated with fat mass, weight and risk of obesity. Nat Genet 40: 768-775.

45. Thorleifsson G, Walters GB, Gudbjartsson DF, Steinthorsdottir V, Sulem P, et al. (2009) Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. Nat Genet 41: 18-24.

46. Willer CJ, Speliotes EK, Loos RJ, Li S, Lindgren CM, et al. (2009) Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. Nat Genet 41: 25-34.

47. Meyre D, Delplanque J, Chevre JC, Lecoeur C, Lobbens S, et al. (2009) Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. Nat Genet 41: 157-159.

48. Cho YS, Go MJ, Kim YJ, Heo JY, Oh JH, et al. (2009) A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. Nat Genet 41: 527-534.

49. Soranzo N, Rivadeneira F, Chinappen-Horsley U, Malkina I, Richards JB, et al. (2009) Meta-analysis of genome-wide scans for human adult stature identifies novel loci and associations with measures of skeletal frame size. PLoS Genet 5: e1000445.

50. Lindgren CM, Heid IM, Randall JC, Lamina C, Steinthorsdottir V, et al. (2009) Genome-wide association scan meta-analysis identifies three loci influencing adiposity and fat distribution. PLoS Genet 5: e1000508.

51. Heard-Costa NL, Zillikens MC, Monda KL, Johansson A, Harris TB, et al. (2009) NRXN3 is a novel locus for waist circumference: a genome-wide association study from the CHARGE Consortium. PLoS Genet 5: e1000539.

52. Yusuf S, Reddy S, Ounpuu S, Anand S (2001) Global burden of cardiovascular diseases: Part II: variations in cardiovascular disease by specific ethnic groups and geographic regions and prevention strategies. Circulation 104: 2855-2864.

53. Unoki H, Takahashi A, Kawaguchi T, Hara K, Horikoshi M, et al. (2008) SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. Nat Genet 40: 1098-1102.

54. Yasuda K, Miyake K, Horikawa Y, Hara K, Osawa H, et al. (2008) Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus. Nat Genet 40: 1092-1097.

55. McCarthy MI (2008) Casting a wider net for diabetes susceptibility genes. Nat Genet 40: 1039-1040.

56. Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, et al. (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. American Journal of Human Genetics 74: 106-120.

57. Service S, DeYoung J, Karayiorgou M, Roos JL, Pretorious H, et al. (2006) Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. Nature genetics 38: 556-560.

58. Plagnol V, Wall JD (2006) Possible ancestral structure in human populations. PLoS Genet 2: e105.

59. Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, et al. (2001) Linkage disequilibrium in the human genome. Nature 411: 199-204.

60. Sawyer SL, Mukherjee N, Pakstis AJ, Feuk L, Kidd JR, et al. (2005) Linkage disequilibrium patterns vary substantially among populations. Eur J Hum Genet 13: 677-686.

61. Ke X, Durrant C, Morris AP, Hunt S, Bentley DR, et al. (2004) Efficiency and consistency of haplotype tagging of dense SNP maps in multiple samples. Human molecular genetics 13: 2557-2565.

62. De La Vega FM, Isaac H, Collins A, Scafe CR, Halldorsson BV, et al. (2005) The linkage disequilibrium maps of three human chromosomes across four

populations reflect their demographic history and a common underlying recombination pattern. Genome research 15: 454-462.

63. Encode Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306: 636-640.

64. International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437: 1299-1320.

65. de Bakker PI, Burtt NP, Graham RR, Guiducci C, Yelensky R, et al. (2006) Transferability of tag SNPs in genetic association studies in multiple populations. Nature genetics 38: 1298-1303.

66. Eyheramendy S, Marchini J, McVean G, Myers S, Donnelly P (2007) A model-based approach to capture genetic variation for future association studies. Genome research 17: 88-95.

67. Paschou P, Mahoney MW, Javed A, Kidd JR, Pakstis AJ, et al. (2007) Intra- and interpopulation genotype reconstruction from tagging SNPs. Genome research 17: 96-107.

68. Bagnall RD, Roberts RG, Mirza MM, Torigoe T, Prescott NJ, et al. (2008) Novel isoforms of the CARD8 (TUCAN) gene evade a nonsense mutation. Eur J Hum Genet 16: 619-625.

69. Clendenning M, Senter L, Hampel H, Robinson KL, Sun S, et al. (2008) A frame-shift mutation of PMS2 is a widespread cause of Lynch syndrome. J Med Genet 45: 340-345.

70. Hani EH, Boutin P, Durand E, Inoue H, Permutt MA, et al. (1998) Missense mutations in the pancreatic islet beta cell inwardly rectifying K+ channel gene (KIR6.2/BIR): a meta-analysis suggests a role in the polygenic basis of Type II diabetes mellitus in Caucasians. Diabetologia 41: 1511-1515.

71. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447: 799-816.

72. Crawford GE, Holt IE, Mullikin JC, Tai D, Blakesley R, et al. (2004) Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. Proc Natl Acad Sci U S A 101: 992-997.

73. Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD (2007) FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res 17: 877-885.

74. Gilbert N, Ramsahoye B (2005) The relationship between chromatin structure and transcriptional activity in mammalian genomes. Brief Funct Genomic Proteomic 4: 129-142.

75. Elnitski L, Jin VX, Farnham PJ, Jones SJ (2006) Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. Genome Res 16: 1455-1464.

CHAPTER II. COMPARISON OF ENCODE REGION SNPS BETWEEN CEBU FILIPINO AND ASIAN HAPMAP SAMPLES

A version of this work was previously published as:

Amanda F. Marvelle, Leslie A. Lange, Li Qin, Yunfei Wang, Ethan M. Lange, Linda S. Adair, Karen L. Mohlke. 2007 Comparison of ENCODE region SNPs between Cebu Filipino and Asian HapMap samples. *Journal of Human Genetics*. 52(9):729-37.

CHAPTER II CREDITS

I would like to thank Sandra German at the Office of Population Studies in Cebu Philippines for blood sample collection and processing, under the direction of Dr. Christopher Kuzawa of Northwest University. I thank Amy Perou of the biospecimen processing facility, Jason Luo of the mammalian genotyping core, and Laura Livingstone of the automated DNA sequencing facility at University of North Carolina at Chapel Hill. Cebu Filipino data collection was supported by TW05596, specimen processing and genotyping was supported by pilot funds from NIH grants RR20649 (Interdisciplinary Obesity Center), ES10126 (Project 7-2004-E of the Center for Environmental Health and Susceptibility), and DK56350 (Clinical Nutrition Research Center), and analysis funded in part by grant DK78150. A.F.M. was supported by an Integrative Vascular Biology Fellowship, NIH grant HL69768.

ABSTRACT

Patterns of linkage disequilibrium (LD) act as the framework for designing efficient association studies; these patterns are being studied and catalogued by The International HapMap Project. The current study assessed the transferability of tag SNPs chosen from HapMap panels to a cohort of 80 individuals from metro Cebu, Philippines, who participated in the Cebu Longitudinal Health and Nutrition Survey (CLHNS). The analyses focused on 627 single nucleotide polymorphisms (SNPs) in the central 40 kb within each of the 10 HapMap ENCODE regions. The similarity between the genetic variants in Cebu Filipino samples and HapMap panels was examined using allele frequency estimates, measures of pairwise LD, and haplotype frequency estimates. For these measures, strong correlations were observed between the Cebu Filipino samples and the Asian panels from HapMap, with the strongest correlations observed with the Han Chinese from Beijing (CHB) panel. Tag SNPs selected using the HapMap CHB panel were particularly effective at representing the genetic variation in Cebu Filipino samples. These results suggest that the HapMap data will be an effective resource for future studies in Cebu Filipino samples.

INTRODUCTION

Genetic association studies make use of patterns of linkage disequilibrium (LD) between genetic polymorphisms for efficient genotyping [1,2]. LD patterns reflect the ancestry of a population and vary considerably throughout the genome [3-5]. The International HapMap Consortium is studying and cataloguing DNA sequence variation and characterizing these patterns of LD across the genome [6], allowing correlated single nucleotide polymorphisms (SNPs) to be excluded from genotyping but considered in analysis. The HapMap is made up of four panels of dense haplotype maps for individuals characterized as Han Chinese from Beijing, China (CHB), Japanese from Tokyo, Japan (JPT), Caucasians from Utah, USA with northern and western European ancestry (CEU), and Yoruban of Ibadan, Nigeria (YRI).

The HapMap is an important resource for choosing tag SNPs in disease association and population studies [6]. However, worldwide population variation is not completely characterized, and an essential question is whether tag SNPs chosen using HapMap panels will adequately capture patterns of genetic variation in other populations [7-17]. Furthermore, for populations similar to those genotyped in the HapMap project, HapMap data may be used to directly predict genotypes of non-tag SNPs for analysis in association studies [18,19]. Previous studies observed that the HapMap CHB and JPT panels have very similar patterns of LD and could act as a proxy for other geographically related populations [20-25]. However, data is currently unavailable to assess the effectiveness of using the existing HapMap data to guide SNP selection and interpretation for samples from the Cebu Longitudinal Health and Nutrition Survey (CLHNS) cohort from metro Cebu in the central Philippines [26,27]. This study assesses the advantages of

having HapMap data from two related Asian panels to evaluate whether the combined CHB and JPT panels would more effectively capture genetic variability in Cebu Filipinos than the CHB or JPT panels alone. In addition, this study develops the most efficient criteria for selecting tag SNPs from the HapMap panels for future genetic association studies in Cebu Filipino samples.

To address these issues, SNPs from within the 10 HapMap ENCyclopedia Of DNA Elements (ENCODE) reference regions were used [28]. These regions were re-squenced in 48 unrelated individuals (8 CHB, 8 JPT, 16 CEU, and 16 YRI) for SNP discovery and reflect the density of SNPs in the genome more accurately than other regions in HapMap. The SNP density in these ENCODE regions is higher than the remainder of HapMap [6]. The similarity of the HapMap samples to 80 Cebu Filipino samples was assessed, using allele frequency estimates, pairwise LD (r^2), and haplotype frequency estimates as measures of similarity. Furthermore, the efficiency of using tag SNPs selected from the HapMap Asian panels for capturing genetic variation in Cebu Filipino samples was studied.

MATERIALS AND METHODS

Samples

Eighty unrelated Cebu Filipino individuals were randomly selected from a cohort of healthy women from the CLHNS (www.cpc.unc.edu/projects/cebu). Informed consent was obtained from all individuals and the study protocol was approved by the University of North Carolina Institutional Review Board for the Protection of Human Subjects.

Genomic DNA was isolated from peripheral blood lymphocytes using automated and manual DNA extraction methods (Puregene, Gentra) by the University of North Carolina, Chapel Hill BioSpecimen Processing Facility. Centre d'Etude du Polymorphisme Humain (CEPH) DNA samples were obtained from Coriell (Camden, NJ, USA).

HapMap genotype data was obtained from the HapMap database (www.hapmap.org) for all available unrelated individuals, including 45 CHB, 44 JPT, 60 CEU parents of trios, and 60 YRI parents of trios. For some analysis, the CHB and JPT samples were combined (indicated as CHB+JPT) [6].

SNP selection and genotyping

To represent the overall complexity of the genome, the central 40 kb region from within each of the ten 500 kb ENCODE regions that have been used for SNP discovery and dense SNP genotyping was chosen for this study (Table 2.1). SNPs were selected if they were polymorphic (minor allele frequency, MAF, >0) in the HapMap CHB, JPT, or CEU panels.

Of the 883 SNPs that met these criteria, 215 were eliminated based on Illumina design score (calculated December 2005). One SNP identified by re-sequencing region ENr213 in Cebu Filipino samples (see below) was included, resulting in a total of 669 SNPs that were genotyped in the Cebu Filipino samples. SNP genotyping was performed at the Mammalian Genotyping Core at the University of North Carolina, Chapel Hill using the Illumina GoldenGate (Illumina Inc., San Diego, USA) genotyping assay [29]. Of the 669 SNPs attempted, 36 SNPs were excluded based on poorly-defined clusters

(n=28), genotyping completeness <90% (n=3), or inconsistency with Hardy-Weinberg Equilibrium (P < .001) (n=5). Six additional SNPs were excluded because of two or more genotype discrepancies between six CEPH DNA samples and equivalent HapMap CEU genotypes. SNPs were also evaluated for two or more genotyping discrepancies between seven duplicate samples, however no SNPs needed to be dropped based on this criterion. The genotyping success rate of the final 627 SNPs was 99.9% and the discrepancy rate was .02%. Of these 627 SNPs, 501 (80%) were polymorphic (MAF > 0) in Cebu Filipino samples. The average marker spacing of these 501 polymorphic SNPs was 1 SNP/798 bp.

Statistical analysis

Tests for consistency of genotype distributions with expected Hardy-Weinberg equilibrium proportions were calculated using standard Pearson's χ^2 statistics. Only markers with a MAF $\geq$ .05 in HapMap panels were analyzed in Cebu Filipino samples. SNPs were matched for the reference allele between all HapMap panels and Cebu Filipino samples. Fisher's exact tests were used to test for allele frequency differences between pairs of samples. Pair-wise LD (r^2) values were calculated using Haploview [30; http://www.broad.mit.edu/mpg/haploview] for adjacent pairs and all pairs of SNPs in each region. Haplotype blocks were defined in Haploview for each HapMap panel based on the default block definition [31]. Identical blocks from each HapMap panel were defined in the Cebu Filipino samples for comparison. Using Haploview, haplotype frequencies were estimated in each haplotype block for every population. Haplotypes with a frequency > .01 were evaluated in the HapMap panels. Haplotypes not observed

in the Cebu Filipino samples were assigned a frequency of zero. Spearman's correlation coefficients were calculated for all comparisons between Cebu Filipino samples and HapMap panels.

In order to evaluate the efficiency of HapMap to choose tag SNPs for Cebu Filipino samples, tag SNPs from HapMap panels were selected using Tagger in pairwise tagging mode with other settings at default values [32; http://www.broad.mit.edu/mpg/tagger/]. Several r^2 thresholds were used to assess the performance of selecting tag SNPs using the HapMap panels: .80, .85, .90, and .95. If a Cebu Filipino SNP exhibited pairwise $r^2 \geq .80$ with at least one tag SNP (selected from the Asian HapMap panels), then the SNP was defined as captured in the Cebu Filipino sample. Percent coverage for a region is defined as the number of captured SNPs in the Cebu Filipino samples divided by the total number of SNPs (with estimated MAF $\geq .05$). Finally, for each Cebu Filipino SNP, the maximum r^2 estimate obtained over all r^2 estimates between that SNP and a tag SNP in the region was identified. For a region, mean maximum r^2 was defined as the average value of the maximum r^2 values obtained over all Cebu Filipino SNPs in the region.

Re-sequencing

Twenty-four randomly chosen Cebu Filipino samples were re-sequenced in the central 800 nucleotide (nt) region within each of the 40 kb ENCODE regions. Primers were selected using Primer3 software [33; http://www.genome.wi.mit.edu/genome_software/other/primer3.html] and sequences were compared using Sequencher 4.2.2 (Gene Codes Corporation, Ann Arbor, MI, USA).

Sequencing was performed at the University of North Carolina, Chapel Hill automated DNA sequencing facility on an ABI Prism 3730 (Applied Biosystems, Foster City, CA, USA) using the Big Dye Terminator Kit.

RESULTS

To determine the extent of similarity between 80 Cebu Filipino samples and HapMap samples, genotype data for 627 SNPs located within the 10 HapMap ENCODE regions was used (Table 2.1).

Allele frequencies

Allele frequency estimates were compared using SNPs with MAF ≥ .05 in the corresponding HapMap panel. A total of 399 SNPs were evaluated when examining CHB, 391 SNPs for JPT, 396 SNPs for CHB+JPT, 431 SNPs for CEU, and 391 SNPs for YRI. The Spearman's correlation coefficients for allele frequency estimates between the Cebu Filipino samples and the HapMap panels were .96, .92, .95, .82, and .65 for CHB, JPT, CHB+JPT, CEU, and YRI, respectively (Figure 2.1). For comparison, the Spearman's correlation coefficient for allele frequency estimates between CHB and JPT samples was .95 for 384 SNPs with MAF ≥ .05 in both panels. The percent of SNPs with significantly different allele frequencies (Fisher's exact p-value < .01) was 5.7% for CHB, 15.6% for JPT, 11.6% for CHB+JPT, 57.7% for CEU, and 60.1% for YRI. Although larger sample sizes should provide greater power to detect statistically significant differences, the 89 CHB+JPT samples showed fewer significant differences with the Cebu Filipino samples than the smaller JPT, CEU, or YRI groups. The allele frequency comparison was repeated using HapMap SNPs with MAF > 0 and slightly

higher Spearman's correlations were obtained with analogous patterns of similarity (data not shown).

Based on the substantially greater similarity in allele frequencies between Cebu Filipino samples and Asian HapMap panels compared to CEU or YRI panels, subsequent analyses were performed using only the HapMap CHB, JPT and CHB+JPT panels.

Linkage disequilibrium

Pairwise r^2 for adjacent pairs and all pairs of SNPs within each HapMap ENCODE region in the Asian HapMap samples and Cebu Filipino samples were estimated to evaluate the extent of LD in each population. Only SNPs with MAF $\geq .05$ in the corresponding HapMap sample were included in comparisons. Analysis was performed for 375, 368, and 373 adjacent pairs of SNPs and 9350, 8912, and 9157 total pairs of SNPs for CHB, JPT, and CHB+JPT, respectively. The Spearman's correlation coefficients of the r^2 estimates for adjacent pairs of SNPs between the Cebu Filipino and Asian HapMap samples were .90 for each of CHB, JPT, and CHB+JPT. The Spearman's correlation coefficients of the r^2 estimates for all pairs were .88 for CHB, .87 for JPT, and .89 for CHB+JPT (Table 2.2). The absolute difference between r^2 estimates of adjacent SNP pairs was calculated. For CHB 51%, 73%, and 85% of the SNPs had absolute differences between r^2 estimates of $\leq .05$, $\leq .10$, and $\leq .15$; for JPT 48%, 67%, and 80% of the SNPs had absolute differences between r^2 estimates of $\leq .05$, $\leq .10$, and $\leq .15$; and for CHB+JPT 50%, 69%, and 83% of the SNPs had absolute differences between r^2 estimates of $\leq .05$, $\leq .10$, and $\leq .15$.

When each of the ten regions was analyzed separately, LD differed both among regions and populations. Region ENr232 varied the most between HapMap panels; Spearman's correlation coefficients of the r^2 estimates, for all pairs of SNPs, between the Cebu Filipino and Asian HapMap samples were .85 for CHB, .63 for JPT, and .77 for CHB+JPT. This region, however, did not differ from the other regions in allele frequency estimates, haplotype frequency estimates (below), and tag SNP analyses (below). The pairwise r^2 analysis was repeated using all HapMap SNPs with MAF > 0 and obtained slightly higher Spearman's correlations but analogous patterns of similarity (data not shown). To confirm that Cebu Filipino sample size did not impact results, the analysis was repeated with three random sets of 45 Cebu Filipino samples. The sets of 45 were compared to CHB and JPT panels, and similar results were observed to the total set of 80 Cebu Filipino samples (data not shown). On average across all regions, Cebu Filipino samples show highly similar patterns of LD compared to all Asian panels, with slightly more similarity observed with CHB+JPT panels and slightly less observed with JPT panels.

Haplotype frequencies

Haplotype frequencies for the Asian HapMap panels and Cebu Filipino samples were estimated for haplotypes comprised of SNPs with MAF ≥ .05 in the HapMap panel. Haplotype blocks were defined using the default block definition used in Haploview [31]. Within the 10 regions, the average number of blocks per region was 3.6, 3.6, and 3.3 for CHB, JPT and CHB+JPT respectively. The blocks ranged in size from 2 to 65 SNPs with an average of 9.7, 10.9, and 9.1 SNPs per block for CHB, JPT, and CHB+JPT. One-

hundred and seventy-eight, 151, and 141 haplotypes were identified with frequency estimates > .01 in CHB, JPT, and CHB+JPT, respectively. The Spearman's correlation coefficient of haplotype frequency estimates between Cebu Filipino and Asian HapMap samples was .95 for CHB, .88 for JPT, and .92 for CHB+JPT (Figure 2.2). Most haplotypes with an estimated frequency > 0 in the Asian samples were also observed (with estimated frequency > 0) in Cebu Filipino samples, demonstrating a high degree of haplotype conservation across the populations. Of the observed haplotypes with estimated frequency ≥ .05 in CHB, JPT, and CHB+JPT, only 2.5% (3 of 119), 2.8% (3 of 107), and 1.8% (2 of 112), respectively, were not observed in Cebu Filipino samples. In addition, of the observed haplotypes with estimated frequency > .01 in CHB, JPT, and CHB+JPT, 23% (41 of 178), 24% (36 of 151), and 11% (16 of 141), respectively, were not observed in Cebu Filipino samples. The greater representation of Cebu Filipino haplotypes in CHB+JPT samples is likely attributed to the larger sample size. Overall, the haplotype frequency differences were modest between Cebu Filipino samples and the Asian HapMap panels, with CHB showing the most similarity and JPT showing the least similarity.

Transferability of tag SNPs

To measure the efficiency of using the HapMap panels for tag SNP selection in the Cebu Filipino population, Tagger was used to select tag SNPs from the CHB, JPT, and CHB+JPT panels for SNPs with MAF ≥ .05. Tag SNP coverage was tested at four r^2 thresholds for selection in the HapMap panels, and the tag SNPs chosen in HapMap panels were applied to SNPs with MAF ≥ .05 in Cebu Filipino samples.

Overall, at each r^2 selection threshold using the CHB, JPT, and CHB+JPT panels the percentages of SNPs captured (with a mean $r^2 \geq .80$) in the Cebu Filipino samples were very similar. Using any of the three panels for SNP selection, the lowest r^2 selection threshold of .80 captured at least 82-83% of Cebu Filipino SNPs (MAF $\geq .05$) across all ten regions (Table 2.3). To obtain this percent coverage 121, 118, and 125 tag SNPs from CHB, JPT, and CHB+JPT, respectively, would need to be genotyped. As expected, increasing the r^2 threshold for selecting tag SNPs in the Asian HapMap samples increased both the number of tag SNPs that needed to be genotyped and the proportion of Cebu Filipino SNPs captured by these tag SNPs. However, the percent coverage of each region varied substantially. At the r^2 selection threshold of .80, the percent coverage ranged over the 10 regions from 54% to 96%, 59% to 96%, and 52% to 94% using CHB, JPT, and CHB+JPT tag SNPs, respectively. This variability between regions was still observed at a r^2 selection threshold of .95, the highest r^2 threshold studied. In addition, at each r^2 selection threshold, the mean maximum r^2 of all Cebu Filipino SNPs (MAF $\geq .05$) was similar between CHB, JPT, and CHB+JPT. Among all SNPs, for the r^2 selection threshold of .80, a mean maximum r^2 of .88 was observed using CHB, JPT, and CHB+JPT tag SNPs. As expected, the mean maximum r^2 increased at each increase of the r^2 selection threshold. Little variability was observed between regions (data not shown).

At each r^2 selection threshold, the SNPs in the Cebu Filipino samples that were not captured by a tag SNP selected in the HapMap panels were evaluated. The percentage of SNPs not captured and the mean maximum r^2 for each SNP were calculated (Table 2.3). Consistent with the sensitivity of tag SNP selection to allele frequency [34],

many of the SNPs not captured were rare (MAF < .10). These rare SNPs had low mean maximum r^2, and were not captured using higher r^2 selection thresholds. Common SNPs (MAF ≥ .10) that were not captured at an r^2 of at least .80 were captured with at least a mean maximum r^2 of .65, .64, and .66 using CHB, JPT, and CHB+JPT tag SNPs, respectively. As the r^2 selection threshold increased, more of these Cebu Filipino common SNPs were captured with an r^2 of at least .80.

Re-sequencing

To assess the frequency of population-specific novel SNPs and to further evaluate the genetic structure in Cebu Filipinos, 24 Cebu Filipino individuals were re-sequenced in an 800 nt region within each of the 10 HapMap ENCODE regions used previously for HapMap re-sequencing [28]. Approximately 184 kb on at least one DNA strand were re-sequenced. Only one novel SNP was detected that was not present in HapMap (data release 21, July 06) or dbSNP (build 126); the SNP was located in region ENr213 (ss69374772) and had a MAF of .05 in 80 Cebu Filipino individuals. Within Cebu Filipino samples, this SNP exhibited a maximum r^2 of .228 with 4 other SNPs in the 40 kb region.

DISCUSSION

The extent of similarity between Cebu Filipino samples and the previously evaluated HapMap samples were examined using measures of allele frequency estimates, pairwise r^2 estimates, and haplotype frequency estimates. Consistent with population migration, mitochondrial DNA, and Y haplotype patterns [35], CEU and YRI samples

were much less similar to Cebu Filipino samples with respect to allele frequency than CHB, JPT, or CHB+JPT samples. All of the analyses showed high similarity between Asian HapMap samples and Cebu Filipino samples.

Because the CHB and JPT samples have similar allele frequencies, these data sets are often combined for analyses [6]. The existence of these two Asian HapMap panels allowed for evaluating the choice of using CHB, JPT, or the larger combined CHB+JPT panel as a resource for choosing haplotype tagging SNPs for Cebu Filipino samples. Among these three panels, JPT samples were the least correlated with Cebu Filipino samples with respect to allele frequency estimates, pairwise r^2 estimates, and haplotype frequency estimates. Cebu Filipino and CHB allele frequency estimates were more closely correlated than CHB and JPT allele frequency estimates. Both CHB and CHB+JPT panels were very similar to Cebu Filipino samples, and it is not clear which panel would act most efficiently as a proxy for the Cebu Filipino samples. The larger CHB+JPT sample size would be expected to decrease the variability in the allele and haplotype frequency estimates; the added JPT samples could decrease accuracy. Indeed, estimated Cebu Filipino allele and haplotype frequencies were slightly more correlated with CHB than CHB+JPT, but Cebu Filipino pairwise r^2 estimates were slightly more similar to CHB+JPT than CHB.

A practical use of HapMap is to select tag SNPs for regional or genome-wide association studies [6]. Evaluation was performed on the transferability of HapMap tag SNPs chosen using the data from CHB, JPT, and CHB+JPT panels at several r^2 selection thresholds, with respect to capturing the genetic variability in samples from Cebu, Philippines. Using these criteria, at an r^2 selection threshold of .80, the HapMap-based

tag SNPs capture 82-83% of the Cebu Filipino SNPs. A majority of the most common SNPs (MAF ≥ .10) in the Cebu Filipino sample that are not captured by the tag SNPs at an r^2 of at least .80 are captured with an r^2 of at least .60. Using higher r^2 thresholds for tag SNP selection in the HapMap samples results in capturing more SNPs in the Cebu Filipino sample, but with the added cost of genotyping more tag SNPs. Increasing the r^2 threshold failed to capture substantially more rare SNPs, most of which exhibited low pairwise LD with other SNPs.

Previously, de Bakker *et al.* [22] showed through extensive SNP discovery and simulations that power to detect disequilibrium-based association is only modestly compromised when an appropriate selection of tag SNPs are chosen from HapMap samples and applied to other case-control samples. Large scale SNP discovery and power simulations were beyond the scope of this study. However, based on the findings from de Bakker *et al.* [22] and the current findings that tag SNPs selected using the Asian HapMap adequately captured common Cebu Filipino SNPs, the average loss in power to detect common casual alleles should be small.

Re-sequencing and genotyping was preformed in the 10 HapMap ENCODE regions that were re-sequenced for SNP discovery and are considered to be a "gold standard" because of the high density of SNP coverage [28]. Only one SNP (estimated MAF = .05) was detected in the Cebu Filipino samples that was not observed in dbSNP or HapMap, suggesting that alleles ascertained from the HapMap ENCODE regions were representative of the common variation in Cebu Filipinos and that additional re-sequencing of these regions would not be required to detect common SNPs in Cebu Filipino samples. While future SNP selection in genome regions that have not been re-

sequenced will be based on less complete SNP identification, the Asian HapMap panels will likely either include or tag most of the common SNPs present in Cebu Filipino samples.

Measures of LD, gene density, and haplotype blocks vary across the genome [10,36], and the HapMap ENCODE regions analyzed represent a range of these and other characteristics [36], suggesting that our results may apply, on average, across the genome. The strong correlations observed between Cebu Filipino samples and HapMap Asian panels are broadly consistent with other assessments of tagging transferability outside the HapMap ENCODE regions [7-17,24].

Our results are consistent with previous studies that compared the Asian HapMap panels to other Eastern Asian samples. Studies that examined many populations worldwide found Asian and Oceania populations to be most similar to the Asian HapMap panel tested [13,23]. Two studies have investigated the tagging transferability between the HapMap CHB, JPT, and CHB+JPT with sample sets from Thailand and from Korea [24,25]. A combination of tag SNPs from CHB+JPT best captured the LD structure of the Thais, while SNP selection based on JPT was most transferable to the Korean samples. In comparison, our results suggest that CHB samples and the combined CHB+JPT samples are most similar to Cebu Filipino samples. Although our results do not necessarily reflect the patterns of genetic variability across the Philippines, our findings will be useful for the future design and analysis of genetic studies in the Cebu Filipino population.

Table 2.1: Number of SNPs successfully genotyped by population and region

Region	40 kb Positions NCBI Build 35 (hg17)	Total number of SNPs (MAF > 0, MAF ≥ .05)					
		Cebu Filipino	CHB	JPT	CHB+JPT	CEU	YRI
ENr112	Chr2:51800356..51840356	76 (52,46)	75 (46,45)	75 (46,45)	75 (47,45)	76 (73,46)	75 (58,51)
ENr131	Chr2:234503825..234543825	96 (91,83)	81 (75,73)	81 (77,72)	81 (77,72)	93 (86,77)	85 (74,70)
ENr113	Chr4:118834259..118874259	99 (84,80)	92 (77,74)	92 (77,71)	92 (77,71)	96 (91,78)	92 (78,64)
ENm010	Chr7:26960761..27000761	40 (27,25)	39 (33,24)	39 (30,23)	39 (36,24)	38 (29,23)	36 (23,20)
ENm013	Chr7:89658340..89698340	64 (42,35)	60 (41,38)	60 (41,37)	60 (42,38)	63 (58,43)	61 (41,34)
ENm014	Chr7:125901178..125941178	53 (41,37)	49 (39,34)	49 (38,33)	49 (41,34)	53 (46,39)	49 (30,30)
ENr321	Chr8:119112221..119152221	57 (47,36)	55 (47,42)	55 (45,36)	55 (48,42)	56 (43,36)	53 (41,41)
ENr232	Chr9:128994856..129034856	34 (30,29)	33 (30,28)	33 (29,29)	33 (30,29)	34 (24,21)	33 (25,21)
ENr123	Chr12:38856477..38896477	52 (40,27)	50 (36,24)	50 (39,23)	50 (42,23)	50 (42,29)	47 (34,26)
ENr213	Chr18:23949232..23989232	56 (47,30)	55 (25,17)	55 (42,22)	55 (45,18)	54 (50,39)	55 (42,34)
Sum across all regions		627 (501,428)	589 (449,399)	589 (464,391)	589 (485,396)	616 (542,431)	616 (542,431)

Table 2.2: Spearman's correlation coefficients of all pairwise r^2 estimates between HapMap Asian panels and Cebu Filipino samples

Region	CHB	JPT	CHB+JPT
ENr112	.886	.896	.897
ENr131	.707	.710	.761
ENr113	.931	.913	.937
ENm010	.857	.874	.886
ENm013	.897	.950	.929
ENm014	.894	.832	.830
ENr321	.845	.812	.844
ENr232	.848	.630	.768
ENr123	.668	.705	.681
ENr213	.755	.766	.701
Average	.877	.870	.888

Table 2.3: Coverage of the Cebu Filipino samples by tag SNPs selected from Asian HapMap panels

r^2 threshold for tag SNP selection	Number of tag SNPs selected	Captured* SNPs		Un-captured* SNPs, MAF < .10		Un-captured* SNPs, MAF ≥ .10	
		Percent[#]	Mean maximum $r^{2\$}$	Percent[#]	Mean maximum $r^{2\$}$	Percent[#]	Mean maximum $r^{2\$}$
CHB							
.80	121	82.4	.877	4.0	.142	13.6	.649
.85	134	85.7	.878	5.4	.113	10.1	.679
.90	152	91.5	.917	4.0	.141	4.7	.702
.95	179	92.0	.923	4.0	.140	4.0	.680
JPT							
.80	118	83.1	.877	4.5	.143	12.4	.642
.85	125	83.1	.876	7.5	.352	9.4	.632
.90	137	90.6	.896	4.5	.147	4.9	.610
.95	159	92.5	.916	4.2	.123	3.3	.518
CHB+ JPT							
.80	125	82.4	.882	4.2	.160	13.4	.656
.85	132	88.5	.900	4.2	.162	7.3	.651
.90	144	89.0	.906	4.2	.166	6.8	.626
.95	170	93.0	.923	4.0	.140	3.1	.589

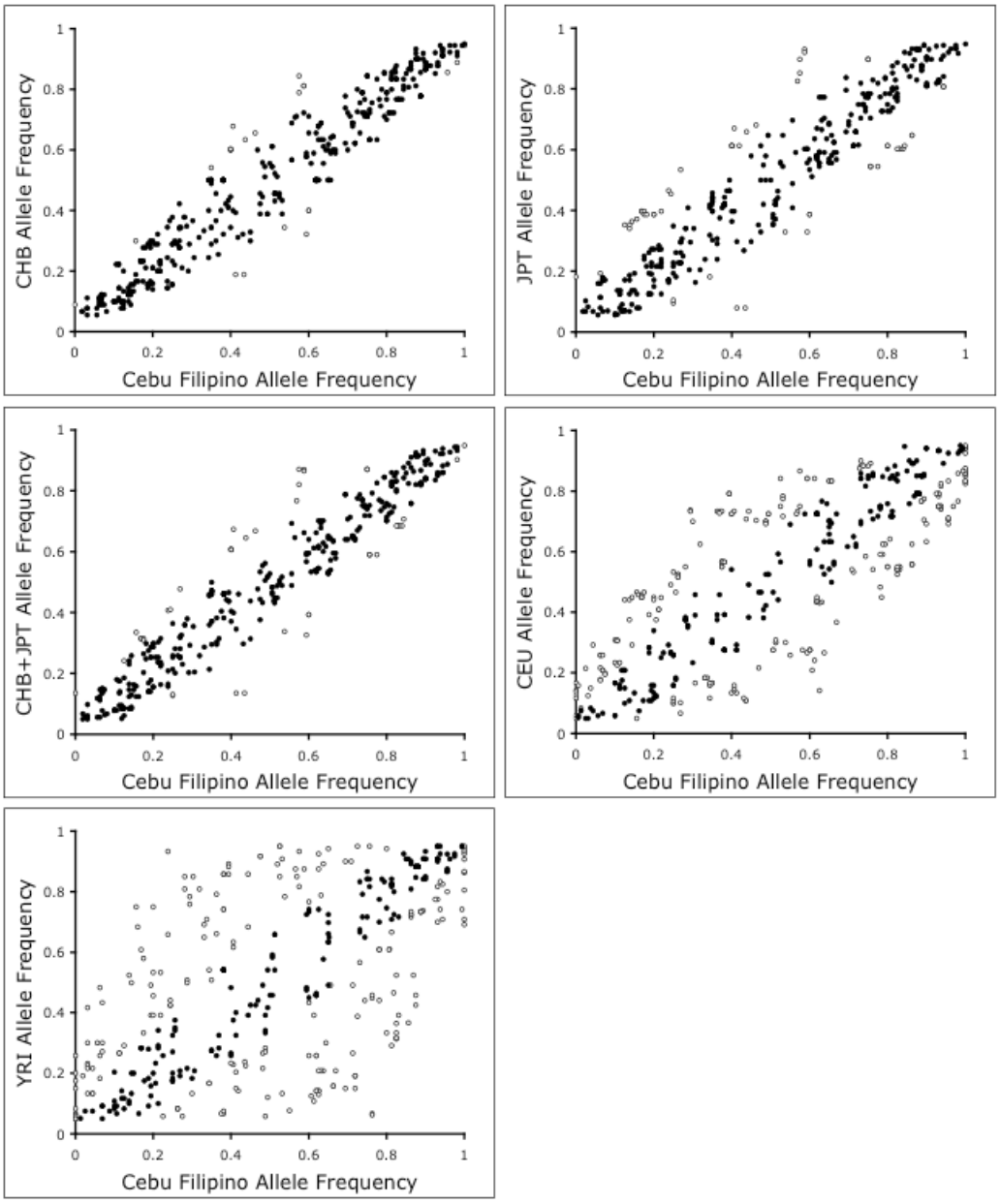
Only SNPs with an allele frequency ≥ .05 in the HapMap panel and the Cebu Filipino samples were analyzed.

*A SNP is considered captured it if exhibited a pairwise r^2 estimate ≥ .80 with at least one tag SNP.

[#]Percent coverage is defined as the number of SNPs in the Cebu Filipino samples captured by a tag SNP divided by the total number of SNPs.

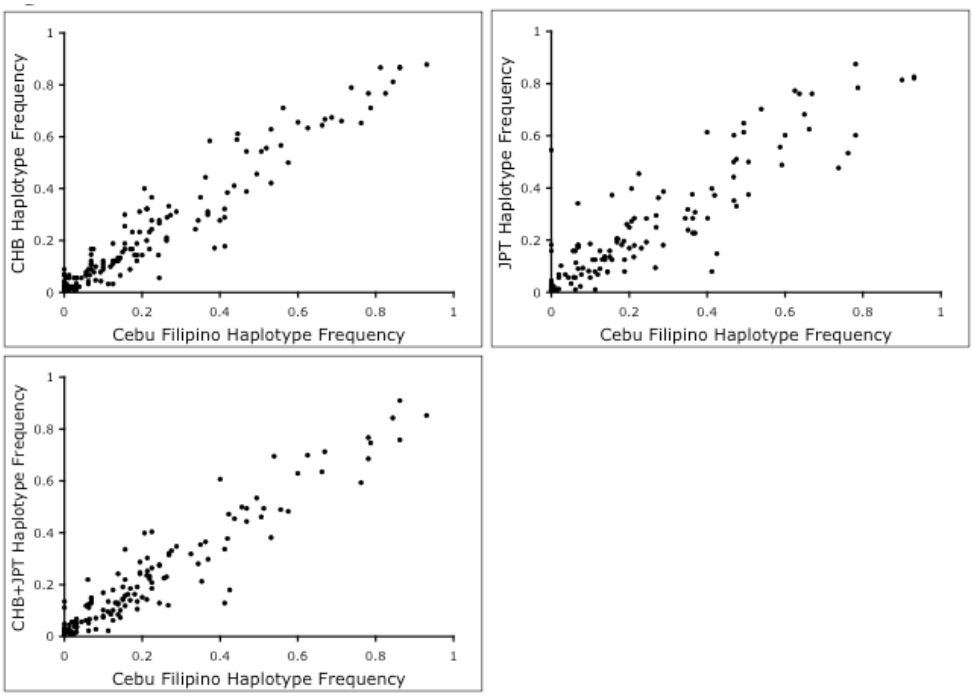
[$]Mean maximum r^2 is the average of the maximum pairwise r^2 estimates obtained between each SNP within a region and a tag SNP.

Figure 2.1: Comparison of allele frequency estimates between Cebu Filipino samples and HapMap samples for SNPs with MAF ≥ .05 in the HapMap sample. Open symbols indicate SNPs with significantly different allele frequencies at a Fisher's exact P value < .01



.

Figure 2.2: Comparison of haplotype frequency estimates between Cebu Filipino samples and Asian HapMap samples for SNPs with MAF ≥ .05 in the HapMap sample and haplotype frequency estimates > .01 in the HapMap sample



.

REFERENCES

1. Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, et al. (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet 74: 106-120.

2. Service S, DeYoung J, Karayiorgou M, Roos JL, Pretorious H, et al. (2006) Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. Nat Genet 38: 556-560.

3. Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, et al. (2001) Linkage disequilibrium in the human genome. Nature 411: 199-204.

4. Sawyer SL, Mukherjee N, Pakstis AJ, Feuk L, Kidd JR, et al. (2005) Linkage disequilibrium patterns vary substantially among populations. Eur J Hum Genet 13: 677-686.

5. Plagnol V, Wall JD (2006) Possible ancestral structure in human populations. PLoS Genet 2: e105.

6. International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437: 1299-1320.

7. Weale ME, Depondt C, Macdonald SJ, Smith A, Lai PS, et al. (2003) Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene SCN1A: implications for linkage-disequilibrium gene mapping. Am J Hum Genet 73: 551-565.

8. Nejentsev S, Godfrey L, Snook H, Rance H, Nutland S, et al. (2004) Comparative high-resolution analysis of linkage disequilibrium and tag single nucleotide polymorphisms between populations in the vitamin D receptor gene. Hum Mol Genet 13: 1633-1639.

9. Evans DM, Cardon LR (2005) A comparison of linkage disequilibrium patterns and estimated population recombination rates across multiple populations. Am J Hum Genet 76: 681-687.

10. Ke X, Miretti MM, Broxholme J, Hunt S, Beck S, et al. (2005) A comparison of tagging methods and their tagging space. Hum Mol Genet 14: 2757-2767.

11. Mueller JC, Lohmussaar E, Magi R, Remm M, Bettecken T, et al. (2005) Linkage disequilibrium patterns and tagSNP transferability among European populations. Am J Hum Genet 76: 387-398.

12. Ramirez-Soriano A, Lao O, Soldevila M, Calafell F, Bertranpetit J, et al. (2005) Haplotype tagging efficiency in worldwide populations in CTLA4 gene. Genes Immun 6: 646-657.

13. Gonzalez-Neira A, Ke X, Lao O, Calafell F, Navarro A, et al. (2006) The portability of tagSNPs across populations: a worldwide survey. Genome Res 16: 323-330.

14. Huang W, He Y, Wang H, Wang Y, Liu Y, et al. (2006) Linkage disequilibrium sharing and haplotype-tagged SNP portability between populations. Proc Natl Acad Sci U S A 103: 1418-1421.

15. Montpetit A, Nelis M, Laflamme P, Magi R, Ke X, et al. (2006) An evaluation of the performance of tag SNPs derived from HapMap in a Caucasian population. PLoS Genet 2: e27.

16. Ribas G, Gonzalez-Neira A, Salas A, Milne RL, Vega A, et al. (2006) Evaluating HapMap SNP data transferability in a large-scale genotyping project involving 175 cancer-associated genes. Hum Genet 118: 669-679.

17. Willer CJ, Scott LJ, Bonnycastle LL, Jackson AU, Chines P, et al. (2006) Tag SNP selection for Finnish individuals based on the CEPH Utah HapMap database. Genet Epidemiol 30: 180-190.

18. Eyheramendy S, Marchini J, McVean G, Myers S, Donnelly P (2007) A model-based approach to capture genetic variation for future association studies. Genome Res 17: 88-95.

19. Paschou P, Mahoney MW, Javed A, Kidd JR, Pakstis AJ, et al. (2007) Intra- and interpopulation genotype reconstruction from tagging SNPs. Genome Res 17: 96-107.

20. Beaty TH, Fallin MD, Hetmanski JB, McIntosh I, Chong SS, et al. (2005) Haplotype diversity in 11 candidate genes across four populations. Genetics 171: 259-267.

21. Lim J, Kim YJ, Yoon Y, Kim SO, Kang H, et al. (2006) Comparative study of the linkage disequilibrium of an ENCODE region, chromosome 7p15, in Korean, Japanese, and Han Chinese samples. Genomics 87: 392-398.

22. de Bakker PI, Burtt NP, Graham RR, Guiducci C, Yelensky R, et al. (2006) Transferability of tag SNPs in genetic association studies in multiple populations. Nat Genet 38: 1298-1303.

23. Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, et al. (2006) A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nat Genet 38: 1251-1260.

24. Mahasirimongkol S, Chantratita W, Promso S, Pasomsab E, Jinawath N, et al. (2006) Similarity of the allele frequency and linkage disequilibrium pattern of single nucleotide polymorphisms in drug-related gene loci between Thai and northern East Asian populations: implications for tagging SNP selection in Thais. J Hum Genet 51: 896-904.

25. Yoo YK, Ke X, Hong S, Jang HY, Park K, et al. (2006) Fine-scale map of encyclopedia of DNA elements regions in the Korean population. Genetics 174: 491-497.

26. Cebu Study Team (L Adair, JS Akin, R Black, J Briscoe, DK Guilkey, BM Popkin and WF Flieger), (1991) Underlying and proximate determinants of child health: the Cebu Longitudinal Health and Nutrition Study. Am J Epidemiol 133: 185-201.

27. Adair LS (2004) Dramatic rise in overweight and obesity in adult Filipino women and risk of hypertension. Obesity Res 12: 1335-1341.

28. Encode Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306: 636-640.

29. Gunderson KL, Kruglyak S, Graige MS, Garcia F, Kermani BG, et al. (2004) Decoding randomly ordered DNA arrays. Genome Res 14: 870-877.

30. Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21: 263-265.

31. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, et al. (2002) The structure of haplotype blocks in the human genome. Science 296: 2225-2229.

32. de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, et al. (2005) Efficiency and power in genetic association studies. Nat Genet 37: 1217-1223.

33. Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132: 365-386.

34. Schulze TG, Zhang K, Chen YS, Akula N, Sun F, et al. (2004) Defining haplotype blocks and tag single-nucleotide polymorphisms in the human genome. Hum Mol Genet 13: 335-342

35. Jin L, Su B (2000) Natives or immigrants: modern human origin in east Asia. Nat Rev Genet 1: 126-133.

36. De La Vega FM, Isaac H, Collins A, Scafe CR, Halldorsson BV, et al. (2005) The linkage disequilibrium maps of three human chromosomes across four populations reflect their demographic history and a common underlying recombination pattern. Genome Res 15: 454-462.

CHAPTER III. ASSOCIATION OF *FTO* AND *ADRB3* WITH OBESITY-RELATED TRAITS IN THE CEBU LONGITUDINAL HEALTH AND NUTRITION SURVEY (CLHNS) COHORT

A version of this work was previously published as:

Amanda F. Marvelle, Leslie A. Lange, Li Qin, Linda S. Adair, Karen L. Mohlke.

Diabetes. (2008) Jul;57(7):1987-91.

CHAPTER III CREDITS

I would like to thank Sandra German at the Office of Population Studies (OPS) in Cebu Philippines for blood sample collection and processing, under the direction of Dr. Christopher Kuzawa of Northwestern University; and the entire staff of OPS for their long term work on the CLHNS. I also thank Amy Perou of the BioSpecimen Processing facility and Jason Luo of the Mammalian Genotyping Core at University of North Carolina at Chapel Hill. This work was supported by NIH R01 DK78150. Cebu Filipino data collection was supported by TW05596, and specimen processing and genotyping was supported by pilot funds from NIH grants RR20649 (Interdisciplinary Obesity Center), ES10126 (Project 7-2004-E of the Center for Environmental Health and Susceptibility), and DK56350 (Clinical Nutrition Research Center). A.F.M. was supported by an Integrative Vascular Biology Fellowship, NIH grant HL69768.

ABSTRACT

OBJECTIVE: The underlying genetic component of obesity-related traits is not well understood, and there is limited evidence to support genetic association shared across multiple studies, populations, and environmental contexts. The present study investigated the association between candidate variants and obesity-related traits in a sample of 1,886 adult Filipino women from the Cebu Longitudinal Health and Nutrition Survey (CLHNS) cohort. RESEARCH DESIGN AND METHODS: We selected and genotyped 19 single nucleotide polymorphisms in 10 genes (*ADRB2, ADRB3, FTO, GNB3, INSIG2, LEPR, PPARG, TNF, UCP2,* and *UCP3*) that had been previously reported to be associated with an obesity-related quantitative trait. RESULTS: We observed evidence for association of the A allele of rs9939609 (*FTO* intron 1) with increased BMI (P = .0072 before multiple test correction), baseline BMI (P = .0015), longitudinal BMI based on eight surveys from 1983 to 2005 (P = .000029), waist circumference (P = .0094), and weight (P = .021). The increase in average BMI was approximately .4 for each additional A allele. We also observed association of the *ADRB3* Trp64Arg variant with BMI, waist circumference, percent body fat, weight, fat mass, arm fat area, and arm muscle area (P < .05), although the direction of effect is inconsistent with the majority of previous reports. CONCLUSIONS: Our study confirms that *FTO* is a common obesity susceptibility gene in Filipinos, with an effect size similar to that seen in samples of European origin.

INTRODUCTION

Obesity is a worldwide epidemic, affecting individuals across all age groups, socioeconomic classes, and ethnicities. In addition, obesity is a risk factor for cardiovascular disease, type 2 diabetes, metabolic syndrome, hypertension, stroke, and some forms of cancer [1]. This complex and heterogeneous disorder arises from interactions between environment, behavior, and genetics.

Although many association studies have attempted to identify genetic variants that influence obesity, replication has been infrequent. As of 2005, 22 candidate genes contained a variant reported to be significantly associated ($P < .05$) with an obesity-related trait in at least five studies [2]. However, additional reports for these genes are inconsistent.

Recently, genome-wide association (GWA) studies have identified variants in additional genes. Single nucleotide polymorphism (SNP) rs7566605, near insulin-induced gene 2 (*INSIG2*), found to be associated with increased body mass index (BMI) [3], has not been consistently replicated [4-7]. Several variants in the fat mass and obesity associated (*FTO*) gene identified through two independent GWA studies [8,9] and a third study [10] were associated with BMI and risk of being overweight in children and adults in cohorts of Europeans, European-Americans, and Hispanic-Americans, but not in African Americans. Significant *FTO* association was also observed for hip circumference, waist circumference, and subcutaneous fat mass assessed using skinfolds [8,9]. Two studies observed *FTO* association with percentage of fat mass and dual energy X-ray absorptiometry derived fat mass in children [8,10]. *FTO* variants have the

most consistent replication across multiple populations to date, suggesting this locus is a likely risk factor for obesity.

In the current study, we examined 19 SNPs previously reported to be associated with obesity-related phenotypes for association with BMI, waist circumference, and percent body fat in 1,886 Filipino women from the Cebu Longitudinal Health and Nutrition Survey (CLHNS). In addition, for SNPs with initial evidence of association, we performed analysis with additional phenotypes and tested for interaction with diet and physical activity.

MATERIALS AND METHODS

Study subjects and traits

We evaluated 1,886 unrelated healthy Cebu Filipino female participants in the ongoing CLHNS [11], mothers of a 1983 to 1984 birth cohort. Trained field staff conducted in-home interviews and collected measurements and comprehensive environmental data (www.cpc.unc.edu/projects/cebu). We used non-pregnant data collected from surveys in 1983-84 ("baseline" = four months post-partum), 1984-85 (one year post-partum), 1985-86 (two years post-partum), 1991, 1994, 1998, 2002, and 2005. For 2005 cross-sectional traits, outcome and covariate measures from the 2002 survey were substituted for 16 women who were pregnant or missing data in 2005.

All outcome and covariate measures, except baseline BMI, were taken from the 2005 survey. Triceps and suprailiac skinfold thicknesses (TSF and SiSF) represent the mean of three consecutive Harpenden caliper measurements. Cross-sectional arm muscle area (AMA) and arm fat area (AFA) were calculated using mid arm circumference and

triceps skinfold thickness. Body density was calculated using the Durnin-Womersley sum of skinfold equation based on TSF and SiSF for adult women from 16 to 68 years of age [12], and percent body fat was derived from body density using the Siri equation [13]. Fat mass was calculated as the product of percent body fat and weight. Height was calculated as an average of eight measures across surveys from 1983-84 to 2005. Dietary intake was assessed by 24-hour dietary recall, with nutrient composition calculated from the Philippines Food Composition Table [14]. Physical activity level was categorized based on time use data, with a focus on occupational activity: each job was classified according to its metabolic equivalents based on field studies in Filipino women [15] and the Compendium of Physical Activity [16].

Informed consent was obtained from all individuals, and the study protocol was approved by the University of North Carolina Institutional Review Board for the Protection of Human Subjects. Genomic DNA was isolated from peripheral blood lymphocytes (Puregene, Gentra) by the University of North Carolina, Chapel Hill BioSpecimen Processing Facility.

SNP selection and genotyping methods

We reviewed genes that exhibited nine or more reports of association with an obesity phenotype as summarized by the 2004 obesity gene map [2]. SNPs within these genes with more than three positive reports of association and a minor allele frequency > .01 in the CHB HapMap samples were subsequently chosen to be genotyped. Variants in *FTO* and *INSIG2* identified through GWA studies were also genotyped [3,8].

Genotyping was performed using TaqMan allelic discrimination (Applied Biosystems, Foster City, CA). The genotype success rate for all SNPs was >98% and the discrepancy rate among duplicate samples was .1%.

Statistical analysis

Tests for consistency of genotype distributions with expected Hardy-Weinberg equilibrium proportions were calculated using Pearson's χ^2 statistic; only rs3856806 was inconsistent (P = .02). All SNPs were tested for association with three primary phenotypes: 2005 measures of BMI, waist circumference, and percent body fat. Analysis of covariance (ANCOVA) models were used to test for association between genotype and the continuously distributed outcomes. Logistic regression models were used for dichotomous outcomes. Only SNPs with evidence for association with at least one trait were tested for association with additional phenotypes. We also examined whether these SNPs modify the effect of established environmental risk factors of diet and physical activity level by performing tests of interaction between genotype and total caloric intake, estimated percent dietary fat and carbohydrates, and activity level. We also performed a longitudinal analysis incorporating all available BMI measurements for the up to eight measurements spanning 22 years using general linear mixed models.

Models were adjusted for age, household assets, natural log of income, number of total pregnancies as a categorical variable (1-4, 5-10, >10), and menopausal status; baseline BMI is not adjusted for menopausal status. Continuously distributed traits were transformed to satisfy the model assumption of normally distributed residuals, conditional on the covariates. The additive mode of inheritance assumption was used

unless fewer than 15 rare homozygotes existed; the dominant mode of inheritance assumption for the minor allele was used for SNPs rs4994, rs8179183, rs1801282, and rs1800629. The rs9939609 SNP in *FTO* was also analyzed under both the additive and dominant models for comparison to previous reports. Because of low linkage disequilibrium ($r^2 < .5$) between pairs of SNPs, Bonferroni adjustment was used to account for multiple SNPs.

RESULTS

Nineteen SNPs described previously to be associated with obesity-related phenotypes were tested for evidence of association with 2005 measures of BMI, waist circumference, and percent body fat in 1,886 unrelated non-pregnant Filipino women in the CLHNS cohort (Table 3.1). Two SNPs were statistically significantly associated ($P < .01$) with at least one trait before adjustment for multiple SNPs (Table 3.2). The A-allele of SNP rs9939609 (*FTO* intron 1) was associated with increased BMI ($P = .0072$) and waist circumference ($P = .0094$). The TT homozygote (Trp64) of SNP rs4994 (*ADRB3* Trp64Arg) was associated with increased BMI ($P = .00069$) and waist circumference ($P = .0013$). After Bonferroni correction for multiple SNPs, only rs4994 in *ADRB3* remained significant ($P < .002$), however, only the *FTO* association was consistent in magnitude and direction of effect with previous reports [8-10].

To further investigate the rs9939609 and rs4994 SNPs, we analyzed additional obesity-related phenotypes of baseline BMI (measured in 1983-84), weight, fat mass, SiSF, TSF, AFA, AMA, and height (Table 3.3). For *FTO* variant rs9939609, significant evidence of association was observed with baseline BMI ($P = .0015$) and weight ($P =$

.021). Marginally significant p-values (.05 < P < .10) were observed for fat mass (P = .055) and AMA (P = .084), with direction of estimated effects consistent with those seen for BMI and weight. For the *ADRB3* variant rs4994, significant evidence of association was observed for weight (P = .0011), fat mass (P = .0036), AFA (P = .016), and AMA (P = .0008), and marginally significant p-values were observed for TSF (P = .068), with the direction of estimated effects consistent with those observed for BMI and weight. Unlike the *FTO* variant, no evidence for association was observed with baseline BMI (P = .55).

We analyzed risk of being either overweight and obese (BMI $\geq$ 25 kg/m^2) or obese (BMI $\geq$ 30 kg/m^2) [17] both in 1983-84 and in 2005 (Table 3.4). Using these criteria, 793 and 178 women had a BMI $\geq$ 25 kg/m^2 or BMI $\geq$ 30 kg/m^2, respectively, in 2005, and 94 women had a BMI $\geq$ 25 kg/m^2 in 1983-84. The A-allele of rs9939609 was associated with increased risk of being overweight in 2005 (P = .0034) and in 1983-84 (P = .023). The TT homozygote of rs4994 was associated with increased risk of being overweight in 2005 (P = .0077) and 1983-84 (P = .044) and obese in 2005 (P = .023).

There was no evidence for interaction between genotype and either the 2005 dietary intake (total calories, estimated percent diet from fat, and estimated percent diet from carbohydrates) or physical activity measures (P > .05 for all tests).

A longitudinal analysis of BMI included an average of 7.3 (range 3 to 8) measurements per individual spanning 22 years. The global P value for the test of association with rs9939609 was .000029 (additive model, Fig. 3.1A) and .016 for rs4994 (Fig. 3.1B). The direction of the genotype-specific least-squares means at each time point was consistent with the cross-sectional analysis. The test of rs9939609 genotype-by-time interaction was significant (P = .047), with a slight increasing effect of genotype

over time. For rs4994, the test of genotype-by-time interaction was highly significant (P = .0065), with evidence for an increasing effect of genotype over time.

DISCUSSION

We evaluated 19 SNPs in a sample of adult Filipino women from the CLHNS cohort, confirmed the association of the A-allele of *FTO* variant rs9939609 with BMI and waist circumference, and observed evidence for an association with the TT homozygote of *ADRB3* rs4994 with BMI, waist circumference, and percent body fat. While only rs4994 reached significance after Bonferroni correction, the direction of effect was not consistent with the majority of previous reports [8-10]. The failure to replicate many of the SNP associations that have been reported previously may reflect environmental and genetic differences between the CLHNS cohort and previously studied populations, limited statistical power, and/or false positive results in the literature.

We also observed evidence for association between the Trp64 allele of rs4994 and increased weight, percent fat mass, AFA, AMA, and longitudinal BMI. However, we did not observe evidence for association with baseline BMI (measured at a time when few women were overweight). In contrast to our study, two meta-analyses with over 35 subgroups each, one in Japanese and one in multiple populations, reported that Arg64 carriers exhibited significantly higher mean BMI than Trp64 homozygotes [18,19]. The *ADRB3* receptor is more abundant and active in visceral adipose tissue than subcutaneous adipose [20]; however, we observed evidence for association between rs4994 and measures of both visceral and subcutaneous fat. The observation of significant

associations, but with opposite alleles associated with increased trait values across studies, suggests that these results should be interpreted with caution.

The *FTO* rs9939609 A-allele was also significantly associated with several obesity-related traits including longitudinal BMI, which reflects a relatively constant genotype effect over 22 years and strengthens the evidence that this locus influences BMI in this population. We observed an effect even at younger ages among women with smaller BMIs, consistent with previous reports in children [8,10]. We observed evidence for an association with waist circumference, but not an association with skinfold thicknesses, a measurement of subcutaneous adiposity, consistent with variation in *FTO* influencing central adiposity to a greater extent than subcutaneous fat.

The minor allele frequency (MAF) of the rs9939609 variant is .18 in the CLHNS samples, less common than observed in European populations (MAF = .45-.48). Recently, Li *et al.* [21] suggested that the lack of association of rs9939609 with obesity in a population of Han Chinese may be due to a decreased allele frequency. This is not consistent with our findings of significant association.

In summary, our results corroborate previous reports that a SNP within the first intron of *FTO* is associated with BMI. The *FTO* SNPs have the most consistent prior evidence for association with obesity-related traits reported to date, and our study replicates this evidence, both in direction and approximate magnitude, in a Filipino population, suggesting *FTO* may be important in many genetic backgrounds.

Table 3.1: Characteristics of 1,886 women in the Cebu Longitudinal Health and Nutrition Survey

	Mean (SD)
BMI (kg/m^2)	24.3 (4.4)
Waist circumference (cm)	81.1 (10.8)
Percent body fat	36.6 (5.4)
BMI (baseline 1983-84) (kg/m^2)	20.6 (2.4)
Arm fat area (mm^2)	9.6 (1.5)
Arm fat mass (mm^2)	60.0 (17.6)
Fat mass (kg)	20.6 (6.3)
Suprailiac skinfold (mm)	28.8 (10.1)
Triceps skinfold (mm)	23.8 (8.0)
Weight (kg)	55.1 (10.9)
Average height (cm)	150.4 (4.9)
Age (years)	48.4 (6.1)
Total number of pregnancies	6.5 (3.0)
Menopausal status (yes/no)	1162/724

All traits are measured from 2005 survey except where indicated. For women who were pregnant or missing data in 2005, measures from the 2002 survey were substituted. Baseline BMI was collected from postpartum surveys in 1983-1984 (see methods).

Table 3.2: Results of SNPs assessed for association in the CLHNS cohort with BMI, waist circumference, and percent body fat

Trait	SNP rs#	Gene	SNP Alias	Minor/ Major Allele	MAF	Minor Allele Homozygotes		Heterozygotes		Major Allele Homozygotes		p-value
						Mean (95% CI)	n	Mean (95% CI)	n	Mean (95% CI)	n	
BMI	rs1042711	*ADRB2*	Cys19Arg	C/T	.147	24.3 (23.0, 25.6)	40	23.8 (23.4, 24.2)	437	24.1 (23.8, 24.4)	1287	.38
(kg/m^2)	rs1042713	*ADRB2*	Gly16Arg	G/A	.463	23.9 (23.5, 24.3)	389	24.1 (23.8, 24.4)	857	24.0 (23.6, 24.4)	518	.67
	rs1042714	*ADRB2*	Gln27Glu	G/C	.146	24.3 (23.0, 25.6)	39	23.8 (23.4, 24.2)	440	24.1 (23.8, 24.4)	1291	.41
	rs4994*	*ADRB3*	Trp64Arg	C/T	.085	22.3 (19.9, 24.7)	11	23.3 (22.8, 23.9)	278	24.2 (23.9, 24.4)	1476	**.0011**
	rs9939609	*FTO*	Intron 1	A/T	.175	24.7 (23.7, 25.7)	64	24.5 (24.1, 24.8)	508	23.9 (23.6, 24.1)	1249	**.0072**
	rs6489738	*GNB3*	Ser275Ser	C/T	.377	24.0 (23.4, 24.5)	241	24.0 (23.7, 24.3)	828	24.1 (23.7, 24.4)	667	.68
	rs7566605	*INSIG2*	-10kb C/G	C/G	.450	24.2 (23.7, 24.6)	355	24.0 (23.7, 24.3)	872	23.9 (23.5, 24.3)	532	.29
	rs1137100	*LEPR*	Lys109Arg	A/G	.224	24.5 (23.7, 25.3)	96	23.6 (23.3, 24.0)	598	24.2 (23.9, 24.5)	1073	.058
	rs1137101	*LEPR*	Gln223Arg	A/G	.143	24.5 (23.1, 25.8)	36	23.8 (23.4, 24.2)	426	24.1 (23.8, 24.4)	1280	.34
	rs3790419	*LEPR*	Ser343Ser	G/A	.108	24.9 (23.1, 26.7)	20	23.8 (23.4, 24.3)	336	24.1 (23.8, 24.4)	1389	.39
	rs8179183*	*LEPR*	Lys656Asn	C/G	.029	22.9 (17.2, 28.6)	2	23.7 (22.9, 24.6)	96	24.0 (23.8, 24.3)	1624	.49
	rs1805096	*LEPR*	Pro1019Pro	C/T	.147	24.6 (23.3, 25.8)	44	23.7 (23.3, 24.1)	432	24.1 (23.9, 24.4)	1287	.14
	rs1801282*	*PPARG*	Pro12Ala	G/C	.048	21.0 (16.9, 25.0)	4	24.2 (23.6, 24.9)	167	24.1 (23.8, 24.3)	1657	.74
	rs3856806	*PPARG*	1431 C/T	T/C	.219	23.4 (22.6, 24.3)	101	23.9 (23.5, 24.3)	570	24.1 (23.9, 24.4)	1091	.053
	rs1799724	*TNF*	-857 C/T	T/C	.127	23.4 (21.8, 24.9)	26	24.2 (23.7, 24.6)	396	24.0 (23.7, 24.3)	1341	.81
	rs1800629*	*TNF*	-308 G/A	A/G	.040	26.1 (20.4, 31.8)	2	24.2 (23.5, 24.9)	137	24.0 (23.8, 24.3)	1612	.74
	rs659366	*UCP2*	-866 G/A	A/G	.297	24.2 (23.6, 24.9)	163	23.9 (23.5, 24.2)	718	24.2 (23.8, 24.5)	875	.5
	rs660339	*UCP2*	Ala55Val	T/C	.299	24.1 (23.5, 24.8)	167	23.8 (23.5, 24.2)	721	24.2 (23.9, 24.5)	876	.32
	rs1800849	*UCP3*	-55 C/T	T/C	.226	24.1 (23.3, 25.0)	91	24.1 (23.7, 24.4)	608	24.0 (23.7, 24.3)	1050	.96
Waist	rs1042711	*ADRB2*	Cys19Arg	C/T	.147	80.7 (77.5, 83.9)	40	79.7 (78.7, 80.8)	437	80.7 (80.1, 81.4)	1286	.20
circ.	rs1042713	*ADRB2*	Gly16Arg	G/A	.463	80.2 (79.1, 81.3)	389	80.5 (79.7, 81.3)	856	80.7 (79.7, 81.7)	518	.42
(cm)	rs1042714	*ADRB2*	Gln27Glu	G/C	.146	80.6 (77.3, 83.9)	39	79.8 (78.8, 80.9)	440	80.7 (80.1, 81.4)	1290	.22
	rs4994*	*ADRB3*	Trp64Arg	C/T	.085	75.2 (69.1, 81.5)	11	79.1 (77.8, 80.3)	278	80.8 (80.1, 81.4)	1475	**.0026**
	rs9939609	*FTO*	Intron 1	A/T	.174	81.1 (78.5, 83.7)	64	81.7 (80.7, 82.7)	507	80.1 (79.4, 80.7)	1249	**.0094**
	rs6489738	*GNB3*	Ser275Ser	C/T	.377	80.3 (79.0, 81.7)	241	80.7 (79.9, 81.5)	827	80.5 (79.6, 81.4)	667	.87
	rs7566605	*INSIG2*	-10kb C/G	C/G	.450	80.6 (79.5, 81.8)	355	80.5 (79.7, 81.3)	871	80.4 (79.5, 81.4)	532	.83
	rs1137100	*LEPR*	Lys109Arg	A/G	.224	81.3 (79.2, 83.4)	96	79.7 (78.8, 80.6)	598	80.9 (80.1, 81.6)	1072	.20
	rs1137101	*LEPR*	Gln223Arg	A/G	.143	81.3 (77.9, 84.7)	36	80.2 (79.1, 81.2)	426	80.5 (79.8, 81.2)	1279	.73
	rs3790419	*LEPR*	Ser343Ser	G/A	.108	80.8 (76.3, 85.4)	20	80.0 (78.8, 81.2)	336	80.7 (80.0, 81.3)	1388	.36
	rs8179183*	*LEPR*	Lys656Asn	C/G	.029	84.6 (70.2, 99.0)	2	80.1 (78.0, 82.2)	96	80.5 (79.8, 81.1)	1623	.83
	rs1805096	*LEPR*	Pro1019Pro	C/T	.148	81.1 (78.0, 84.2)	44	80.2 (79.1, 81.2)	432	80.6 (79.9, 81.3)	1286	.53
	rs1801282*	*PPARG*	Pro12Ala	G/C	.048	73.0 (62.8, 83.1)	4	80.7 (79.1, 82.3)	167	80.6 (80.0, 81.3)	1656	.92
	rs3856806	*PPARG*	1431 C/T	T/C	.219	78.8 (76.7, 80.8)	101	80.3 (79.3, 81.2)	570	80.8 (80.1, 81.5)	1090	.059
	rs1799724	*TNF*	-857 C/T	T/C	.127	77.3 (73.3, 81.3)	26	80.9 (79.8, 82.0)	396	80.4 (79.8, 81.1)	1340	.88
	rs1800629*	*TNF*	-308 G/A	A/G	.040	85.0 (70.6, 99.4)	2	81.5 (79.7, 83.3)	137	80.5 (79.8, 81.1)	1611	.28
	rs659366	*UCP2*	-866 G/A	A/G	.297	80.7 (79.1, 82.3)	163	80.5 (79.6, 81.3)	718	80.6 (79.8, 81.3)	874	.99

	rs660339	*UCP2*	Ala55Val	T/C	.299	80.5 (78.9, 82.1)	167	80.4 (79.5, 81.2)	721	80.6 (79.8, 81.4)	875	.68
	rs1800849	*UCP3*	-55 C/T	T/C	.226	80.6 (78.4, 82.8)	91	80.8 (79.9, 81.7)	608	80.4 (79.6, 81.1)	1049	.58
Percent	rs1042711	*ADRB2*	Cys19Arg	C/T	.146	35.9 (34.4, 37.5)	40	36.0 (35.5, 36.5)	434	36.4 (36.1, 36.7)	1285	.17
body fat	rs1042713	*ADRB2*	Gly16Arg	G/A	.463	35.8 (35.2, 36.3)	388	36.5 (36.1, 36.8)	854	36.3 (35.8, 36.8)	517	.19
	rs1042714	*ADRB2*	Gln27Glu	G/C	.146	35.9 (34.4, 37.5)	39	36.0 (35.5, 36.5)	437	36.4 (36.1, 36.7)	1289	.18
	rs4994*	*ADRB3*	Trp64Arg	C/T	.085	35.6 (32.7, 38.5)	11	35.7 (35.1, 36.3)	278	36.4 (36.1, 36.7)	1471	.050
	rs9939609	*FTO*	Intron 1	A/T	.175	36.8 (35.6, 38.0)	64	36.4 (35.9, 36.9)	508	36.3 (36.0, 36.6)	1249	.43
	rs6489738	*GNB3*	Ser275Ser	C/T	.378	36.3 (35.6, 36.9)	241	36.2 (35.8, 36.6)	826	36.4 (36.0, 36.8)	664	.64
	rs7566605	*INSIG2*	-10kb C/G	C/G	.450	36.3 (35.8, 36.9)	354	36.3 (35.9, 36.7)	870	36.2 (35.7, 36.7)	530	.60
	rs1137100	*LEPR*	Lys109Arg	A/G	.224	36.7 (35.7, 37.7)	96	35.9 (35.5, 36.4)	597	36.4 (36.1, 36.8)	1069	.49
	rs1137101	*LEPR*	Gln223Arg	A/G	.143	36.5 (34.9, 38.1)	36	36.2 (35.7, 36.8)	425	36.3 (35.9, 36.6)	1276	.92
	rs3790419	*LEPR*	Ser343Ser	G/A	.108	37.4 (35.2, 39.5)	20	35.9 (35.3, 36.5)	335	36.4 (36.1, 36.7)	1385	.54
	rs8179183*	*LEPR*	Lys656Asn	C/G	.029	36.9 (30.0, 43.8)	2	36.0 (35.0, 37.0)	96	36.3 (35.9, 36.6)	1619	.68
	rs1805096	*LEPR*	Pro1019Pro	C/T	.147	36.6 (35.2, 38.1)	44	36.1 (35.6, 36.6)	430	36.4 (36.0, 36.7)	1284	.62
	rs1801282*	*PPARG*	Pro12Ala	G/C	.048	35.4 (30.6, 40.3)	4	36.7 (35.9, 37.4)	167	36.3 (36.0, 36.6)	1648	.55
	rs3856806	*PPARG*	1431 C/T	T/C	.220	36.0 (35.1, 37.0)	101	36.3 (35.9, 36.8)	570	36.3 (35.9, 36.6)	1086	.98
	rs1799724	*TNF*	-857 C/T	T/C	.127	36.2 (34.3, 38.1)	26	36.2 (35.7, 36.8)	395	36.3 (36.0, 36.6)	1337	.80
	rs1800629*	*TNF*	-308 G/A	A/G	.04	38.3 (31.4, 45.2)	2	36.1 (35.3, 37.0)	137	36.3 (36.0, 36.6)	1608	.62
	rs659366	*UCP2*	-866 G/A	A/G	.298	36.4 (35.7, 37.2)	163	36.2 (35.8, 36.6)	716	36.3 (36.0, 36.7)	872	.92
	rs660339	*UCP2*	Ala55Val	T/C	.299	36.3 (35.6, 37.1)	167	36.1 (35.7, 36.5)	719	36.4 (36.0, 36.8)	873	.59
	rs1800849	*UCP3*	-55 C/T	T/C	.226	36.1 (35.1, 37.1)	91	36.2 (35.8, 36.7)	607	36.3 (36.0, 36.7)	1046	.60

Untransformed means are reported. Models were adjusted for age, assets, number of pregnancies, income, and menopausal status. *All tests were performed under the additive model except for SNPs with fewer than 15 minor allele homozygotes, in which case a dominant mode of inheritance assumption for the minor allele was used.

Table 3.3: Association of *FTO* and *ADRB3* SNPs with obesity-related traits

	FTO rs9939609			Additive	Dominant	*ADRB3* rs4994		Dominant
	TT	TA	AA	p-value	p-value	TT	TC/CC	p-value
BMI (kg/m^2)	23.9 (23.6, 24.1)	24.5 (24.1, 24.8)	24.7 (23.7, 25.7)	**.0072**	**.0080**	24.2 (23.9, 24.4)	23.3 (22.8, 23.8)	**.0011**
Waist circumference (cm)	80.1 (79.4, 80.7)	81.7 (80.7, 82.7)	81.1 (78.5, 83.7)	**.0094**	**.0040**	80.8 (80.1, 81.4)	78.9 (77.7, 80.2)	**.0026**
Percent body fat	36.3 (36.0, 36.6)	36.4 (35.9, 36.9)	36.8 (35.6, 38.0)	.43	.47	36.4 (36.1, 36.7)	35.7 (35.1, 36.3)	**.0499**
Baseline BMI (kg/m^2)	20.5(20.3, 20.6)	20.9 (20.7, 21.1)	21.0 (20.4, 21.6)	**.0015**	**.0013**	20.6 (20.5, 20.8)	20.5 (20.3, 20.8)	.55
Weight (kg)	54.2 (53.5, 54.8)	55.3 (54.3, 56.2)	55.9 (53.4, 58.4)	**.021**	**.024**	54.8 (54.1, 55.4)	52.6 (51.4, 53.9)	**.0011**
Fat mass (kg)	20.1 (19.7, 20.5)	20.6 (20.1, 21.2)	20.9 (19.5, 22.3)	.055	.06	20.4 (20.0, 20.7)	19.3 (18.6, 20.0)	**.0036**
Suprailiac skinfold thickness (mm)	28.3 (27.6, 28.9)	28.8 (27.9, 29.7)	28.5 (26.2, 30.9)	.37	.31	28.4 (27.8, 29.0)	27.4 (26.2, 28.5)	.104
Triceps skinfold thickness (mm)	23.5 (23.0, 24.0)	23.6 (22.9, 24.3)	24.0 (22.2, 25.8)	.64	.87	23.6 (23.1, 24.0)	22.8 (21.9, 23.7)	.0682
Arm fat area (mm^2)	9.5 (9.4, 9.6)	9.5 (9.4, 9.7)	9.6 (9.3, 10.0)	.33	.45	9.5 (9.4, 9.6)	9.3 (9.1, 9.5)	**.0157**
Arm muscle area (mm^2)	58.7 (57.6, 59.8)	60.2 (58.6, 61.7)	61.5 (57.4, 65.6)	.084	.11	59.7 (58.6, 60.7)	56.1 (54.1, 58.1)	**.0008**
Height (cm)	150.4 (150.1, 150.8)	150.2 (149.7, 150.6)	150.4 (149.2, 151.6)	.42	.32	150.4 (150.1, 150.7)	150.0 (149.4, 150.6)	.21

Reported values are untransformed means (95% confidence interval). All data except baseline BMI were collected in the 2005 survey. For women who were pregnant or missing data in 2005, measures from the 2002 survey were substituted. Baseline BMI and covariates were collected from postpartum surveys in 1983-1984. Models were adjusted for age, assets, total number of past pregnancies, income, and menopausal status (except baseline BMI).

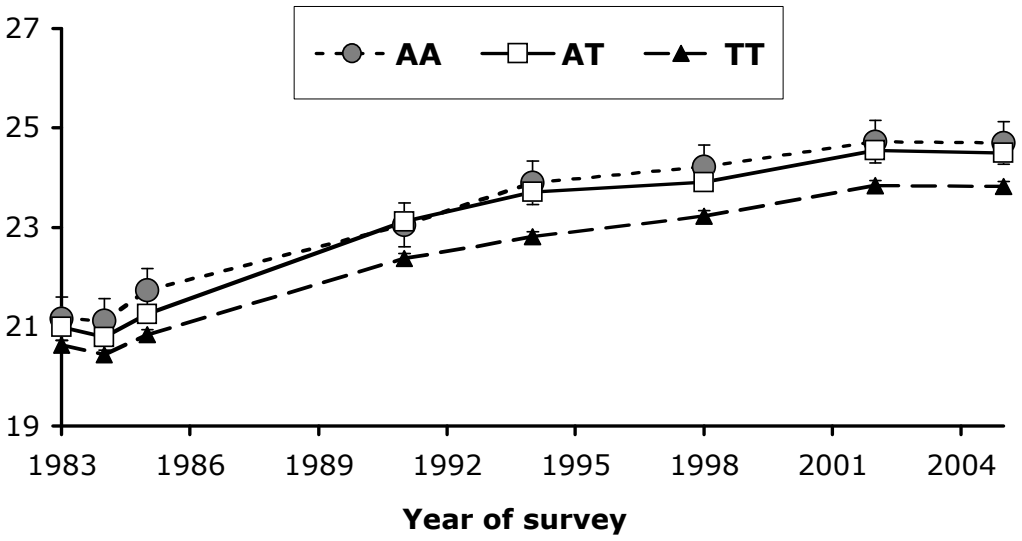
Table 3.4: Association of *FTO* and *ADRB3* SNPs with overweight and obesity status

	FTO rs9939609		*ADRB3* rs4994	
	Odds Ratio (95% CI)	p-value	Odds Ratio (95% CI)	p-value
2005 overweight and obese (BMI ≥ 25 kg/m^2)	1.30 (1.09, 1.55)	**.0034**	1.33 (1.07, 1.63)	**.0077**
2005 obese (BMI ≥ 30 kg/m^2)	1.31 (1.00, 1.72)	.054	1.46 (1.05, 2.02)	**.023**
1983-84 overweight and obese (BMI ≥ 25 kg/m^2)	1.50 (1.06, 2.12)	**.023**	1.27 (1.01, 1.61)	**.044**

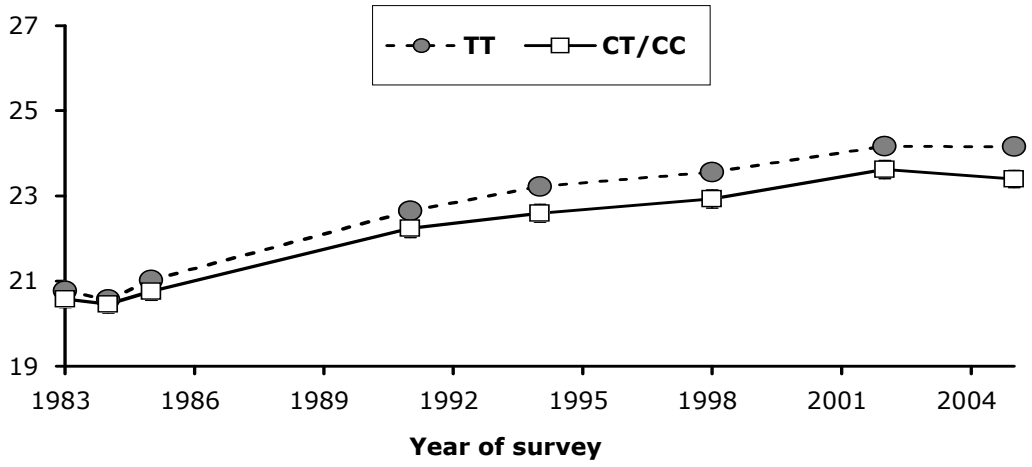
1983-84 obese (BMI≥ 30 kg/m2) is not reported because only 2 people were observed with BMI≥30 kg/m2. Models were adjusted for age, assets, total number of past pregnancies, income, and menopausal status (except 1983-84 model).

Figure 3.1: Longitudinal analysis of BMI using measurements across eight surveys from 1983-84 to 2005 of A) *FTO* rs9939609 (p-value .000029) and B) *ADRB3* rs4994 (p-value .016). BMI is reported as the least-squares means at each time point.

A



B



REFERENCES

1. Yang W, Kelly T, He J (2007) Genetic epidemiology of obesity. Epidemiol Rev 29: 49-61.

2. Rankinen T, Zuberi A, Chagnon YC, Weisnagel SJ, Argyropoulos G, et al. (2006) The human obesity gene map: the 2005 update. Obesity (Silver Spring, Md) 14: 529-644.

3. Herbert A, Gerry NP, McQueen MB, Heid IM, Pfeufer A, et al. (2006) A common genetic variant is associated with adult and childhood obesity. Science (New York, NY) 312: 279-283.

4. Lyon HN, Emilsson V, Hinney A, Heid IM, Lasky-Su J, et al. (2007) The association of a SNP upstream of *INSIG2* with body mass index is reproduced in several but not all cohorts. PLoS Genet 3: e61.

5. Smith AJ, Cooper JA, Li LK, Humphries SE (2007) *INSIG2* gene polymorphism is not associated with obesity in Caucasian, Afro-Caribbean and Indian subjects. Int J Obesity (2005) 31: 1753-1755.

6. Rosskopf D, Bornhorst A, Rimmbach C, Schwahn C, Kayser A, et al. (2007) Comment on "A common genetic variant is associated with adult and childhood obesity". Science (New York, NY) 315: 187; author reply 187.

7. Dina C, Meyre D, Samson C, Tichet J, Marre M, et al. (2007) Comment on "A common genetic variant is associated with adult and childhood obesity". Science (New York, NY) 315: 187; author reply 187.

8. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, et al. (2007) A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science 316: 889-894.

9. Scuteri A, Sanna S, Chen WM, Uda M, Albai G, et al. (2007) Genome-wide association scan shows genetic variants in the *FTO* gene are associated with obesity-related traits. PLoS Genet 3: e115.

10. Dina C, Meyre D, Gallina S, Durand E, Korner A, et al. (2007) Variation in *FTO* contributes to childhood obesity and severe adult obesity. Nat Genet 39: 724-726.

11. Adair LS (2004) Dramatic rise in overweight and obesity in adult Filipino women and risk of hypertension. Obesity Res 12: 1335-1341.

12. Durnin JV, Womersley J (1974) Body fat assessed from total body density and its estimation from skinfold thickness: measurements on 481 men and women aged from 16 to 72 years. The British J of Nut 32: 77-97.

13. Siri WE, Brozek J, Hanschel A (1961) Body composition from fluid space and density. Techniques for measuring body composition. Washington, DC: National Academy of Science. pp. 223-244.

14. Food and Nutrition Research Institute of the Philippines (FNRI). Food Composition Tables: Recommended for use in the Philippines, Handbook 1. Manilla, Philippines: Department of Science and Technology (DOST). (1990)

15. Tuazon MA, van Raaij JM, Hautvast JG, Barba CV (1987) Energy requirements of pregnancy in the Philippines. Lancet 2: 1129-1131.

16. Ainsworth BE, Haskell WL, Whitt MC, Irwin ML, Swartz AM, et al. (2000) Compendium of physical activities: an update of activity codes and MET intensities. Med Sci Sports Exerc 32: S498-504.

17. Physical status: The use and interpretation of anthropometry. Report of a WHO Expert Committee. (1995) World Health Organization Technical Report Series 854: 1-452.

18. Kurokawa N, Nakai K, Kameo S, Liu ZM, Satoh H (2001) Association of BMI with the beta3-adrenergic receptor gene polymorphism in Japanese: meta-analysis. Obesity Res 9: 741-745.

19. Fujisawa T, Ikegami H, Kawaguchi Y, Ogihara T (1998) Meta-analysis of the association of Trp64Arg polymorphism of beta 3-adrenergic receptor gene with body mass index. The Journal of clinical endocrinology and metabolism 83: 2441-2444.

20. Emorine L, Blin N, Strosberg AD (1994) The human beta 3-adrenoceptor: the search for a physiological function. Trends Pharmacol Sci 15: 3-7.

21. Li H, Wu Y, Loos RJ, Hu FB, Liu Y, et al. (2007) Variants in FTO gene are not associated with obesity in a Chinese Han population. Diabetes.

CHAPTER IV. GENOME-WIDE ASSOCIATION STUDY OF LIPOPROTEIN CHOLESTEROL AND TRIGLYCERIDES IN THE CEBU LONGITUDINAL HEALTH AND NUTRITIONAL SURVEY (CLHNS) COHORT

CHAPTER IV CREDITS

The work described in this chapter was performed in collaboration with others. Damien C. Croteau-Chonka performed data manipulation, statiscial analysis, and interpreted results. Ethan M. Lange performed statistical analysis and interpreted results. Anh Le and Christopher Kuzawa measured lipid levels and analyzed lipid data. Leslie A. Lange designed the study, performed statistical analysis, and interpreted results. Linda S. Adair and Karen L. Mohlke designed the study and interpreted results.

I would like to thank the entire staff of the Office of Population Studies Foundation (OPS) and the participants of CLHNS for their long-term work and dedication to this study. I would also like to thank Yun Li for her work on the imputation and Shawn Levy at the Vanderbilt Microarray Shared Resource Center for directing the genotyping. This work was supported by NIH R01 DK78150 and TW05596, and sample collection was supported by pilot funds from RR20649, ES10126, and DK56350. A.F.M. was supported by an Integrative Vascular Biology Fellowship, NIH grant HL69768. D.C.C.-C. was supported by an NRSA in Genetics (GM007092).

ABSTRACT

We performed a genome-wide association (GWA) study for HDL-cholesterol (HDL-C), triglycerides, LDL-cholesterol (LDL-C), and total cholesterol in a cohort of 1,780 adult Filipino women from the Cebu Longitudinal Health and Nutrition Survey (CLHNS). Among the ~2.1 million single nucleotide polymorphisms (SNPs) analyzed, the most significant associations after adjusting for age, age^2, total assets, income, number of previous pregnancies, menopausal status, and principal components included rs5882 at *CETP* with HDL-C ($P = 4.01 \times 10^{-8}$) and rs662799 at *APOA1* with triglycerides ($P = 1.23 \times 10^{-14}$). Evidence of association ($P < .05$) was observed at eleven loci previously described in GWA studies of lipoprotein levels, and evidence for two independent signals was detected at *LIPC*, *CETP*, and *GCKR*. Loci with suggestive evidence of association ($P < 10^{-5}$) and not previously described include Collecting-12 (*COLEC12*) associated with total cholesterol and Tankyrase (*TNKS*) associated with LDL-C. These two loci have potential biological relevance to lipid metabolism. The results confirm that even in a divergent population with limited power to detect association we can extend the evidence of association of many previously reported loci. The suggestive evidence of association of the two novel loci should be the basis for follow-up studies in subsequent populations.

INTRODUCTION

Plasma lipoprotein levels are associated with risk of coronary artery disease (CAD), one of the leading causes of death worldwide [1]. Low levels of high density lipoprotein cholesterol (HDL-C) and high levels of triglycerides and low density lipoprotein cholesterol (LDL-C) are associated with increased risk of CAD [2-4]. While diet, exercise, weight, alcohol consumption, and smoking affect levels of lipoproteins, heritability estimates for lipoprotein concentrations are as high as 75% [5,6]. The influence of genetic variants on lipoprotein levels may differ across populations. For these reasons, there has been considerable interest in understanding genetic factors that contribute to inter individual variation.

Recent genome-wide association (GWA) studies in populations of primarily European descent have identified many loci that are associated with blood plasma lipoprotein levels [7-19]. Some identified loci correspond to previously known genes and others suggest new pathways relevant to lipoprotein metabolism . Despite the large number of loci identified in these studies (>30), the associated single nucleotide polymorphisms (SNPs) only explain 5-8% of the variation in HDL-C, LDL-C, or triglycerides, indicating that most of the heritability of these traits remains to be explained [20].

Association studies of lipoprotein levels in Asian populations have replicated some but not all loci from European GWA studies [21-23]. Some Asian populations are undergoing socio-economic development and lifestyle changes that are increasing CAD burden [24]. The environmental differences and the potential differences in genetic architecture between ethnic groups within and between European and Asian populations

may modify the effect of genetic variants on lipoprotein levels.

In the current study, we performed a GWA analysis to investigate genetic factors influencing lipoprotein and triglyceride levels in mothers from the Cebu Longitudinal Health and Nutrition Survey (CLHNS), using data from 1,780 unrelated healthy women from Cebu, Philippines.

MATERIALS AND METHODS

Study population and phenotypes

We initially evaluated 1,895 healthy Cebu Filipino female participants from the ongoing CLHNS, mothers of a 1983 to 1984 birth cohort [25]. Trained field staff conducted in-home interviews and collected quantitative anthropometric measurements, blood samples for DNA and biomarkers, and comprehensive environmental data (available on-line at www.cpc.unc.edu/projects/cebu/).

Outcomes and covariates were measures were during the 2005 survey. Lipid levels were collected from women who fasted overnight (>8 hours) and were not taking statins. Details of the measurement of lipid profiles in this cohort have been reported previously [26]. Total cholesterol was determined by enzymatic methods (Beckman Diagnostics, Palo Alto, CA) on a CX5 chemistry analyzer, HDL-C was determined using a homogenous assay (Genzyme, Exton, PA), and triglyceride (TG) concentrations were measured with a glycerol blank as a 2-step reaction (Beckman Coulter Diagnostics, Fullerton, CA). LDL-C was determined using the Friedewald formula, except if TG exceeded 400 mg/dl then LDL-C was directly determined using a homogenous assay (Genzyme, Exton, PA).

Except for the correlation between LDL-C and total cholesterol ($r^2 = .81$), the four traits show low pair-wise correlation with one another ($r^2 < .1$) (Table 4.1). Body mass index (BMI) was calculated as weight/height^2 (kg/m^2).

Informed consent was obtained from all CLHNS subjects, and the University of North Carolina Institutional Review Board for the Protection of Human Subjects approved the study protocol.

SNP genotyping

Genome-wide SNP genotyping was performed with the commercial release of the Affymetrix Genome-Wide Human SNP Array 5.0. Genotyping was carried out at the Vanderbilt Microarray Shared Resource at Vanderbilt University Medial Center, Nashville, TN, using the standard protocol recommended by the manufacturer. Genotype calling was performed using the Birdseed calling algorithm (version 2). Genotyping was attempted on 1,895 unique CLHNS samples, 40 CLHNS duplicates, and 5 HapMap CEPH trios (whose genotypes were downloaded from HapMap.org). After sample quality control (QC) checks, ten CLHNS samples could not be genotyped because one failed DNA fragmentation and nine failed an array quality control check (DM algorithm). An additional four CLHNS samples were removed after genotyping because their overall genotyping call rate was < 97%. The final Birdseed call rate across 1,881 samples was 99.6%.

We applied SNP QC checks in PLINK v1.02 on the 1,881 CLHNS samples that were successfully genotyped. Of the initial 424,670 SNPs, we discarded 16,564 SNPs due to poor mapping, call rates < 90%, deviation from Hardy-Weinberg equilibrium (P <

10^{-6}), inconsistent genotypes in duplicate samples and Mendelian inheritance errors (combined ≥ 3 discrepancies among 40 duplicate pairs and 5 CEPH trios), and/or ≥ 3 genotype discrepancies with HapMap genotypes.

Based on identity-by-descent (IBD) and identity-by-state (IBS) estimates calculated in PLINK in combination with prior knowledge of the CLHNS samples, eighty-one samples were excluded from pairs or trios of likely first-degree relatives (either mother/daughter or sisters). The final sample set consisted of 1,780 CLHNS women with available genotypes and phenotypes of lipoprotein levels and covariates from the 2005 survey.

Additional genotyping of 2 SNPs at the *APOE* locus (rs7412 Arg158Cys and rs429358 Cys112Arg) was performed using a TaqMan allelic discrimination assay (Applied Biosystems, Foster City, CA). The genotype success rate for both SNPs was >98%, and no discrepancies were observed among 78 duplicate sample pairs per SNP.

SNP imputation

Using a Hidden Markov Model algorithm implemented in the MACH software version 1.0 [27] (www.sph.umich.edu/csg/abecasis/mach/), genotype imputation was conducted using 352,264 directly genotyped SNPs polymorphic in both the 60 HapMap CEU founders and the 89 combined CHB+JPT samples. We pooled haplotypes from phased chromosomes in these populations to better capture the linkage disequilibrium (LD) structure in the CLHNS samples; we have shown previously that the CHB HapMap population is a reasonable proxy for the CLHNS cohort [28]. Imputation yielded genotype data for 1,878,188 additional SNPs; the 352,264 directly genotyped SNPs were

also assigned imputed genotypes in this process. Out of a now total 2,230,452 SNPs, we discarded 150,177 low-quality imputations (Rsq ≤ .3) and 30,351 SNPs with estimated minor allele frequencies (MAF) ≤ .01, retaining 2,049,924 imputed SNPs. For each subject, the imputed genotype at each SNP was reported as a dosage value (a continuous number between 0 and 2), reflecting the expected number of copies of the arbitrary allele at that SNP conditional on the directly observed genotypes in both the subject and all other CLHNS subjects and the phased haplotype assignments in the CEU and CHB+JPT HapMap samples. Finally, discrete dosage values of 0, 1 or 2 were assigned for 23,750 directly genotyped SNPs non-polymorphic in either HapMap populations (and therefore not imputed) but with MAF > .01 in the CLHNS samples. In total, 2,073,674 directly genotyped (n = 23,750) and imputed (n = 2,049,924) SNPs were tested for association.

Population substructure

We used two approaches to consider possible effects of population stratification. First, we constructed principal components (PCs) using the software EIGENSOFT [29,30] to identify population substructure among our CHLNS subjects. We used a set of 13,972 independent SNPs (estimated r^2 < .005 between all pairs of SNPs within 1Mb) with observed MAF > .05 and 1,571 CHLNS subjects with estimated pair-wise IBD < 0.1 to construct PCs. The PCs for the remaining 228 subjects were subsequently calculated using these parameter values. The corresponding eigenvalues for the individual PCs were plotted and the "elbow" of the corresponding eigenvalues for the individual PCs was used to select seven PCs to be included as covariates in our linear regression genotype association analyses to account for genetic ancestry differences among study

subjects. In addition, we tested for an association between each of the first 10 PCs and the four outcomes of interest to ensure that any important ancestry explanatory PC were included in the analyses (Table 4.2).

Second, we used the genomic control method to examine systematic p-value inflation due to cryptic relatedness between samples not accounted for by the analytic adjustment for PCs in the regression analyses. The genomic controls values are .988, 1.01, 1.01, and 1.00 for HDL-C, LDL-C, triglycerides, and total cholesterol, respectively, suggesting good control of the overall family-wise type I error rate.

GWA analysis

All phenotypes were natural log-transformed to satisfy the model assumption of normally distributed residuals, conditional on the covariates. Covariate adjustment was made for 7 PCs of population structure, age, age^2, total assets, natural log-transformed income, number of previous pregnancies (three categories: 0–4, 5–10, >10), and menopausal status. Each of these predictors was significantly associated ($P < .05$) with at least one trait in our sample (Table 4.2). Additional analysis adjusting for BMI excluded women who were pregnant at the time of the survey. Multivariable linear regression association analysis was performed using Array Studio version 3.1 (Omicsoft Corporation, Research Triangle Park, NC). We assumed an additive mode of inheritance and report β coefficients representing the estimated change in mean transformed trait value associated with each additional copy of an allele. A 1 degree-of-freedom likelihood ratio test was used to assess statistical significance. Quantile-quantile (Q-Q) plots, mapping observed versus expected $-\log_{10}$(P values), and Manhattan plots were

constructed to graphically display results across the genome and to help assess for any cumulative inflation or deflation of statistical significance estimates compared to expectation under the null hypothesis. Conditional analysis was performed by re-testing all SNPs with a 2 Mb region centered on the SNP representing the strongest primary signal using genotypes of the main effect SNP as an additional covariate in the linear regression.

APOE haplotype analysis

Haplotypes were estimated using Haplo.Stats [31] (Version 1.4.3 http://mayoresearch.mayo.edu/mayo/research/schaid_lab/software.cfm) and correspond to the three common protein isoforms of apolipoprotein E (*APOE*), encoded by the ε2, ε3, and ε4 alleles of the *APOE* gene. We used analysis of covariance (ANCOVA) models to test whether *APOE* haplotype was associated with the lipoprotein phenotypes, adjusting for the same covariates as the genome wide analysis. *APOE* haplotypes were tested using 2 approaches: 1) collapsing into three categories of the ε2 carriers, ε3/ε3, and the ε4 carriers, excluding ε2/ε4 individuals; and 2) the six estimated haplotypes. We used 2- and 5-degree of freedom F tests for these two approaches, respectively.

Selection of previously reported SNPs

To test evidence of association with previously reported loci, we selected one SNP from each loci reported in GWA studies with $P < 5 \times 10^{-8}$. When multiple non-independent SNPs were reported for a given locus we included the most strongly associated one from the previous reports. If two signals were reported for a locus, we

also chose a SNP to represent the second signal. If a previously reported SNP was not present in our dataset, we sought to identify a proxy SNP in high LD ($r^2 > .8$ in both CEU and CHB+JPT, HapMap Release 22).

RESULTS

We tested 2,073,674 SNPs for association with plasma levels of HDL-C, triglycerides, LDL-C, and total cholesterol in 1,780 Filipino women in the CLHNS. Basic descriptive characteristics of the cohort are summarized in Table 4.3. GWA analyses for all lipoprotein levels were adjusted for 7 PCs of population structure, age, age^2, total assets, natural log-transformed income, number of previous pregnancies, and menopausal status is summarized in Figure 4.1. Q-Q plots, mapping observed versus expected $-\log_{10}$ (P values), show little cumulative inflation or deflation of statistical significance estimates compared to expectation under the null hypothesis (Figure 4.2).

The CLHNS GWA results support ten loci previously described in GWA studies. We chose a single previously reported SNP from each locus and defined positive evidence for replication in CLHNS as $P < .05$ and an effect in the same direction as observed in the original study (Table 4.4**)**. The CLHNS study supports three previously reported loci associated with HDL-C at *MMAB/MVK*, *LIPC*, and *CETP*, six loci associated with triglycerides at *ANGPTL3/DOCK7*, *GCKR*, *LPL*, *TRIB1*, *APOA1*, and *LIPC*, and two loci associated with LDL-C at *HMGCR* and *HNF1A*. No loci previously reported to be associated with total cholesterol were significant in the CLHNS. Several of the loci previously reported for one trait also showed evidence of association ($P < .05$) with other traits in the CLHNS (Table 4.4).

For each of the ten previously reported loci with evidence of association in the CLHNS, we performed conditional analyses to evaluate whether any additional SNPs were associated with an outcome after accounting for the primary associated SNP. Evidence from these analyses suggests the presence of secondary association signals with HDL-C at both *LIPC* and *CETP*, consistent with previous reports [12]. For the *LIPC* locus, we observed a P value of .0047 for rs10468017, the SNP representing the primary previously reported signal, and a P value of 2.5×10^{-5} for rs2070895, the SNP from the previously reported secondary signal. When we adjusted for both of these SNPs, there was still a strong association with rs2043085 in the location of the first signal ($P = 2.8 \times 10^{-6}$) and with rs8023503 in the second signal ($P = 6.6 \times 10^{-7}$). These two SNPs are the most highly associated in the CLHNS data for this locus. For the SNP reported as the *CETP* first signal (rs173539), we did not have a genotyped or imputed proxy in high LD ($r^2 > .8$). However, after correcting for the second signal (rs289714 for which we observed a P value of 1.2×10^{-6}), there was still an association with rs711752. This SNP is in $r^2 = .5$ with rs173539 in the CEU HapMap population, and is the most significant SNP for what we believe is a proxy, albeit not perfect, for the first signal ($P = 5.6 \times 10^{-7}$). For the *GCKR* locus and association with triglycerides, the previously reported SNP rs1260326 had a P value of .0019 in the CLHNS. After correcting for this SNP, we still observed association with rs814295, which had a more significant association of $P = 1.4 \times 10^{-7}$ in the original analysis.

A representative SNP ($r^2 < .8$) for the previously reported LDL-C association at the *APOE* locus was not genotyped or imputed in the CLHNS. We independently genotyped in 1656 CLHNS women the two SNPs (rs429358 and rs7412) that compose

the common ApoE haplotypes (ε2, ε3, and ε4) [32]. Estimated haplotype frequencies were .11, .80, and .08 for ε2, ε3, and ε4, respectively. The overall genotype frequencies were .01, .19, .02, .64, .13, and .01 for ε2/ε2, ε2/ε3, ε2/ε4, e3/e3, ε3/ε4, and ε4/ε4, respectively. The P value for association between the 3-category ApoE haplotype variable and LDL-C was strongly significant ($P < .0001$). In Bonferroni-corrected *post hoc* pair-wise tests, the ε4 carriers had significantly higher LDL-C levels than the ε3 homozygotes and ε2 carriers ($P < .0001$ and $P = .0068$). The ε3 homozygotes were significantly higher than the ε2 carriers ($P < .0001$). Associations with total cholesterol and HDL were also significant ($P < .05$), but not with triglycerides (Table 4.5). When the haplotypes were analyzed as six categories, associations with LDL-C were highly significant ($P < .0001$) and approximately linear when ordered ε2/ε2, ε2/ε3, ε2/ε4, ε3/ε3, ε3/ε4, and ε4/ε4 with the least-square adjusted mean log-LDL-C levels of 4.41 (± .05), 4.59 (± .02), 4.70 (± .05), 4.76 (± .01), 4.81 (± .02), and 4.91 (± .07), respectively. The least-square adjusted mean log-LDL-C level correspond to raw LDL-C values of 87.1, 102.4, 114, 121.3, 127, and 140.4, respectively. As with the three-category haplotype analysis, associations were also observed for triglycerides, HDL-C, and total cholesterol (Table 4.5). Together, these association results are consistent with what have been observed by others [32].

For each of the four lipoprotein traits, we observe suggestive evidence of association ($P < 10^{-5}$) at loci not described previously (Table 4.6). Most notably, SNPs in intron 3 of tankyarase (*TNKS*) are associated with LDL-C ($P = 7.4 \times 10^{-6}$) and SNPs in intron 2 of collectin-12 (*COLEC12*) are associated with total cholesterol ($P = 5.3 \times 10^{-6}$). We performed conditional analyses on all the most significant main effect associations to

evaluate whether any additional SNPs were associated with a given outcome after accounting for the associated SNP. We found no evidence for a strong secondary signal at these loci.

Finally, we tested whether genotype associations with lipoprotein levels are mediated by BMI to further evaluate the characteristics of loci exhibiting evidence of association. However, adjusting for BMI did not considerably attenuate or increase the evidence of association with most variants, suggesting the changes in lipoprotein levels due to these variants and obesity are independent (data not shown). However, for the locus *ANGPTL3-DOCK7* with triglycerides we observed a change in P value from .04 with a ß of .038 ± .018 to .006 with a ß of .048 ± .017 after adjusting for BMI. However, this change in P value may be the result of stochastic variation.

DISCUSSION

We performed the first GWA study for lipoprotein levels in a cohort from the Philippines, a country undergoing socio-economic and nutrition transition. The strongest main effect association in the CLHNS was observed for the *APOA1-APOC3-APOA4-APOC2* locus for triglycerides (rs662799, $P = 1.2 \times 10^{-14}$), one of the most commonly replicated triglyceride signals. Additionally, these results support association of other loci: *MMAB/MVK*, *LIPC*, and *CETP* with HDL-C; *ANGPTL3/DOCK7*, *GCKR*, *LPL*, *TRIB1*, *APOA1*, and *LIPC* with triglycerides; and *HMGCR*, *HNF1A*, and *APOE* with LDL-C. Many of these genes were initially identified in candidate gene studies and have shown evidence of association across many populations. Together, all of the previously reported SNPs for each trait explain 4.2%, 5.5%, 1%, .5% of the variation for HDL–C,

triglycerides, LDL-C, and total cholesterol, respectively, leaving most of the heritability to be explained. Despite a lack of significant statistical association, the direction of association was consistent with previous reports for many of the other previously reported loci. Many of the SNPs that show association with the original reported trait are also associated with another trait, consistent with the correlations between traits. Lack of trait association in our study for some SNPs that have been previously reported likely primarily reflects a smaller sample size but may also be due to different LD between analyzed SNPs and the underlying functional variant. Using our replication criteria, the study had > 80% power to detect SNPs that explained .44% of the total variation in our lipoprotein traits after adjustment for covariates.

Compared to Western populations, CLHNS participants have relatively HDL-C values. HDL-C levels have been previously reported to be lower in Asians compared to European populations, even though Asians are less obese [33]. Low HDL-C is defined as < 50 mg/dL, and the mean level in the CLHNS women was 40.9 mg/dL, similar to that reported in another study of Filipino women [34]. Despite differing HDL-C levels in the CLHNS population compared to the previous European populations studied, we have replicated several previously reported loci associated with plasma levels of HDL-C. For the *CETP*, *LIPC*, and *MMAB/MVK* loci, the CLHNS samples have slightly higher allele frequencies for the HDL-C lowering alleles. These may contribute slightly to a decrease the population mean HDL-C because the risk alleles are affecting more members of the population. We further examined individuals in the CLHNS with isolated low HDL-C and observed similar results (data not shown). These results suggest that certain loci still seem important even in a different environmental/genetic background.

Some of the associations described first in European populations have also been replicated in other Asian populations. Hiura *et al.* conducted a GWA study for HDL-C in 900 Japanese individuals and observed evidence of association with *CETP*, but not with other previously identified loci [22]. Tai *et al*. examined the association with LDL-C, HDL-C and total cholesterol for four polymorphisms identified in previous GWA study of lipoproteins in 2932 Malay participants from Singapore [21]. The authors observed evidence of association for *TRIB1* with total cholesterol and LDL-C, but not HDL-C and *CILP2/PBX4* with HDL-C, a trait not previously reported to be associated with this locus. Of these two sets of findings, only the association of *CETP* with HDL-C is consistent with our results.

Among the loci with suggestive evidence of association ($P < 10^{-5}$) two signals are located within genes that that have potential biological relevance to metabolism, Collectin-12 (*COLEC12*) and Tankyrase (*TNKS*). SNPs within intron 2 of *COLEC12* are associated with total cholesterol and LDL-C levels. *COLEC12* mediates the recognition, internalization and degradation of oxidatively modified LDL by vascular endothelial cells [35]. SNPs within intron 3 of *TNKS* are associated with LDL-C and total cholesterol. TNKS-deficient mice exhibit an increase in energy expenditure, fatty-acid oxidation, and insulin-stimulated glucose utilization, and adiposity is substantially decreased even with excessive food intake [36]. While these genes seem biologically plausible, the evidence of association needs to be confirmed in additional populations. Each of these loci account to ~1% of the estimated proportion of variance in the CLHNS.

Our results show that multiple genetic risk factors identified in other world populations are also associated with lipid traits in Filipinos. These observations should

form the basis of understanding the mechanistic roles of these loci and identifying functional or causative variants.

Table 4.1. **Spearman's correlations between traits measured in 1780 CLHNS women**

	TG	LDL-C	Total Cholesterol
HDL-C	0.038	0.016	0.082
TG	-	0.00058	0.100
LDL-C	-	-	0.811

Table 4.2. Effect of potential covariates on lipoprotein outcomes for 1798 CLHNS women

	Covariates included in tests of effect	P-value for effect			
		Log HDL-C	**Log TG**	**Log LDL-C**	**Log Total Cholesterol**
Age	-	**.0040**	**<.0001**	**<.0001**	**<.0001**
Age^2	Age	.40	**.00020**	**<.0001**	**<.0001**
Assets	Age, Age^2	**<.0001**	**<.0001**	**<.0001**	**<.0001**
Log income	Age, Age^2	**.0019**	**<.0001**	**<.0001**	**<.0001**
#Pregnancies (3 categories)	Age, Age^2	**<.0001**	.076	**.00030**	**<.0001**
Menopause	Age, Age^2	**.016**	.11	.27	.16
PC1	Age, Age^2, All other PCs	**.0001**	.60	.0679	**.0070**
PC2	Age, Age^2, All other PCs	**.0028**	.15	.1001	.88
PC3	Age, Age^2, All other PCs	**.041**	.19	.77	.996
PC4	Age, Age^2, All other PCs	.34	.22	.34	.81
PC5	Age, Age^2, All other PCs	.31	.28	.54	.59
PC6	Age, Age^2, All other PCs	.27	.65	**.0085**	**.014**
PC7	Age, Age^2, All other PCs	**.0277**	.84	.44	.18
PC8	Age, Age^2, All other PCs	.61	.79	.24	.19
PC9	Age, Age^2, All other PCs	.27	.44	.999	.98
PC10	Age, Age^2, All other PCs	.74	.15	.39	.31

All factors and outcomes correspond to the 2005 survey. P-values in **bold** denote $P < .05$. PC1-PC10 are the Eigenstrat principal components; PC1-PC7 were included in SNP association analyses.

Table 4.3. Descriptive characteristics of CLHNS women

Trait	N	Mean ± SD
LDL-cholesterol (mg/dL)	1780	119.6 ± 33.7
HDL-cholesterol (mg/dL)	1780	40.9 ± 10.37
Triglycerides (mg/dL)	1780	130.8 ± 83.5
Total cholesterol (mg/dL)	1780	186.8 ± 39.3
Age (years)	1780	48.4 ± 6.1
Total number of pregnancies	1780	6.5 ± 3.0
Body mass index (kg/m^2)*	1768	24.3 ± 4.4

Values correspond to the 2005 survey.
*BMI reported only for women who were not pregnant.

Table 4.4. Evidence of association in CLHNS women at previously reported GWA loci

SNP	Gene	Chr	Pos	Allele 1[a]	Allele 2	Freq Allele 1	HDL-C P value	HDL-C (β ± s.e.m.)	Triglycerides P value	Triglycerides (β ± s.e.m.)	LDL-C P value	LDL-C (β ± s.e.m.)	Total P value
HDL-C													
rs4846914	*GALNT2*	1	228362314	A	G	0.29	0.36	.008 (.009)	0.34	-.017 (.017)	0.75	.003 (.01)	0.92
rs6754295[b]	*APOB*	2	21059688	G	T	0.69	0.60	-.005 (.010)	0.19	.024 (.018)	0.20	-.014 (.011)	0.43
rs10503669[d]	*LPL*	8	19891970	A	C	0.05	0.90	.002 (.020)	0.023	-.088 (.039)	0.088	-.039 (.023)	0.032
rs1883025	*ABCA1*	9	106704122	C	T	0.58	0.35	.009 (.009)	0.045	.036 (.018)	0.17	-.015 (.011)	0.83
rs471364	*TTC39B*	9	15279578?	T	C	0.97	0.24	.032 (.027)	0.46	.039 (.052)	0.51	.02 (.031)	0.22
rs7395662[b]	*MADD-FOLH1-NR1H3*	11	48475469	G	A	0.21	0.80	-.003 (.010)	0.92	.002 (.020)	0.75	-.004 (.012)	0.74
rs174547	*FADS1-FADS2-FADS3*	11	61327359	T	C	0.07	0.25	-.020 (.017)	0.23	.039 (.033)	0.21	-.024 (.019)	0.41
rs964184	*APOA1-APOC3-APOA4-APOA5*	11	116154127	C	G	0.76	0.053	.019 (0.01)	2.0E-13	-.14 (.019)	0.53	.007 (.011)	0.12
rs10892044	*APOA1-APOC3-APOA4-APOA5**	11	116272109	C	T	0.02	0.80	.007 (.027)	0.033	-.112 (.052)	0.57	-.018 (.031)	0.26
rs2338104	*MMAB-MVK*	12	108379551	G	C	0.39	**0.02**	**.021 (.009)**	0.88	-.003 (.017)	0.015	.024 (.010)	0.0069
rs10468017	*LIPC*	15	56465804	T	C	0.21	**0.0047**	**.031 (.011)**	0.0076	.057 (.021)	0.029	.027 (.013)	7.8E-04
rs2070895	*LIPC**	15	56511231	A	G	0.37	**2.5E-05**	**.046 (.011)**	0.028	.047 (.021)	0.44	.01 (.013)	0.014
rs289714	*CETP**	16	55564952	A	G	0.80	**1.2E-06**	**.083 (.017)**	0.49	-.023 (.033)	0.11	.031 (.020)	0.015
rs2271293	*LCAT-CTCF-PRMT8*	16	66459571	A	G	0.04	0.12	.034 (.022)	0.98	-.001 (.043)	0.86	.005 (.025)	0.39
rs4939883	*LIPG*	18	45421212	C	T	0.86	0.072	.022 (.012)	0.73	.008 (.023)	0.97	.0004 (.014)	0.46
rs2967605	*ANGPTL4*	19	8375738	C	T	0.42	0.064	.021 (.011)	0.085	-.037 (.022)	0.44	.010 (.013)	0.59
rs7679	*PLTP*	20	44009909	T	C	0.97	0.091	-.044 (.026)	0.75	-.016 (.050)	0.039	.062 (.03)	0.53
rs173539	*CETP*	N/A											
rs1800961	*HNF4A*	N/A											
Triglycerides													
rs1167998[b]	*ANGPTL3-DOCK7*	1	62704220	A	C	0.71	0.11	-.015 (.009)	**0.035**	**.038 (.018)**	0.79	.003 (.011)	0.45
rs4846914[e]	*GALNT2*	1	228362314	A	G	0.29	0.36	.008 (.009)	0.34	-.017 (.017)	0.75	.003 (.01)	0.92
rs7557067	*APOB*	2	21061717	A	G	0.30	0.58	.005 (.009)	0.17	-.025 (.018)	0.19	.014 (.011)	0.42
rs1260326	*GCKR*	2	27584444	T	C	0.42	0.67	.003 (.009)	**0.0019**	**.055 (.018)**	0.86	.002 (.011)	0.18
rs714052	*MLXIPL*	7	72502805	A	G	0.89	0.79	.0038 (.014)	0.19	.035 (.027)	0.42	.013 (.016)	0.25
rs7819412	*XKR6-AMAC1L2*	8	11082571	A	G	0.10	0.084	-.0253 (.0146)	0.042	-.057 (.028)	0.71	-.006 (.017)	0.24
rs12678919	*LPL*	8	19888502	A	G	0.95	0.87	-.003 (.020)	**0.019**	**.091 (.039)**	0.081	.04 (.023)	0.027
rs2954029	*TRIB1*	8	126560154	A	T	0.44	0.81	-.002 (.008)	**0.030**	**.035 (.016)**	0.39	.008 (.010)	0.15
rs174547	*FADS1-FADS2-FADS3*	11	61327359	C	T	0.93	0.25	.020 (.017)	0.23	-.039 (.033)	0.21	.024 (.019)	0.41
rs964184	*APOA1-AOC3-APOA4-APOA5*	11	116154127	G	C	0.24	0.053	-.019 (.010)	**2.0E-13**	**.14 (.019)**	0.53	-.007 (.011)	0.12
rs4775041[d]	*LIPC*	15	56461987	C	G	0.19	0.005	.030 (.011)	**0.0084**	**.054 (.021)**	0.021	.028 (.012)	5.5E-04
rs17216525	*NCAN-CILP2-PBX4*	19	19523220	C	T	0.93	0.96	.001 (.016)	0.94	.002 (.031)	0.20	-.024 (.019)	0.45
rs7679	*PLTP*	20	44009909	C	T	0.03	0.091	.044 (.026)	0.75	.016 (.050)	0.039	-.062 (.030)	0.53
rs157580[b]	*CEACAM16-TOMM40*	N/A											

LDL-C													
rs11206510	*PCSK9*	1	55268627	T	C	0.94	0.087	.032 (.018)	0.51	.024 (.035)	0.61	.011 (.021)	0.43
rs12740374	*CELSR2-PSRC1-SORT1*	1	109619113	G	T	0.95	0.24	-.026 (.022)	0.15	.06 (.042)	0.062	.046 (.025)	0.10
rs4844614[c]	*CR1L*	1	205941798	T	G	0.23	0.48	.009 (.013)	0.98	.001 (.025)	0.90	-.002 (.015)	0.69
rs515135	*APOB*	2	21139562	C	T	0.92	0.18	.022 (.016)	0.38	-.026 (.030)	0.70	.007 (.018)	0.74
rs3846663	*HMGCR*	5	74691482	T	C	0.54	0.63	-.004 (.008)	0.51	.011 (.016)	**0.034**	**.020 (.010)**	0.036
rs1501908	*TIMD4-HAVCR1*	5	156330747	C	G	0.65	0.85	.002 (.009)	2.6E-04	.063 (.017)	0.57	.006 (.010)	0.059
rs12670798[b]	*DNAH11*	7	21573877	C	T	0.56	0.15	.018 (.012)	0.77	.007 (.024)	0.67	.006 (.014)	0.31
rs174570[b]	*FADS3-FADS2*	11	61353788	C	T	0.16	0.96	-.001 (.014)	0.14	.04 (.027)	0.23	-.020 (.016)	0.69
rs2650000	*HNF1A*	12	119873345	A	C	0.42	0.71	.003 (.009)	0.96	.001 (.017)	**0.014**	**.025 (.010)**	0.023
rs10401969	*NCAN-CILP2-PBX4*	19	19268718	T	C	0.95	0.34	-.029 (.031)	0.21	.074 (.059)	0.59	-.019 (.035)	0.82
rs10402271[f]	*BCAM*	19	50021054	G	T	0.17	0.82	.003 (.011)	0.24	.025 (.021)	0.040	-.026 (.013)	0.13
rs6102059	*MAFB*	20	38662198	C	T	0.58	0.83	-.002 (.009)	0.20	-.022 (.017)	0.63	-.005 (.010)	0.61
rs6544713	*ABCG8*	N/A											
rs4420638	*LDLR*	N/A											
rs2075650[f]	*TOMM40*	N/A											
rs6511720	*APOE-APOC1-APOC4-APOC2*	N/A											
rs4803750[f]	*BCL3*	N/A											
Total Cholesterol													
rs10903129[b]	*TMEM57*	1	25641524	G	A	0.16	0.62	-.006 (.012)	0.49	-.015 (.022)	0.97	-.001 (.013)	0.82
rs10889353[b]	*DOCK7*	1	62890784	A	C	0.76	0.24	-.012 (.010)	0.057	.036 (.019)	0.80	-.003 (.011)	0.78
rs646776[b]	*SARS-CELSR2-MYBPHL*	1	109620053	T	C	0.95	0.25	-.025 (.022)	0.16	.059 (.041)	0.068	.045 (.025)	0.10
rs693[b]	*APOB*	2	21085700	A	G	0.08	0.14	.024 (.016)	0.42	.025 (.030)	0.55	.011 (.018)	0.21
rs3846662[b]	*GCNT4-HMGCR-POLK*	5	74686840	G	A	0.55	0.60	-.004 (.008)	0.51	.011 (.016)	0.078	.017 (.010)	0.069
rs6987702[b]	*TRIB1-FAM84B*	8	126573908	C	T	0.64	0.18	.014 (.011)	0.22	-.024 (.020)	0.31	.012 (.012)	0.56
rs174570[b]	*FADS3-FADS2*	11	61353788	C	T	0.16	0.96	-.001 (.014)	0.14	.04 (.027)	0.23	-.02 (.016)	0.69
rs2304128[b]	*NCAN-ZNF104*	19	19607151	G	T	0.96	0.66	.015 (.035)	0.51	.043 (.067)	0.88	-.006 (.040)	0.60
rs6756629[b]	*ABCG5*	N/A											
rs2228671[b]	*LDLR*	N/A											

Chromosomal positions are NCBI Build 36. SNPs are arranged in ascending chromosomal order.
SNPs previously reported to be associated with lipoprotein levels in GWA studies ($P < 10^{-8}$) are included. SNPs are from Kathiresanat al. 2009 unless otherwise indicated. **Bold** represents SNPs with a P value < .05 and direction of effect consistent with previous reports.

*Asterisks represent second signals reported in Kathiresan *et al.* 2009.
a Allele associated with increased trait value in prior report
b Aulchenko *et al.* 2009
c Sabatti *et al.* 2009
d Willer *et al.* 2008
e Kathiresan *et al.* 2008
f Sandhu *et al.* 2008

Table 4.5. *APOE* haplotype association analysis in CLHNS

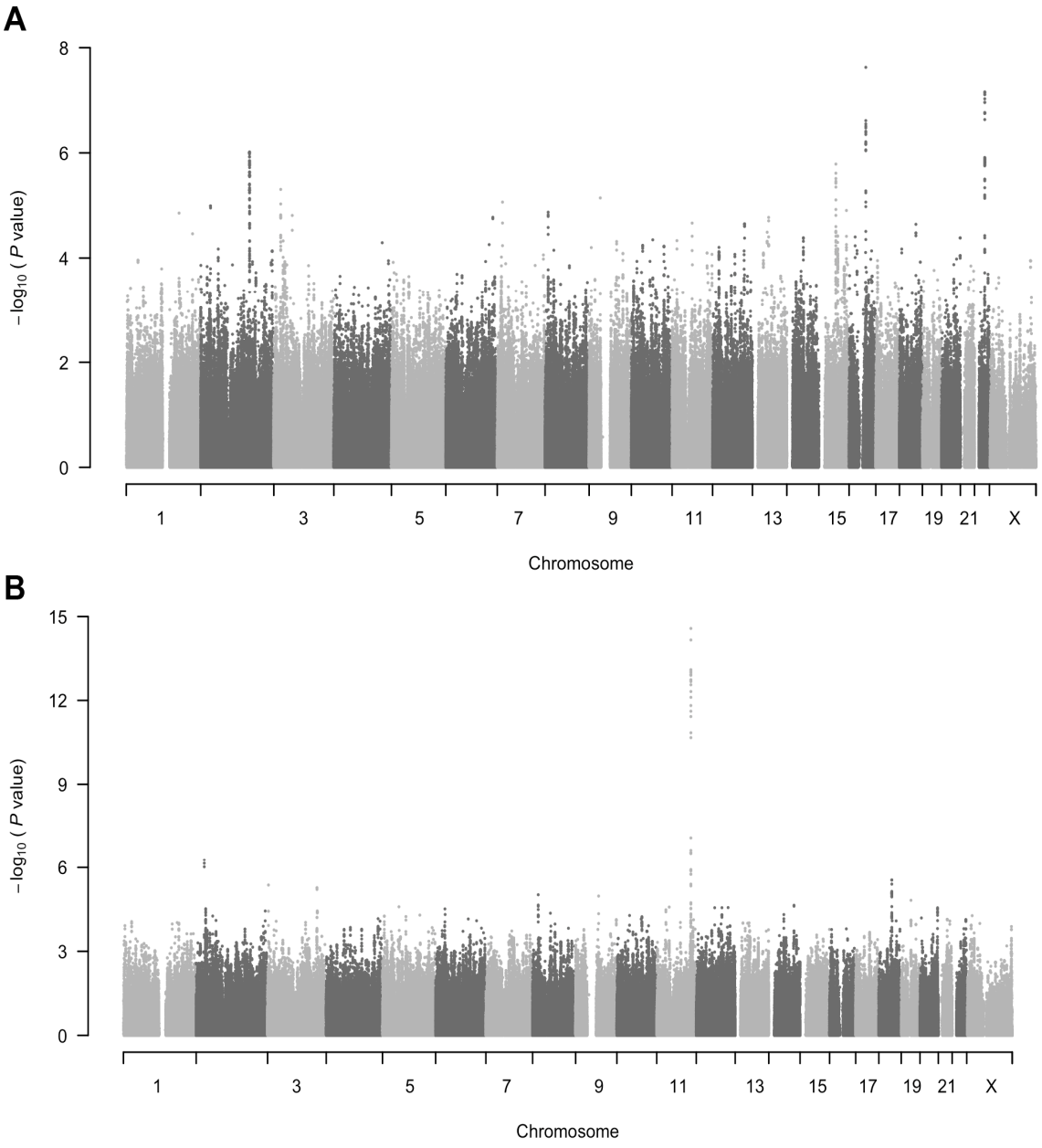
***APOE* haplotype**	**Genotype frequency**	**HDL-C**	**Triglycerides**	**LDL-C**
3 categories				
ε2 carriers	.20	3.71 (0.01)	4.76 (0.03)	4.57 (0.02)
ε3/ ε3	.64	3.67 (0.01)	4.72 (0.02)	4.76 (0.01)
ε4 carriers	.14	3.66 (0.02)	4.73 (0.03)	4.82 (0.02)
P value		.019	.53	<.0001
6 categories				
ε2/ ε2	.01	3.78 (0.05)	5.11 (0.1)	4.41 (0.06)
ε2/ ε3	.19	3.71 (0.02)	4.73 (0.03)	4.59 (0.02)
ε2/ ε4	.02	3.67 (0.05)	114.01 (5.78)	4.7 (0.05)
ε3/ ε3	.64	3.67 (0.01)	4.72 (0.02)	4.76 (0.01)
ε3/ ε4	.13	3.65 (0.02)	4.73 (0.03)	4.81 (0.02)
ε4/ ε4	.01	3.75 (0.06)	4.81 (0.12)	4.91 (0.07)
P value		.035	.0008	<.0001

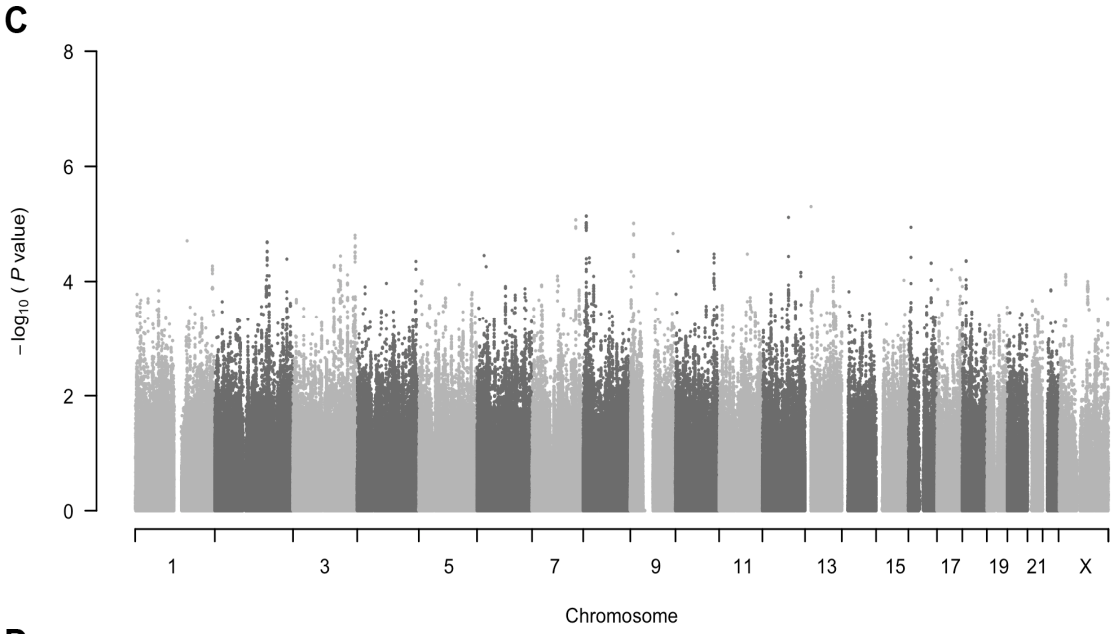
Least-square means (SE) are shown.

Table 4.6. SNPs with suggestive evidence of association with lipoprotein levels in CLHNS ($P < 10^{-5}$)

SNP	Gene	Chr	Pos	Allele1	Allele2	Freq Allele1	HDL P value	HDL (β ± s.e.m.)	TG P value	TG (β ± s.e.m.)	LDL P value	LDL (β ± s.e.m.)	Total Cholesterol P value	Total Cholesterol (β ± s.e.m.)
TG														
rs17023681	*CNTN4*	3	2978279	T	G	0.72	0.87	-0.002 (0.013)	4.1E-06	-0.114 (0.025)	0.50	-0.010 (0.015)	0.0088	-0.027 (0.010)
rs7644509	*Chr3q26.1*	3	1.68E+08	C	G	0.85	0.21	0.015 (0.012)	6.3E-06	-0.104 (0.023)	0.88	0.002 (0.014)	0.22	-0.012 (0.010)
rs2453464	*PCSK5*	9	78109296	A	G	0.16	0.11	-0.020 (0.012)	5.5E-06	0.107 (0.024)	0.45	-0.011 (0.014)	0.38	0.009 (0.010)
rs1893838	*ZBTB7C*	18	44132146	T	C	0.65	0.29	-0.010 (0.009)	5.0E-06	-0.078 (0.017)	0.16	-0.014 (0.010)	7.1E-04	-0.024 (0.007)
HDL														
rs17548357	*BIRC6*	2	32510121	G	A	0.99	8.5E-06	0.174 (0.039)	0.93	-0.007 (0.075)	0.80	0.011 (0.045)	0.28	0.034 (0.032)
rs1544857	*SLC4A10*	2	1.62E+08	G	C	0.83	1.6E-06	0.056 (0.012)	0.88	-0.003 (0.022)	0.56	0.008 (0.013)	0.13	0.014 (0.009)
rs3739440	*PAX5*	9	37021074	C	T	0.84	5.4E-06	0.059 (0.013)	0.79	0.007 (0.025)	0.053	-0.029 (0.015)	0.71	-0.004 (0.011)
rs11227643	*11q13.1*	11	66516071	G	C	0.78	7.7E-06	-0.055 (0.012)	0.27	0.026 (0.024)	0.38	-0.012 (0.014)	0.18	-0.013 (0.010)
rs138779	*TOM1*	22	34042177	T	C	0.59	1.9E-07	0.046 (0.009)	0.66	-0.008 (0.017)	0.72	0.004 (0.010)	0.12	0.011 (0.007)
LDL														
rs4570159	*TNKS*	8	9568712	G	A	0.69	0.17	0.012 (0.009)	0.16	-0.025 (0.017)	7.4E-06	0.046 (0.010)	9.4E-05	0.029 (0.007)
rs4787103	*A2BP1*	16	7982718	G	A	0.64	0.66	-0.005 (0.012)	0.45	0.017 (0.023)	6.5E-06	-0.061 (0.014)	5.6E-04	-0.033 (0.010)
TC														
rs1414513	*HLX*	1	2.19E+08	A	G	0.90	0.097	-0.024 (0.015)	0.080	-0.049 (0.028)	6.1E-04	-0.057 (0.017)	7.5E-06	-0.053 (0.012)
rs551314	*TBA3C*	13	18620448	A	C	0.98	0.32	0.038 (0.039)	0.11	0.118 (0.074)	1.2E-05	0.193 (0.044)	4.7E-06	0.142 (0.031)
rs2032158	*COLEC12*	18	349196	C	T	0.94	0.23	-0.022 (0.018)	0.019	-0.080 (0.034)	2.9E-04	-0.074 (0.020)	5.3E-06	-0.065 (0.014)

Figure 4.1. Genome-wide evidence of association with lipoproteins in 1,780 CLHNS women. A HDL-C, B. triglycerides, C. LDL-C, and D. total cholesterol levels. Each plot represents 2,073,674 SNPs that were tested for association adjusted for age, age^2, number of previous pregnancies, menopause status, and 7 PCs for population substructure.




C
8
6
4
2
0
– log10 (P value)
1
3
5
7
9
11
13
15
17
19
21
X
Chromosome

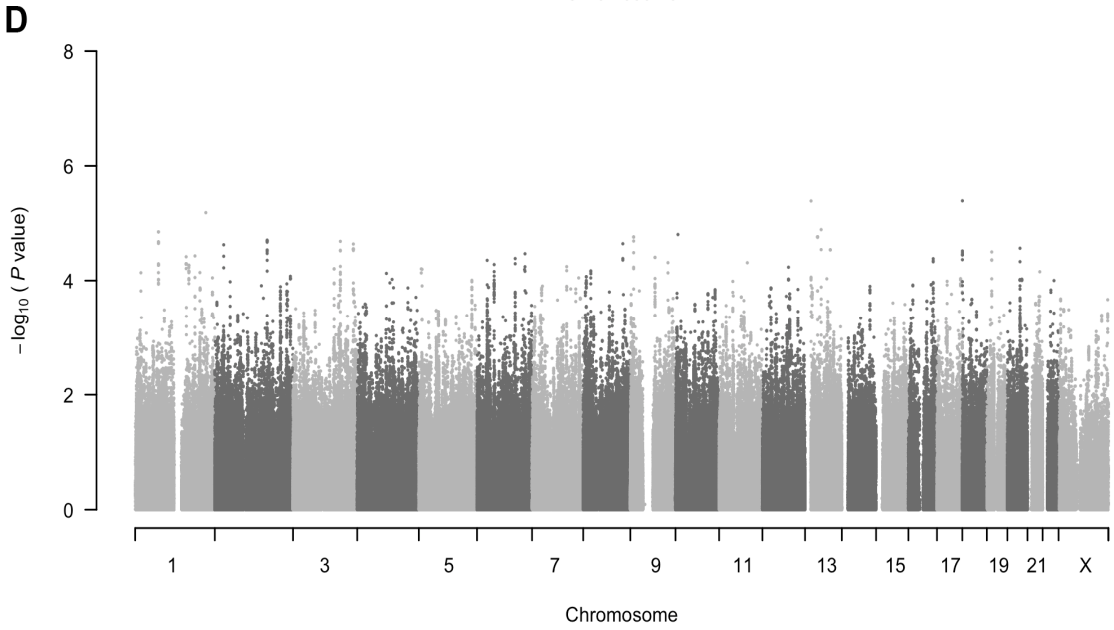
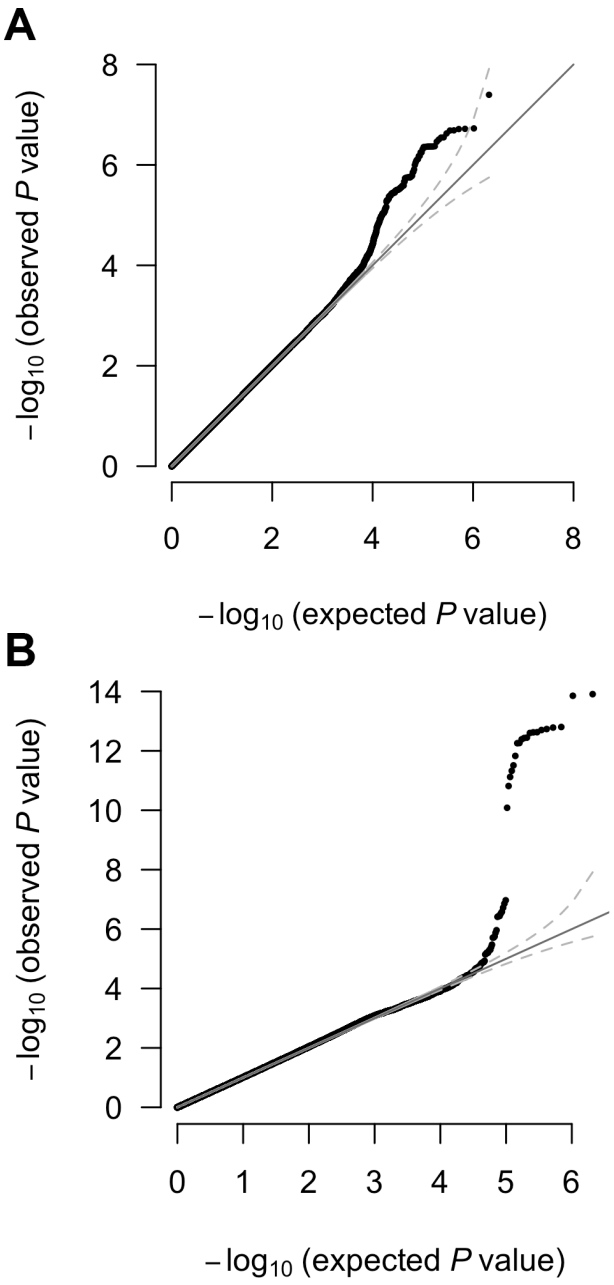
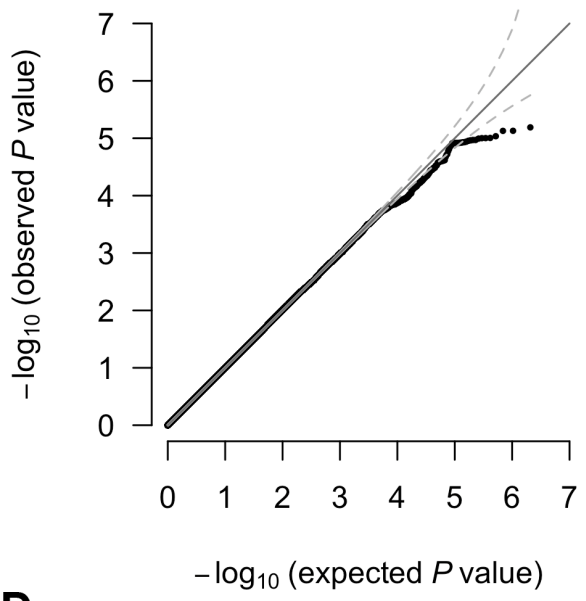

D
8
6
4
2
0
– log10 (P value)
1
3
5
7
9
11
13
15
17
19
21
X
Chromosome

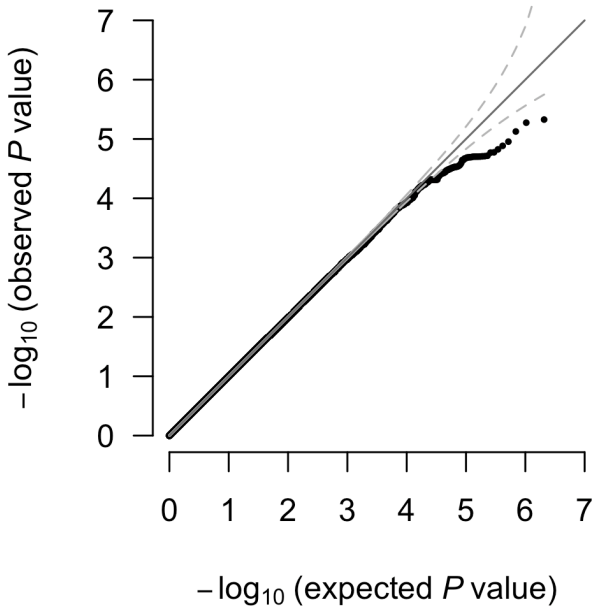
Figure 4.2. Quantile-Quantile plots for tests of association with lipoproteins in 1,780 CLHNS women A. HDL-C, B. triglycerides, C. LDL-C, and D. total cholesterol. Dotted lines indicate 95% confidence intervals



C


7
6
5
4
3
2
1
0
0 1 2 3 4 5 6 7
−log10 (observed P value)
−log10 (expected P value)

D


7
6
5
4
3
2
1
0
0 1 2 3 4 5 6 7
−log10 (observed P value)
−log10 (expected P value)

REFERENCES

1. Gaziano TA, Bitton A, Anand S, Abrahams-Gessel S, Murphy A Growing epidemic of coronary heart disease in low- and middle-income countries. Curr Probl Cardiol 35: 72-115.

2. Grundy SM, Cleeman JI, Merz CN, Brewer HB, Jr., Clark LT, et al. (2004) Implications of recent clinical trials for the National Cholesterol Education Program Adult Treatment Panel III guidelines. Circulation 110: 227-239.

3. Gotto AM, Jr., Brinton EA (2004) Assessing low levels of high-density lipoprotein cholesterol as a risk factor in coronary heart disease: a working group report and update. J Am Coll Cardiol 43: 717-724.

4. Nordestgaard BG, Benn M, Schnohr P, Tybjaerg-Hansen A (2007) Nonfasting triglycerides and risk of myocardial infarction, ischemic heart disease, and death in men and women. JAMA 298: 299-308.

5. O'Connell DL, Heller RF, Roberts DC, Allen JR, Knapp JC, et al. (1988) Twin study of genetic and environmental effects on lipid levels. Genet Epidemiol 5: 323-341.

6. Pilia G, Chen WM, Scuteri A, Orru M, Albai G, et al. (2006) Heritability of cardiovascular and personality traits in 6,148 Sardinians. PLoS Genet 2: e132.

7. Chasman DI, Pare G, Zee RYL, Parker AN, Cook NR, et al. (2008) Genetic loci associated with plasma concentration of low-density lipoprotein cholesterol, high-density lipoprotein cholesterol, triglycerides, apolipoprotein a1, and apolipoprotein b among 6382 white women in genome-wide analysis with replication. Circ Cardiovasc Genet 1: 21-30.

8. Heid IM, Boes E, Muller M, Kollerits B, Lamina C, et al. (2008) Genome-wide association analysis of high-density lipoprotein cholesterol in the population-based KORA study sheds new light on intergenic regions. Circ Cardiovasc Genet 1: 10-20.

9. Aulchenko YS, Ripatti S, Lindqvist I, Boomsma D, Heid IM, et al. (2009) Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. Nat Genet 41: 47-55.

10. Burkhardt R, Kenny EE, Lowe JK, Birkeland A, Josowitz R, et al. (2008) Common SNPs in HMGCR in micronesians and whites associated with LDL-cholesterol levels affect alternative splicing of exon13. Arterioscler Thromb Vasc Biol 28: 2078-2084.

11. Kathiresan S, Melander O, Guiducci C, Surti A, Burtt NP, et al. (2008) Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. Nat Genet 40: 189-197.

12. Kathiresan S, Willer CJ, Peloso GM, Demissie S, Musunuru K, et al. (2009) Common variants at 30 loci contribute to polygenic dyslipidemia. Nat Genet 41: 56-65.

13. Kooner JS, Chambers JC, Aguilar-Salinas CA, Hinds DA, Hyde CL, et al. (2008) Genome-wide scan identifies variation in *MLXIPL* associated with plasma triglycerides. Nat Genet 40: 149-151.

14. Pollin TI, Damcott CM, Shen H, Ott SH, Shelton J, et al. (2008) A null mutation in human *APOC3* confers a favorable plasma lipid profile and apparent cardioprotection. Science 322: 1702-1705.

15. Sabatti C, Service SK, Hartikainen AL, Pouta A, Ripatti S, et al. (2009) Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. Nat Genet 41: 35-46.

16. Sandhu MS, Waterworth DM, Debenham SL, Wheeler E, Papadakis K, et al. (2008) LDL-cholesterol concentrations: a genome-wide association study. Lancet 371: 483-491.

17. Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, et al. (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 316: 1331-1336.

18. Wallace C, Newhouse SJ, Braund P, Zhang F, Tobin M, et al. (2008) Genome-wide association study identifies genes for biomarkers of cardiovascular disease: serum urate and dyslipidemia. Am J Hum Genet 82: 139-149.

19. Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, et al. (2008) Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet 40: 161-169.

20. Mohlke KL, Boehnke M, Abecasis GR (2008) Metabolic and cardiovascular traits: an abundance of recently identified common genetic variants. Hum Mol Genet 17: R102-108.

21. Tai ES, Sim XL, Ong TH, Wong TY, Saw SM, et al. (2009) Polymorphisms at newly identified lipid-associated loci are associated with blood lipids and cardiovascular disease in an Asian Malay population. J Lipid Res 50: 514-520.

22. Hiura Y, Shen CS, Kokubo Y, Okamura T, Morisaki T, et al. (2009) Identification of genetic markers associated with high-density lipoprotein-cholesterol by genome-wide screening in a Japanese population: the Suita study. Circ J 73: 1119-1126.

23. Nakayama K, Bayasgalan T, Yamanaka K, Kumada M, Gotoh T, et al. (2009) Large scale replication analysis of loci associated with lipid concentrations in a Japanese population. J Med Genet 46: 370-374.

24. Yusuf S, Reddy S, Ounpuu S, Anand S (2001) Global burden of cardiovascular diseases: Part II: variations in cardiovascular disease by specific ethnic groups and geographic regions and prevention strategies. Circulation 104: 2855-2864.

25. Adair LS (2004) Dramatic rise in overweight and obesity in adult Filipino women and risk of hypertension. Obesity Research 12: 1335-1341.

26. Kuzawa CW, Adair LS (2003) Lipid profiles in adolescent Filipinos: relation to birth weight and maternal energy status during pregnancy. Am J Clin Nutr 77: 960-966.

27. Li Y, Willer C, Sanna S, Abecasis G (2009) Genotype imputation. Annu Rev Genomics Hum Genet 10: 387-406.

28. Marvelle AF, Lange LA, Qin L, Wang Y, Lange EM, et al. (2007) Comparison of ENCODE region SNPs between Cebu Filipino and Asian HapMap samples. Journal of human genetics 52: 729-737.

29. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2: e190.

30. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904-909.

31. Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 70: 425-434.

32. Bennet AM, Di Angelantonio E, Ye Z, Wensley F, Dahlin A, et al. (2007) Association of apolipoprotein E genotypes with lipid levels and coronary risk. JAMA 298: 1300-1311.

33. Kim SM, Han JH, Park HS (2006) Prevalence of low HDL-cholesterol levels and associated factors among Koreans. Circ J 70: 820-826.

34. Morales DD, Punzalan FE, Paz-Pacheco E, Sy RG, Duante CA (2008) Metabolic syndrome in the Philippine general population: prevalence and risk for atherosclerotic cardiovascular disease and diabetes mellitus. Diab Vasc Dis Res 5: 36-43.

35. Ohtani K, Suzuki Y, Eda S, Kawai T, Kase T, et al. (2001) The membrane-type collectin *CL-P1* is a scavenger receptor on vascular endothelial cells. J Biol Chem 276: 44222-44228.

36. Yeh TY, Beiswenger KK, Li P, Bolin KE, Lee RM, et al. (2009) Hypermetabolism, hyperphagia, and reduced adiposity in tankyrase-deficient mice. Diabetes.

CHAPTER V. TRANSCRIPTIONAL ACTIVITY OF SNPS AT THE *GALNT2* LOCUS ASSOCIATED WITH HUMAN HIGH DENSITY LIPOPROTEIN CHOLESTEROL LEVELS

CHAPTER V CREDITS

The work described in this chapter was performed in collaboration with others. Kyle J. Gaulton performed study design and analysis of the resquencing data. Arlene J. González assisted in laboratory experiments. Jason D. Lieb, Terrence S. Furey, Gregory E. Crawford, designed the studies and performed analysis of the functional annotation of open chromatin (DNase hypersensitivity and FAIRE) data. Karen L. Mohlke designed the study and interpreted results.

Resequencing of 188 individuals was supported by NHLBI's Resequencing and Genotyping program and was performed by Timothy Stockwell, Dana Busam, and Samuel Levy.

ABSTRACT

Recent genome-wide association studies have identified several novel risk loci for plasma lipoprotein levels. One of these loci, *GALNT2*, contains single nucleotide polymorphisms (SNPs) in intron one associated with high density lipoprotein cholesterol (HDL-C) ($P < 5 \times 10^{-8}$). Our goal is to identify functional variant(s) at this locus responsible for the association signal. Because no other genes are located close to the associated SNPs, we hypothesize that a functional variant(s) acts on *GALNT2* or a putative novel intronic transcript. We identified a total of 24 SNPs that are in at least moderate linkage disequilibrium (LD) ($r^2 > .5$) with an HDL-C associated SNP. Additionally, we determined that a 21 bp deletion, rs6143660, was in strong LD ($r^2 = .88$) with an associated SNP and had a minor allele frequency of .36. Of these 25 variants, five (rs2144300, rs4846913, rs6143660, rs17315646, rs4846914) are located within a 1 kb region that overlaps experimental functional annotation of open chromatin (DNase hypersensitivity and FAIRE) and histone modifications (H3K4me3) in a human hepatocellular carcinoma cell line (HepG2), suggesting regulatory function. To test the effects of the haplotypes on transcriptional activity, the 1 kb segment was divided into two separate regions. Each region was cloned in both the forward and reverse orientation into a luciferase reporter vector and transfected into HepG2 cells. Region 1 (rs2144300, rs4846913, rs6143660) demonstrated an approximate two-fold increase in transcriptional activity for one haplotype containing the deletion of rs6143660, the T allele of rs2144300, and the A allele of rs4846913, which have been associated with increased HDL-C levels. The effect was consistent in both a promoterless and a minimal promoter vector in both the forward and reverse orientation ($P < .02$). Region 2 did not show any

allelic differences in transcriptional activity ($P > .05$) in either the forward and reverse orientation in a minimal promoter luciferase reporter vector. These results suggest that the 21 bp deletion, rs4849913, and/or rs2144300 may act to increase the transcriptional activity of *GALNT2* or an unknown novel intronic transcript to increase HDL-C.

INTRODUCTION

Coronary artery disease (CAD) is the leading cause of death in the United States in both men and women [1]. Low levels of high density lipoprotein cholesterol (HDL-C) are associated with increased risk of CAD, thus understanding the basis of inter-individual variation in HDL-C is essential to understanding CAD [2]. Many factors influence HDL-C, including diet, physical activity, and genetics [3]. However, most of the heritability of HDL-C is left to be explained [4].

In recent genome-wide association (GWA) studies of Caucasian populations [5,6], a novel locus in intron one of *GALNT2* was identified to be strongly associated ($P < 9.4 \times 10^{-8}$) with HDL-C levels as well as triglyceride levels. Additionally, an unpublished meta-analysis of approximately 100,000 Caucasian individuals confirms the association with HDL-C levels with *P* values as significant as 2.97×10^{-21} for rs4846914 (Teslovich *et al*., Abstract American Society for Human Genetics 2009). For each additional A allele of the single nucleotide polymorphism (SNP) rs4846914, individuals have an ~1.15 mg/dL increase of HDL-C [5]. However, in the GWA study of the CLHNS women (Chapter IV), *GALNT2* SNPs were not associated ($P > .05$) with HDL-C or triglycerides (index SNP rs4846914, $P = .36$ for HDL-C in the direction of effect previously reported and $P = .34$ for triglycerides in the opposite direction of effect previously reported).

The genomic region around the association signal contains few genes upon which these associated SNPs may be acting (Figure 5.1A). The signal is within intron one of *GALNT2*, N-acetylgalactosaminyltransferase 2. Other genes are relatively far away from the associated region. The nearest genes are *PGBD5* and *KIAA0133*. *PGBD5,* piggyBac

transposable element derived 5, is 148 kb downstream of the association signal. *PGBD5* appears to have been derived from the piggyBac transposons and has no known function. *KIAA0133*, which encodes hypothetical protein LOC9816, is ~407 kb upstream of the 5' untranslated region (UTR) of *GALNT2*. Alternatively, the HDL-C-associated variants may affect an unidentified or more distant transcript.

The association signal spans a moderately small region of at least 15 kilobases (kb) and contains relatively few variants. Of the reported SNPs within *GALNT2*, eight are highly associated with HDL-C levels ($P < 1 \times 10^{-7}$) and are in strong linkage disequilibrium ($r^2 > .8$) with each other [7,8; http://www.sph.umich.edu/csg/abecasis/public/lipids2008/]. These characteristics make this region a prime candidate for identifying the functional variant(s).

There are many potential molecular mechanisms for a functional variant(s). In the simplest scenario, a SNP allele can alter a key amino acid residue that is critical to protein function and activity. However, like many of the recent association signals identified in GWA studies, the associated SNPs in *GALNT2* are located in non-coding sequence, suggesting that non-coding variation plays an important role in common disease. These non-coding SNPs may influence biological processes by reducing transcription factor binding affinity or modifying RNA splicing.

Most DNA is wound around a histone core, forming nucleosomes. When DNA is in this state it is inaccessible to transcription factors, RNA polymerase, or trans-regulatory factors. DNA segments that are actively regulating transcription are characterized by depletion of nucleosomes from the chromatin. The distribution of nucleosomes along DNA can be mapped using DNase I hypersensitivity [9] and

Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE) [10]. Additionally, in eukaryotes modifications of histone tails such as histone H3 trimethylation at lysine 4 (H3K4me3) are associated with active chromatin and gene expression [11]. These modifications can differ between tissues, making it valuable to study the specific tissue in which a gene may act.

To identify additional HDL-C-associated variant(s) that may be responsible for the association signal, we used LD data generated in our laboratory from resequencing the association region as well as data from the 1000 Genomes project (http://browser.1000genomes.org). We then used computational and experimental data of open chromatin and histone modifications in the human hepatocellular carcinoma cell line (HepG2) to identify regions that are highly suggestive of regulatory function. We tested variants in the identified regions using a luciferase reporter assay to identify effects on transcription. One or more variants may act to increase or decrease the transcription of *GALNT2* or an unknown transcript.

MATERIALS AND METHODS

Resequencing *GALNT2* exon 2

Forty-eight unrelated individuals from the Finland-United States Investigation of NIDDM Genetics (FUSION) [12] were resequenced for 538 bp including exon 2, 291 bp upstream and 153 bp downstream. Primers were selected using Primer3 software [13; http://www.genome.wi.mit.edu/genome_software/other/primer3.html]: Forward: 5'-CCCTGGGAGTTTTTGGAGTA-3' and Reverse: 5'-ACTGCTTTGCCAACTTCCAT-3'. Sequencing was performed at the University of North Carolina, Chapel Hill

automated DNA sequencing facility on an ABI Prism 3730 (Applied Biosystems, Foster City, CA, USA) using the Big Dye Terminator Kit. Sequences were compared using Sequencher 4.2.2 (Gene Codes Corporation, Ann Arbor, MI, USA).

Additional sequencing of the *GALNT2* exons and the ~15 kb associated region defined by recombination hotspots was performed on 188 individuals at the J. Craig Venter Institute as part of the NHLBI genotyping and re-sequencing program. The 188 individuals were selected from the extremes of the HDL-C distribution among FUSION individuals without type 2 diabetes and not taking lipid-lowering medications.

Construction of luciferase reporter plasmids

Four individuals from the FUSION samples who were homozygous for the alleles of the desired haplotypes were used as templates for PCR. Primers were selected using Primer3 software. The PCR primers were as follows: Region 1 Forward: 5'-GGCTCTGGCAAAGTGTCTTG-3' Reverse: 5'-TGAATTTCTCCGGTTGACCT-3', Region 2 Forward: 5'-TTAGTTGAGGATCAGATGTGTCA-3' Reverse: 5'-TCAGTGAGCAGAAACTAAGGACA-3'. The primers were tailored to incorporate a Kpn I site and an Xho I site at the end of the amplified regions. Fragments were inserted in both the forward and reverse orientation into the vector.

The amplified regions were purified with the Promega Wizard SV Gel and PCR CleanUp System (Promega Corporation, Madison, WI, USA), digested with Xho I and Kpn I restriction endonucleases (New England Biolabs, Ipswich, MA, USA), and subcloned into the Kpn I and Xho I sites of firefly luciferase-expressing pGL4 vectors (Promega Corporation). Region 1 was subcloned into the basic pGL4.10 and the minimal

promoter pGL4.23 vectors. Region 2 was subcloned into the minimal promoter pGL4.23 vector. All constructs were sequenced to verify nucleotides at variant positions and to identify any additional variant nucleotides in the region.

Cell cultures, transfections, and luciferase assays

Human HepG2 cells were grown in Dulbecco's modified Eagle medium with 10% fetal bovine serum, and 1% sodium pyruvate at 37°C and 5% CO_2. Cells were seeded at 8 x 10^4 to 10 x 10^4 cells per well in a 24-well collagen-coated plate and allowed to grow undisturbed for 24 hours prior to transfection. Transient transfections were performed with FuGENE6 (Roche Applied Science, Indianapolis, IN, USA) according to the manufacturer's protocol. Specifically, 720 ng of the luciferase reporter construct was cotransfected with 80 ng of Renilla luciferase (Promega Corporation) to control for variation in transfection efficiency. Each clone and the corresponding empty vector were transfected in duplicate.

After 48 hours, the transfected cells were lysed with 90 ul of 1x lysis buffer for 45 minutes by gently rocking at room temperature. Fifteen microliters of cell lysate was harvested and luciferase activity was measured with Dual-Luciferase Reporter Assay System (Promega Corporation) using a 96-well microplate luminometer.

Statistical analysis

Relative luciferase activity is reported as luciferase divided by Renilla relative to empty vector. Results are represented at ± standard deviation of the 2 to 7 independent clones. For Region 1, differences between haplotypes in the luciferase assay data were

analyzed by ANOVA and a Tukey post hoc pairwise analysis between haplotypes. For Region 2, luciferase assay data were assessed by a two-sided Student's t-test. A $P < .05$ was considered statistically significant.

Genotyping rs6143660 insertion-deletion variant

Eighty-seven random FUSION samples were chosen for genotyping. Primers for amplification were selected using Primer3 software: Forward: 5'-CTCATCTTTGCACACGAAGG-3' Reverse: 5'-GAGACCCTGAGTGTGAGGCT-3'. The products for chromosomes with and without the deletion were 91 bp and 112 bp, respectively. The products were run on a 3% low melting point agarose gel, and two researchers scored genotypes independently.

RESULTS

Identifying possible functional variants

To identify the full set of possible functional SNPs in the *GALNT2* HDL-C association region (chr1:228,360,704-228,375,798) (Figure 5.1B), we used four approaches to perform resequencing and evaluate LD compared to the known HDL-C associated SNPs (Table 5.1). First, to identify novel coding or splice site variants, I performed targeted resequencing of *GALNT2* exon 2 located downstream of the 15 kb region of HDL-C associated SNPs. No sequence variations were found in the exon. By using 48 FUSION samples, I had 99% probability of detecting a variant with a minor allele frequency >.05. Second, additional resequencing of the association region was performed in 188 individuals. Eighty percent of the region was successfully resequenced,

and successful amplicons are pictured in Figure 5.1B. Nine SNPs were identified to be in moderate LD (r^2>.5) with one of the eight original associated SNPs. Of these nine, one had been previously tested for association in the Kathiresan *et al.* 2009 GWA study and had a *P* of 3.1 x 10^{-5}. Third, further examination of HapMap (http://www.hapmap.org) CEU variants in strong LD (r^2 = 1) with one of the eight original associated SNPs identified two additional SNPs. These two SNPs were not analyzed in the Kathiresan *et al.* 2009 GWA study because they failed quality control and were not identified by resequencing because they were located in an amplicon that failed. Finally, using preliminary data from the 1000 Genomes Project (May 2009; http://browser.1000genomes.org) we identified five SNPs that were in moderate LD (r^2>.5) with the best SNP reported by Kathiresan *et al.* 2009, rs4846914. In total, 24 SNPs were identified through resequencing and LD analysis to be potential functional SNPs [7]. Detailed descriptions of this set are described in Table 5.1.

Selection of regions with potential regulatory function

Of the 24 SNPs identified in the association region, four SNPs are located in genomic regions that are highly suggestive of regulatory function (rs2144300, rs4846913, rs17315646, and rs4846914). These SNPs are located on the furthest upstream side of the association region, within 1 kb (976bp) of one another, and overlap many functional annotations including that of open chromatin (DNase HS, FAIRE), histone modification (H3K4me3), and chromatin immunoprecipitation (ChIP) with regulatory proteins, Sterol regulatory element-binding protein-1A and RNA polymerase II, in HepG2 cells (Figure 5.1B and 5.1C). These SNPs thus represent the best candidates at this locus.

This 976 bp section was divided based on the functional annotations into two regions to test in the luciferase expression reporter assay (Figure 5.1C). Region 1 is 780 bp and includes seven variants, five of which are common (minor allele frequency > .05). Region two is 565 bp and includes five variants, three of which are common. Fifty base pairs of Region 1 was predicted to fall within a possible promoter by Berkeley Drosophila Genome Project Eukaryote Promoter Prediction (http://www.fruitfly.org/seq_tools/promoter.html). Therefore, this region was tested in both pGL4.10, a promoterless vector, and pGL4.23, a minimal promoter vector. Three and two haplotypes were tested for functional activity in Region 1 and 2, respectively (Table 5.2 and 5.3).

Analyzing the 21 nucleotide deletion - rs6143660

During the resequencing of the reporter constructs, a 21 nucleotide deletion, rs6143660, was found to be common. To further evaluate this variant, we genotyped the deletion in 87 individuals with known genotypes at HDL-C-associated variants. We found that the MAF was .36 and the r^2 with the HDL-C-associated SNP rs4846914 was .88. While we did not directly test the deletion for association with HDL-C levels, the correlation of it with the strongest associated SNP suggests that this deletion may also be associated with HDL-C levels.

Allele-specific effect of Haplotype 1 on transcriptional activity in Region 1

The genomic fragments from Region 1 with alleles specific to the haplotypes described in Table 5.2 were inserted into the firefly luciferase expressing pGL4.10 basic

promoterless vector or pGL4.23 minimal promoter vector. The activity of the constructs was assessed by transiently transfecting HepG2 cells.

The luciferase activity of Haplotype 1 was 1.6 to 3-fold higher than that of the other haplotypes in the basic vector, in both the forward and reverse orientations (*P* overall < .0001, *P* overall = .0009; Figure 5.2A). We performed a Tukey post-hoc test to access the differences between individual haplotypes; the *P* values for the forward orientation were *P* < .0001, *P* < .0001, and *P* = .0009 for Haplotype 1 to 2, 1 to 3, and 2 to 3, respectively. For the reverse orientation, the *P* values were *P* = .02, *P* = .0007, and *P* = .2 for Haplotype 1 to 2, 1 to 3, and 2 to 3, respectively.

In the minimal promoter vector, the luciferase activity of Haplotype 1 was also 1.8 to 3-fold higher than that of the other haplotypes, in both the forward and reverse orientations (*P* overall < .0001, *P* = .0001; Figure 5.2B). The *P* values for the forward orientation were *P* < .0001, *P* < .0001, and *P* = .03 for Haplotype 1 to 2, 1 to 3, and 2 to 3, respectively. For the reverse orientation, the *P* values were *P* = .0008, *P* = .0002, and *P* = .6 for Haplotype 1 to 2, 1 to 3, and 2 to 3, respectively.

No allele-specific effects in Region 2

The genomic fragments from Region 2 with alleles specific to the haplotypes described in Table 5.3 were also tested for luciferase activity in the minimal promoter, pGL4.23, vector. The luciferase activity demonstrated no difference between the two haplotypes in either the forward or reverse orientation (*P* = .06, *P* = .1; Figure 5.3).

DISCUSSION

HDL-C is a complex trait that can be affected by multiple genetic and environmental factors [3]. In the present study, we explored the functional potential of some of the alleles that may be underlying the HDL-C association signal at the *GALNT2* locus by investigating the transcriptional effects of two genomic regions. We show that the SNPs in Haplotype 1 of Region 1 consistently cause a significant increase in transcription compared to Haplotype 2 and Haplotype 3 in both a minimal promoter vector and a basic vector in the forward and reverse orientation. The results indicate increased transcriptional activity of Haplotype 1 containing the 21 bp deletion of rs6143660, the T allele of rs2144300, and the A allele of rs4846913. These alleles have been associated with increased HDL-C levels and may be the potential functional variants causing Haplotype 1 to have an increase in transcriptional activity. Haplotype 2 shows a significant increase in transcription compared to Haplotype 3 in the forward direction of both vectors but does not show an increase in the reverse direction. These results suggest that the A allele of rs1555290, which differentiates Haplotype 2 from Haplotype 3, may also be a potentially functional variant with a modest effect.

This study is the first known attempt to pinpoint the functional variant(s) contributing to the HDL-C association at the *GALNT2* locus. Based on these results, one or more of the alleles in Region 1 appear to increase transcription of *GALNT2* or another transcript. The luciferase activity for Region 1 is consistently higher than the empty vector, with relative luciferase activity as high as 48-fold in the minimal promoter vector in the forward direction and activity nearly as high in the promoterless vector. The luciferase data as well as the open chromatin data suggest that Region 1 may contain an

enhancer or promoter. If the potentially functional variants act directly on *GALNT2* then increased GALNT2 may lead to increased HDL-C. On the other hand, these variants may also or instead increase transcription of an unknown transcript that leads to increased HDL-C.

Our results should be interpreted cautiously in light of a recent abstract by Edmondson *et al.* (Abstract American Society of Human Genetics 2009), which contradicts the direction of our results. Edmondson overexpressed mouse and human *GALNT2* in mouse hepatocytes, which led to a significant decrease in HDL-C by approximately 20% at 28 days. Additionally, the knockdown endogenous *Galnt2* lead to a 37% increase in HDL-C. These results suggest that *GALNT2* is the casual gene at this locus and that changes in hepatic *GALNT2* expression are associated with inverse changes in HDL-C levels.

While there are no established connections between the function of *GALNT2* and HDL-C levels, the localized association signal suggests *GALNT2* may be the most likely gene affecting levels of HDL-C. *GALNT2* codes for a widely expressed N-actetylgalactosaminyltransferase, which is involved in O-linked glycosylation of proteins. GALNT2 is located on the trans side of the golgi stack, and is known to transfer N-acetylgalactosamine to serine or threonine residues [14]. Because O-linked glycosylation can regulate protein function, changes in *GALNT2* expression may indirectly affect HDL-C levels through glycosylating and modifying a lipoprotein or receptor involved in metabolism. Potential target candidates may include LCAT, apoCIII, VLDLR, or LDLR because these have been reported to be O-glycosylated with N-acetylgalactosamine residues [15-18].

The luciferase assay experimental approach has several limitations. Luciferase reporter assays are an artificial *in vitro* system that may not represent fully all the interacting nuances in the human body. While *GALNT2* is expressed in liver tissue [14] and liver is very critical to HDL-C metabolism, GALNT2 may not have its functional effects on HDL-C within the liver. Additionally, we are using HepG2 cells as a proxy for *in vivo* liver tissues, and these cells may not contain all of the regulatory proteins biologically relevant to human HDL-C metabolism.

While Regions 1 and 2 of the association region seem to be the most obvious for a predicted regulatory element, they do not contain all possible functional alleles, and we may have not identified the full set of possible functional alleles. Through the large scale resequencing and data from the 1000 Genomes Project, we identified 16 other SNPs and one deletion. The likelihood that we have identified the full set of possible functional alleles depends on the unknown frequency of the underlying functional variant(s).

The functional activity of the variants in Region 1 can be analyzed in future experiments. The variants could be tested in other biological relevant cell lines, such as primary liver cells or muscle-derived cell lines. In addition, the three variants on Haplotype 1 in Region 1 could be isolated to determine if a single variant affects transcriptional activity. Because these SNPs are in high linkage disequilibrium, only very rare individuals would have the needed haplotypes.; we could perform site-directed mutagenesis to separate the alleles. Region 1 might also be tested as smaller sub-regions to isolate the SNP effects.

In addition to the luciferase assays, there are other methods to determine potential functional alleles. One method is an electrophoretic mobility shift assay. This method

allows for the study of protein-DNA interactions to identify if one allele differentially binds a specific protein or any protein from cell lystates. The future directions include testing other variants in the association signal that are highly associated or that fall within in a predicated regulatory region. We have the potential to test rare variants within *GALNT2* that are specific to high or low HDL-C individuals and test to see if these variants perturb function. Additionally, it would be helpful to investigate potential proteins that may be modified by GALNT2 and identify the pathway that effects HDL-C metabolism.

In summary, we found that constructs containing the 21 bp deletion of rs6143660, the T allele of rs2144300, and the A allele of rs4846913 increased transcriptional activity compared to the alternative alleles. Nonetheless, additional research will be required to better define the functional significance of these variants and to clarify the mechanism of *GALNT2* with HDL-C. Ultimately, by identifying the causative variant(s) at this locus, we may aid in the discovery of therapeutic targets for altering HDL-C levels.

Table 5.1: Candidate functional SNPs with annotation in the *GALNT2* HDL-C associated locus.

SNP	Source	*P* value for HDL-C association[1]	r^2 with an original HDL-C associated SNP	Potential regulatory functional annotation	Tested Region
rs2144300	Original HDL-C associated SNP[1]	1.03E-07		HepG2 DNase HS, FAIRE, PolII, SREBP1A, H3K4me3	1
rs17315646	Original HDL-C associated SNP[1]	7.24E-08		HepG2 DNase HS, H3K4me3	2
rs4846914	Original HDL-C associated SNP[1]	3.91E-08		HepG2 FAIRE, H3K4me3	2
rs10127775	Original HDL-C associated SNP[1]	6.86E-08			2
rs2281719	Original HDL-C associated SNP[1]	1.02E-07			
rs10779835	Original HDL-C associated SNP[1]	5.25E-08		HepG2 H3K4me3	
rs10489615	Original HDL-C associated SNP[1]	9.40E-08			
rs1321257	Original HDL-C associated SNP[1]	8.23E-08			
rs2281721	188 individual re-sequencing[2]		1		
rs10864727	188 individual re-sequencing[2]		1		
rs10864728	188 individual re-sequencing[2]		1		
rs11122456	188 individual re-sequencing[2]		1		
rs4846921	188 individual re-sequencing[2]		.989		
rs2281718	188 individual re-sequencing[2]		.966		
rs4846922	188 individual re-sequencing[2]		.915	Most conserved mammal and vertebrate	
rs609526	188 individual re-sequencing[2]	3.10E-05	.729		
rs4846923	188 individual re-sequencing[2]		.616		
rs11122450	HapMap[3]		1		
rs4846913	HapMap[3]		1	HepG2 DNase HS, FAIRE, PolII, SREBP1A	1
rs34996149	1000 Genomes Project[4]		.9604		
rs11122454	1000 Genomes Project[4]		.9604		
rs1546954	1000 Genomes Project[4]		.6079		
rs11122453	1000 Genomes Project[4]		.6766		
rs10864726	1000 Genomes Project[4]		.6079		

1. Definition and *P* values from Kathiresan *et al.* 2009
2. Linkage disequilbrium (LD) was obtained from K. Gaulton *et al*. unpublished
3. LD determined by HapMap CEU population
4. LD with rs4846914 determined by data from May 2009 of 1000 Genomes Project data

Table 5.2: Haplotypes tested in Region 1

	rs4846913*	rs2144300*	rs1555290	rs6143660*	Association of alleles
Haplotype 1:	A	T	A	-	Increase HDL-C
Haplotype 2:	C	C	A	+	Decrease HDL-C
Haplotype 3:	C	C	C	+	Decrease HDL-C

* Direction of association was based upon these SNPs.

Table 5.3: Haplotypes tested in Region 2

	rs17315646	rs4846914	rs10127775	Association of alleles
Haplotype 1:	C	G	A	Decreases HDL
Haplotype 2:	G	A	T	Increases HDL

Figure 5.1: Evidence for association with HDL-C and potential regulatory regions at the *GALNT2* locus

A. Evidence for association with high density lipoprotein cholesterol (HDL-C) at the *GALNT2* locus. Evidence for association is shown for evaluated single nucleotide polymorphisms (SNPs) in 19,794 GWA samples [7]. The top panel depicts the locations of genes.

B. Zoom (chr1: 228,360,704-228,375,798) on the region of association in intron one with experimental and computationally predicted evidence of regulatory regions. The first track indicates single nucleotide polymorphisms (SNP) within the region. The red ovals indicate SNPs evaluated for association with $P < 1.02 \times 10^{-7}$ [7]. Predicted regions of open chromatin are indicated by DNase hypersensitivity and formaldehyde-assisted identification of regulatory elements using the human hepatocellular carcinoma HepG2 cell line. Additional potential regulatory predictors include markers of histone modification (H3K4me3), chromatin immunoprecipitation (ChIP) with regulatory proteins, and conservation regions. The bottom track indicates the successful amplicons of the 188 individual re-sequencing.

C. Zoom on Region 1 and Region 2 with experimental and computationally predicted evidence of regulatory regions.

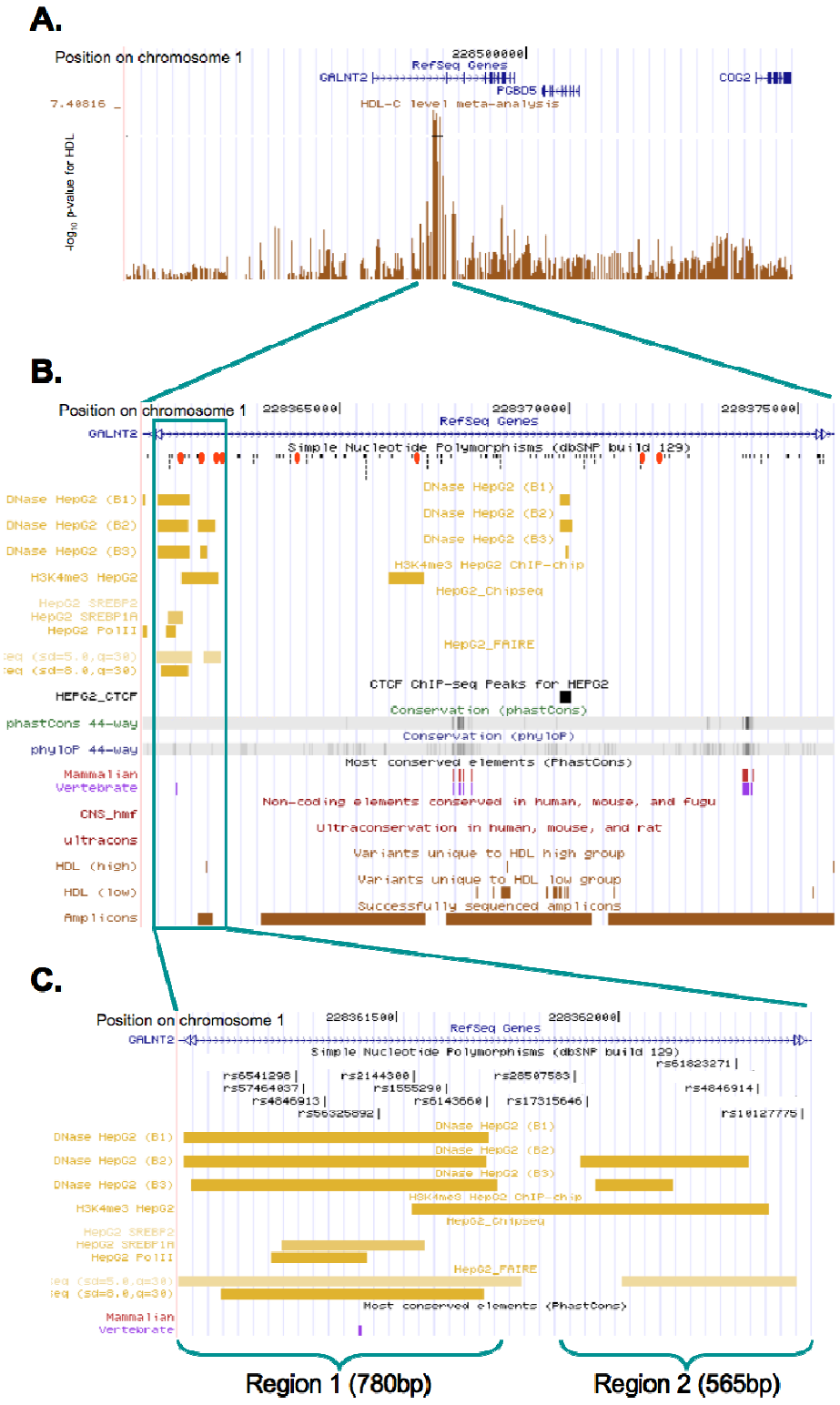
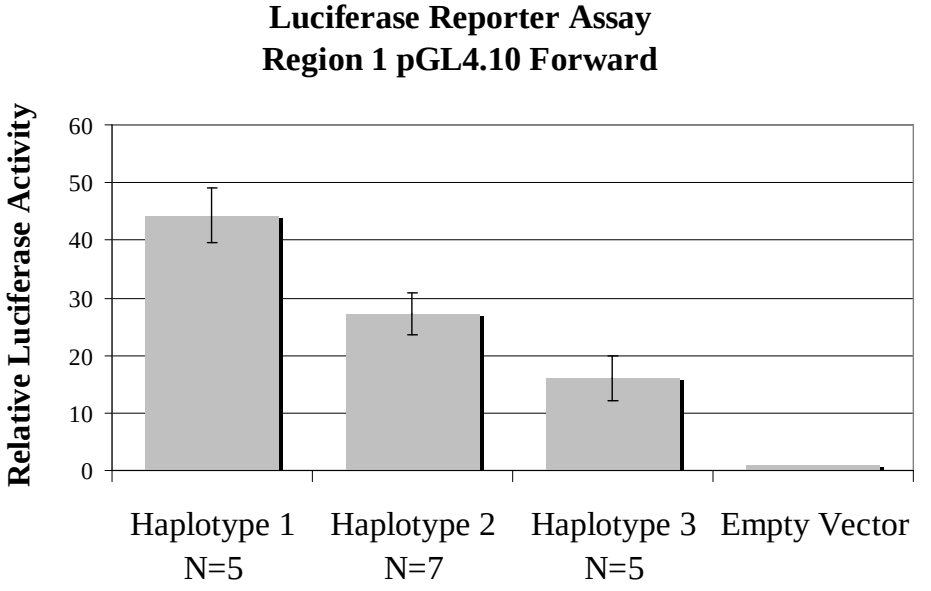
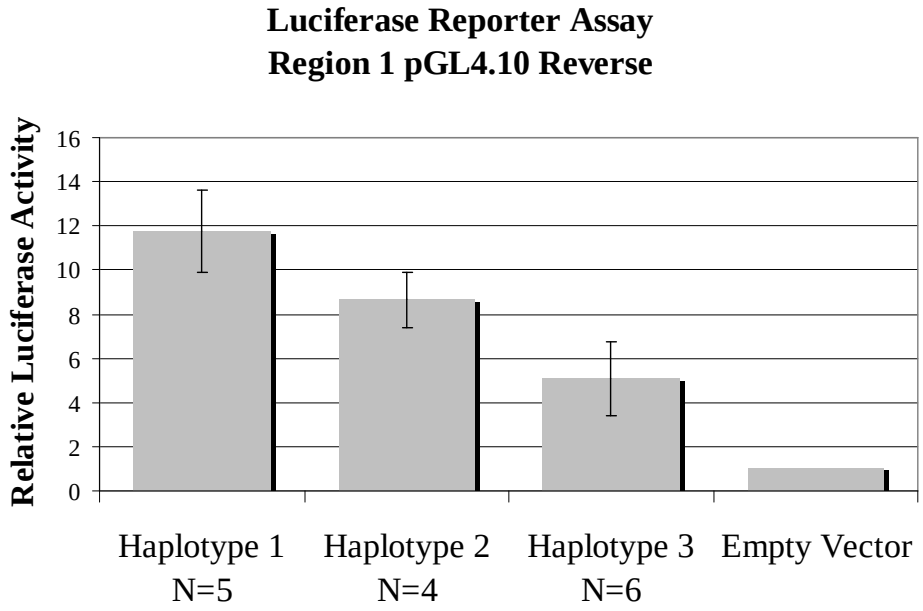

A.
Position on chromosome 1
228500000
RefSeq Genes
GALNT2
PGBD5
COG2
7.40816
HDL-C level meta-analysis
-log10 p-value for HDL
B.
Position on chromosome 1
228365000
228370000
228375000
RefSeq Genes
GALNT2
Simple Nucleotide Polymorphisms (dbSNP build 129)
DNase HepG2 (B1)
DNase HepG2 (B2)
DNase HepG2 (B3)
H3K4me3 HepG2 ChIP-chip
H3K4me3 HepG2
HepG2_Chipseq
HepG2 SREBP2
HepG2 SREBP1A
HepG2 PolII
HepG2_FAIRE
HEPG2_CTCF
CTCF ChIP-seq Peaks for HEPG2
Conservation (phastCons)
phastCons 44-way
Conservation (phyloP)
phyloP 44-way
Most conserved elements (PhastCons)
Mammalian
Vertebrate
Non-coding elements conserved in human, mouse, and fugu
CNS_hmf
Ultraconservation in human, mouse, and rat
Ultracons
Variants unique to HDL high group
HDL (high)
Variants unique to HDL low group
HDL (low)
Successfully sequenced amplicons
Amplicons
C.
Position on chromosome 1
228361500
228362000
RefSeq Genes
GALNT2
Simple Nucleotide Polymorphisms (dbSNP build 129)
rs61823271
rs6541298
rs2144300
rs28507583
rs57464037
rs1555290
rs4846914
rs4846913
rs6143660
rs17315646
rs56325892
rs10127775
DNase HepG2 (B1)
DNase HepG2 (B2)
DNase HepG2 (B3)
H3K4me3 HepG2 ChIP-chip
H3K4me3 HepG2
HepG2_Chipseq
HepG2 SREBP2
HepG2 SREBP1A
HepG2 PolII
HepG2_FAIRE
Most conserved elements (PhastCons)
Mammalian
Vertebrate
Region 1 (780bp)
Region 2 (565bp)

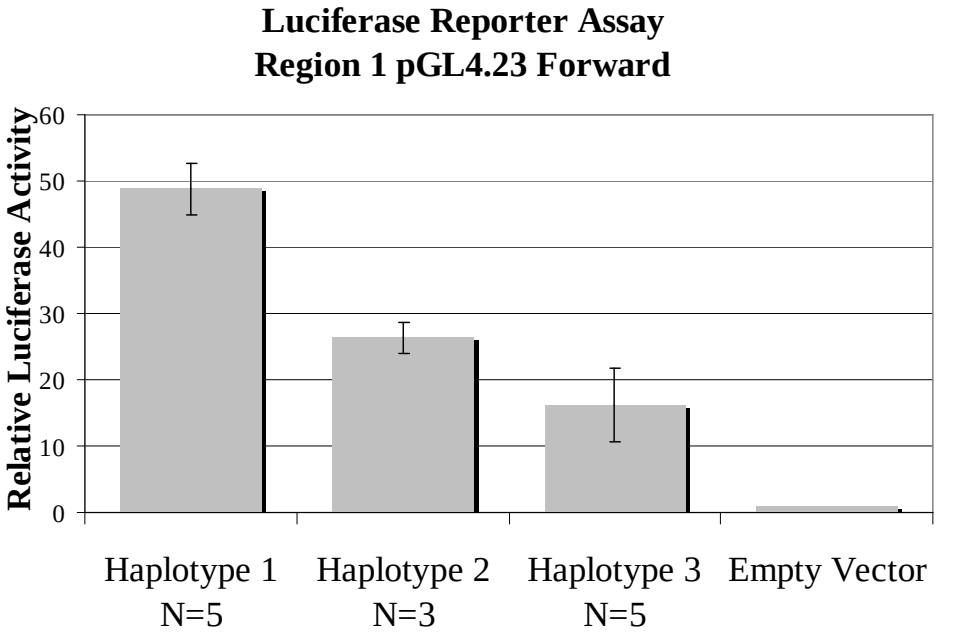
Figure 5.2: Allelic differences in transcriptional activity between haplotypes of Region 1 in the human hepatocellular carcinoma HepG2 cell line. Relative luciferase activity (firefly luciferase/Renilla luciferase) was determined and fold luciferase activity is presented as compared to vector without insert. Error bars indicated ± standard deviation of N independent biological replicate clones. **A.** Luciferase activity was tested in the basic, pGL4.10, vector. **B.** Luciferase activity was tested in the minimal promoter, pGL4.23, vector.

A





B


Luciferase Reporter Assay
Region 1 pGL4.23 Forward
Relative Luciferase Activity
60
50
40
30
20
10
0
Haplotype 1
N=5
Haplotype 2
N=3
Haplotype 3
N=5
Empty Vector

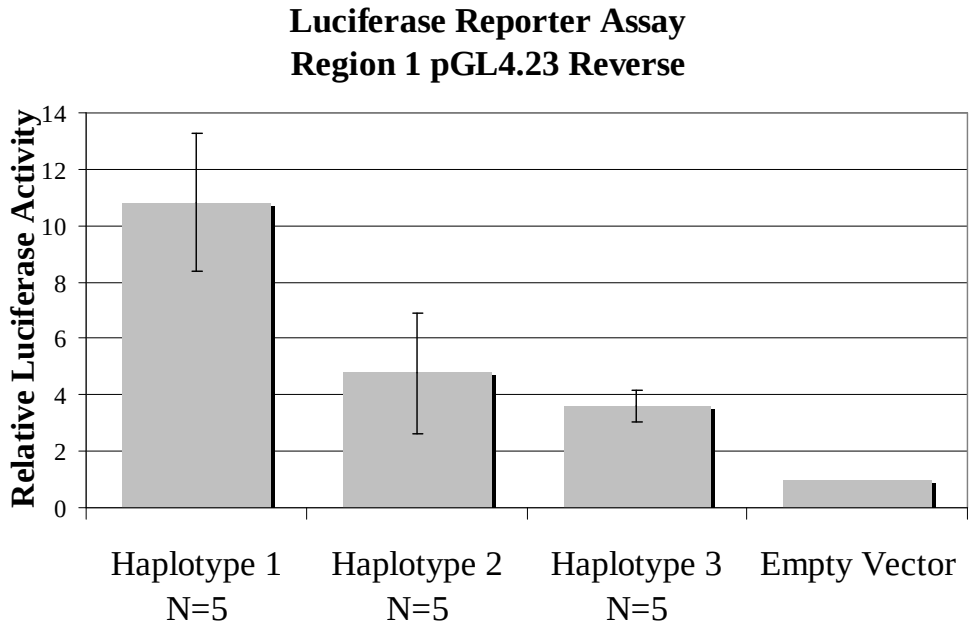
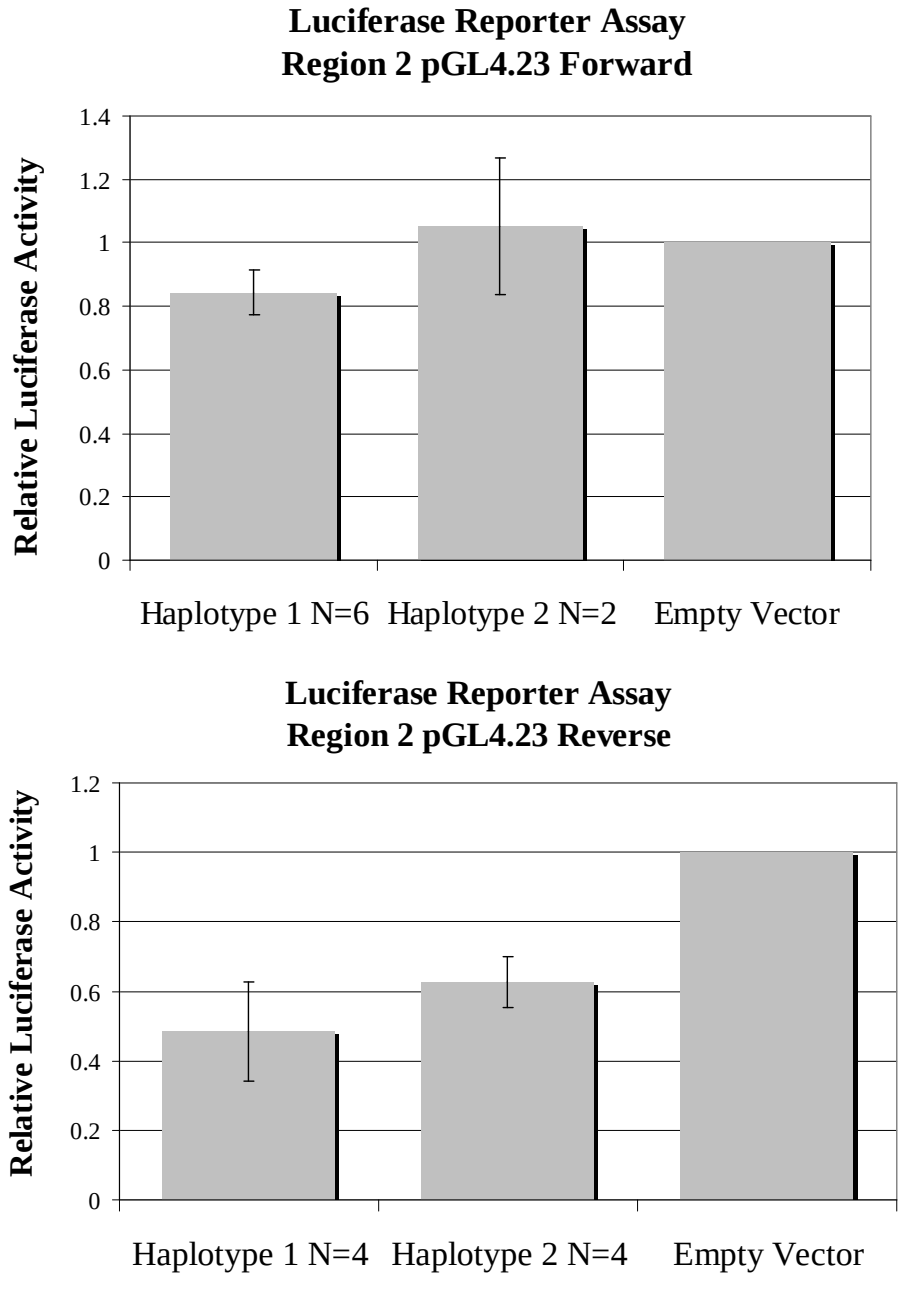

Luciferase Reporter Assay
Region 1 pGL4.23 Reverse
Relative Luciferase Activity
14
12
10
8
6
4
2
0
Haplotype 1
N=5
Haplotype 2
N=5
Haplotype 3
N=5
Empty Vector

Figure 5.3: Allelic differences in transcriptional activity between haplotypes of Region 2 in the human hepatocellular carcinoma HepG2 cell line. Relative luciferase activity (firefly luciferase/Renilla luciferase) was determined and fold luciferase activity is presented as compared to vector without insert. Error bars indicated ± standard deviation of N independent biological replicate clones. Luciferase activity was tested in the minimal promoter, pGL4.23, vector.



REFERENCES

1. Lopez AD, Mathers CD, Ezzati M, Jamison DT, Murray CJ (2006) Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data. Lancet 367: 1747-1757.

2. Gotto AM, Jr., Brinton EA (2004) Assessing low levels of high-density lipoprotein cholesterol as a risk factor in coronary heart disease: a working group report and update. J Am Coll Cardiol 43: 717-724.

3. O'Connell DL, Heller RF, Roberts DC, Allen JR, Knapp JC, et al. (1988) Twin study of genetic and environmental effects on lipid levels. Genet Epidemiol 5: 323-341.

4. Mohlke KL, Boehnke M, Abecasis GR (2008) Metabolic and cardiovascular traits: an abundance of recently identified common genetic variants. Hum Mol Genet 17: R102-108.

5. Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, et al. (2008) Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet 40: 161-169.

6. Kathiresan S, Melander O, Guiducci C, Surti A, Burtt NP, et al. (2008) Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. Nat Genet 40: 189-197.

7. Kathiresan S, Willer CJ, Peloso GM, Demissie S, Musunuru K, et al. (2009) Common variants at 30 loci contribute to polygenic dyslipidemia. Nat Genet 41: 56-65.

8. International HapMap Constortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851-861.

9. Crawford GE, Holt IE, Mullikin JC, Tai D, Blakesley R, et al. (2004) Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. Proc Natl Acad Sci U S A 101: 992-997.

10. Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD (2007) FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res 17: 877-885.

11. Gilbert N, Ramsahoye B (2005) The relationship between chromatin structure and transcriptional activity in mammalian genomes. Brief Funct Genomic Proteomic 4: 129-142.

12. Valle T, Tuomilehto J, Bergman RN, Ghosh S, Hauser ER, et al. (1998) Mapping genes for NIDDM. Design of the Finland-United States Investigation of NIDDM Genetics (FUSION) Study. Diabetes Care 21: 949-958.

13. Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132: 365-386.

14. Rottger S, White J, Wandall HH, Olivo JC, Stark A, et al. (1998) Localization of three human polypeptide GalNAc-transferases in HeLa cells suggests initiation of O-linked glycosylation throughout the Golgi apparatus. J Cell Sci 111 (Pt 1): 45-6.

15. Goldstein JL, Brown MS, Anderson RG, Russell DW, Schneider WJ (1985) Receptor-mediated endocytosis: concepts emerging from the LDL receptor system. Annu Rev Cell Biol 1: 1-39.

16. Schindler PA, Settineri CA, Collet X, Fielding CJ, Burlingame AL (1995) Site-specific detection and structural characterization of the glycosylation of human plasma proteins lecithin:cholesterol acyltransferase and apolipoprotein D using HPLC/electrospray mass spectrometry and sequential glycosidase digestion. Protein Sci 4: 791-803.

17. Vaith P, Assmann G, Uhlenbruck G (1978) Characterization of the oligosaccharide side chain of apolipoprotein C-III from human plasma very low density lipoproteins. Biochim Biophys Acta 541: 234-24.

18. Magrane J, Casaroli-Marano RP, Reina M, Gafvels M, Vilaro S (1999) The role of O-linked sugars in determining the very low density lipoprotein receptor stability or release from the cell. FEBS Lett 451: 56-62.

CHAPTER VI. CONCLUSIONS

OVERVIEW OF FINDINGS AND SYNTHESIS

This dissertation identifies variants that are associated with obesity-related traits and lipoprotein levels in the Cebu Longitudinal Health and Nutrition Study (CLHNS), a cohort of adult women from the Philippines. One of the loci identified to influence the interindividual variation of high density lipoprotein cholesterol (HDL-C) was further investigated to detect the functional variant(s) that are contributing to the association. Together, these findings represent important contributions to the field of human genetics and will serve as building blocks for future studies. While each chapter of this dissertation can be seen as a standalone paper, the analyses and findings fit together to form a cohesive picture of genetic risk factors for cardiovascular disease (CVD). This discussion highlights key points from and relationships between each chapter, the significance of the research, and explores future endeavors, which may prove beneficial in expanding the understanding of complex traits.

Chapter II: Comparison of ENCODE region SNPs between Cebu Filipino and Asian HapMap samples

Prior to beginning this study, limited genetic data was available on Filipinos. The detailed population study described in **Chapter II** was motivated by our plan to select genetic polymorphisms representing common variation across the genome for future CLHNS studies. We, therefore, set out to assess the effectiveness of using the Hapmap data to guide single nucleotide polymorphism (SNP) selection for the Cebu Filipino CLHNS samples.

In order to determine whether a large number of unknown variants would be identified, we resequenced a portion of each of the ten ENCODE regions [1] revealing only one SNP (minor allele frequency, MAF > .05) unique to the CLHNS sample and not present in dbSNP or HapMap suggesting that SNPs ascertained from the ENCODE regions are representative of common variation in Cebu Filipinos. From this result, we concluded that no further resequencing would be needed for typical candidate gene studies of common (MAF > .05) variants. Nonetheless, in **Chapter V**, we resequenced exon 2 of *GALNT2* in order to identify all potential variants that could be contributing to the association signal.

We were also interested in identifying which of four HapMap populations (CEU, CHB, JPT, YRI) [2] would be most closely related to the CLHNS. The identified population could be used for imputation to infer genotypes of un-genotyped SNPs in the genome wide association (GWA) study conducted in **Chapter IV**. Consistent with major population migration patterns, mitochondrial DNA, and Y haplotype patterns, CHB, JPT, and CHB+JPT HapMap samples were more similar with respect to allele frequency than the CEU and YRI samples. Based on estimates of pairwise r^2 and haplotype frequencies of the Asian HapMap samples, the CHB and CHB+JPT panels were used as an efficient proxy when choosing haplotype tagging SNPs for CLHNS samples. Despite Spanish presence in the Philippines from the late 16th century to late 19th century, these analyses did not detect a significant genetic contribution from the Spanish in the individuals examined from Cebu. The individuals were most similar to the other Asian samples.

Because the ENCODE regions represent a range of measures of linkage disequilibrium (LD), gene density, and haplotype blocks, the results gleaned from these

ten regions may apply, on average, across the genome. These data are consistent with the likelihood that Filipinos from Cebu would not show dramatically different LD compared to individuals from China and Japan, suggesting that SNP selection based on CHB and JPT HapMap samples is satisfactory. These findings were useful for the design and analysis of genetic studies in the Cebu Filipino population, as presented in **Chapter IV**.

Chapter III: Association of FTO with obesity-related traits in the Cebu Longitudinal Health and Nutrition Survey (CLHNS) cohort

Leading up to the work for this chapter, the underlying genetic components of obesity-related traits were not well understood, and there was limited evidence to support genetic association shared across multiple studies, populations, and environmental contexts. During this time GWA studies were just beginning, and most of the prior focus was on candidate gene studies. As of 2005, the Human Obesity Gene Map reported 426 statistically significant associations between obesity-related phenotypes and a DNA variant in 127 candidate genes. Of these genes, 22 were supported by positive replication in at least five studies [3]. However, additional reports existed for these loci that were inconsistent with preliminary data. Using the well-characterized CLHNS samples from an understudied population provided us with a unique data set to analyze association of variants with previous evidence of association with an obesity-related quantitative trait in other populations. The longitudinal anthropometry, diet, and physical activity measures provided an outstanding resource to analyze cross-sectional endpoint and longitudinal traits and test for interaction effects between SNPs and diet and physical activity.

We selected and genotyped 19 single nucleotide polymorphisms in ten genes (*ADRB2, ADRB3, FTO, GNB3, INSIG2, LEPR, PPARG, TNF, UCP2,* and *UCP3*) that had the strongest prior evidence (at the time) of being associated with an obesity-related quantitative trait in other populations. Two of the variants were chosen based on GWA studies: the *FTO* variant rs9939609 and the *INSIG2* variant rs7566605. The other variants were identified from candidate gene association studies.

Of the SNPs tested, only the A allele of *FTO* variant rs9939609 showed significant association in the same direction as seen in other previously reported populations. We observed evidence of association with body mass index (BMI), waist circumference (measurement of central adiposity), and weight, but not with skinfold thicknesses (measurements of subcutaneous adiposity). These data suggest that variants in *FTO* influence central adiposity to a greater extent than subcutaneous fat. We also observed evidence for association with longitudinal BMI, which reflects a relatively constant genotype effect over 22 years and strengthens the evidence that this locus influences BMI in this population. We observed an effect even with baseline BMI when the women were young and had with smaller BMIs, consistent with an effect over a wide age range, which is further supported by previous reports of an *FTO* effect in children [4,5]. There was no evidence for interaction between genotype and either the 2005 dietary intake (total calories, estimated percent diet from fat, and estimated percent diet from carbohydrates) or physical activity measures. This study replicates this evidence in a Filipino population suggesting *FTO* may be important in many genetic backgrounds.

Additionally, we observed evidence of association with the Trp64 allele of rs4994 in *ADRB3* with increased weight, percent fat mass, arm fat area (AFA), arm muscle area

(AMA), and longitudinal BMI, but not with baseline BMI. However, the direction of effect was not consistent with the majority of previous reports suggesting that these results should be interpreted with caution.

There have been many GWA studies published subsequent to this paper that have examined obesity-related traits [6-14], and to date, of the 19 tested variants, SNPs in *FTO* have the most consistent and strongest evidence for association with obesity-related traits; the variant rs9939609 within *FTO* has been identified to be associated with obesity-related traits in these GWA studies at a genome wide significance threshold ($P < 10^{-8}$). Variants within *FTO* have evidence of association with many measures of obesity in both adults and children and across diverse populations. One of the reasons for the consistent association may be that variants in *FTO* have one of the strongest effect-sizes observed for obesity-related traits. The common associated variants in intron 1 of *FTO* have an effect of ~3 kilograms for the homozygotes of the risk allele compared to the homozygotes of the non-risk allele.

FTO encodes a 2-oxoglutarate–dependent nucleic acid demethylase [15] and may play a role in DNA or RNA repair [16]. FTO is expressed in several tissues including adipose, pancreatic islets, skeletal muscle, and hypothalamus suggesting a role in regulation of appetite [4,5,15,17]. Recently a study found that Fto-deficient mice have growth retardation and reduced adipose and lean body mass caused by increased energy expenditure. This study suggests *FTO* may play a role in energy homeostatis by regulating energy expenditure [18]. However, the exact variant(s) within *FTO* as well as the exact pathway through which *FTO* is influencing obesity have yet to be identified. This uncertainty is true for many loci identified to be associated with obesity-related

traits and lipoproteins. I have attempted to identify which variant(s) is involved within the HDL-C association signal in **Chapter V** when investigating variants within GALNT2 that are associated with HDL-C levels.

The inability of our study and recent GWA studies to corroborate associations of the other 18 SNPs identified in previous genetic studies of obesity-related traits demonstrates the difficultly of identifying genes using a candidate approach and suggests that there may be variable effects across populations. While variants in the genes from candidate gene studies have not reached the stringent GWA threshold of 5×10^{-8} in recent GWA studies, these genes may have less significant associations within these studies and may still be contributing to obesity-related traits. As more consortiums are formed and GWA studies become larger, these genes may be identified. However, it is notable that some genes involved in lipoprotein levels (**Chapter IV**) were successfully identified in candidate gene studies and have consistent evidence of association in recent lipid GWA studies. These genes include: *APOB, PCSK9, ABCA1, ANGPTL4, HMGCR, LDLR, APOA, CETP, LIPC, LIPG, LPL, GCKR,* and *APOE* [19]. These genes and many others were examined for association in **Chapter IV**.

The differing association patterns of variants in *INSIG2* between studies illustrate that SNPs may have variable effects in different populations. The initial association was identified in a GWA study and has been replicated [20-24] in some but not all cohorts [25-37]. Recently, I was involved in a GWA study of obesity-related traits in the CLHNS lead by D.C. Croteau-Chonka [38]. This study has replicated several, but not all, of the findings from current GWA studies for both BMI and weight with *MC4R* and *FTO*; and for BMI with *BDNF*. The differing evidence of associations between

populations may be a result of ascertainment or study design, population substructure, genotype call rates, and the degree of LD between the variant being tested and the true causative variant. Subsequently, there may be heterogeneity between populations, due to underlying and unknown genetic or environmental factors, that can account for the differences in associations.

Chapter IV: Genome-wide association study of lipoprotein cholesterol and triglycerides in the Cebu Longitudinal Health and Nutrition Survey (CLHNS) cohort

The goal of this study was to utilize the CLHNS to capture additional genetic diversity and identify important variants that are involved in lipoprotein metabolism that might not have been identified in previous studies. Prior to this research, the majority of studies identifying genes for complex diseases were dominated by samples from European populations. However, in order to have a greater understanding of the interindividual variation of disease, genetic studies need to include populations with a wide range of geographic origins, environmental conditions, and disease burden as it is not yet possible to make generalizations about associations of markers across populations.

Although a targeted effort was made in **Chapter II** to address the genetic structure among the CLHNS women, the GWA study described in this chapter allowed for a more thorough analysis of all the samples to identify population admixture and unknown relatedness. Using a set of independent SNPs ($r^2 < .005$), we constructed principal components to be included as covariates in the association analyses to account

for genetic ancestry differences among study subjects. Additionally, we determined the genomic control values from association scans to examine systematic p-value inflation due to cryptic relatedness between samples. A genomic control value of 1 was determined for HDL-C, low density lipoprotein cholesterol (LDL-C), triglycerides, and total cholesterol, suggesting good control of the overall family-wise type I error rate. Finally, based on identity-by-descent (IBD) and identity-by-state (IBS) estimates, we excluded 81 samples that were likely first-degree relatives with another individual in our sample.

We employed the knowledge gleaned from **Chapter II** to decide how to conduct genotype imputation, an analytic approach that is based on more comprehensively genotyped Hapmap data and LD patterns. Imputation allowed us to evaluate the evidence of association at SNPs that were not directly genotyped. Our data suggested that the use of the CHB+JPT as a reference population was appropriate. Additionally, an evaluation had been done to determine "optimal" mixtures of HapMap panels to maximize imputation accuracy [39]. Huang *et al*. recommends using a combination of two or more HapMap panels for most populations. An independent analysis done by Yun Li (data not shown) retrospectively on our GWA imputation concluded that using the CEU population along with the CHB+JPT population from HapMap gave better imputation results than just CHB+JPT alone. The increase in sample size allowed for a better representation of the CLHNS haplotypes across the genome.

Our study observed evidence of association with many well-known loci implicated in lipid metabolism in previous GWA studies: *MMAB/MVK*, *LIPC* (two association signals), and *CETP* (two association signals) with HDL-C; *APOA1-APOC3-*

APOA4-APOC2, *ANGPTL3/DOCK7*, *GCKR*, *LPL*, *TRIB1*, and *LIPC* with triglycerides; and *APOE*, *HMGCR* and *HNF1A* with LDL-C. As stated above, many of these genes were initially identified in candidate gene studies and have shown evidence of association across many populations. Many of these SNPs show an association with more than one trait, consistent with the correlations between traits. Together, all of the previously reported SNPs for each trait explain 4.2%, 5.5%, 1%, and .5% of the variation for HDL–C, triglycerides, LDL-C, and total cholesterol, respectively, leaving most of the heritability to be explained. One possibility is that other genes are playing a role in the CLHNS population that have not previously been identified in GWA studies, each loci has immeasurable contribution.

We observed two loci with suggestive evidence of association ($P < 10^{-6}$) at genes that have implications in lipid metabolism: Tankyrase (*TNKS*) and Collectin-12 (*COLEC12*). Each of these loci account for 1% of the variation in the CLHNS. *TNKS*-deficient mice exhibit increase in energy expenditure, fatty-acid oxidation, and insulin-stimulated glucose utilization. In these mice, adiposity is substantially decreased even with excessive food intake [40]. *COLEC12* mediates the recognition, internalization and degradation of oxidatively modified LDL-C by vascular endothelial cells [41]. Both of these loci are unique to our study and have not been identified in other GWA studies. Because of their biological likelihood, these genes should be studied in additional samples for replication in both populations of European and Asian descent.

Furthermore, we tested whether genotype associations with lipoprotein levels are mediated by BMI. Obesity is largely responsible for antherogenic dyslipidemia - characterized by low levels of HDL-C, increased triglycerides, and increased LDL-C

[42]. However, adjusting for BMI did not considerably attenuate or increase the evidence of association with most variants, suggesting the changes in lipoprotein levels due to these variants and obesity are independent. However, for the locus *ANGPTL3-DOCK7* with triglycerides we observed a change in P value from .04 with a ß of .038 ±.018 to .006 with a ß of .048 ± .017 after adjusting for BMI. However, this change in P value may be the result of stochastic variation. While *ANGPTL3* or *DOCK7* are unlikely functional candidates, this locus may also be negatively confounded by BMI. However, this locus has not been identified in other genetic studies of obesity-related traits.

Due to the many challenges in understanding the joint effect of genes and environment, we did not analyze these interactions within this study. There are unlimited possible exposures that may be relevant to disease, each of which are difficult to measure in detailed repeatable fashion. These environmental contributions can reduce the power to detect a signal and make it hard to distinguish between real heterogeneity and failure to replicate. Additionally, when performing gene by environment interactions, one must assume that the environmental variable has no genetic component; however, dietary preference and inclination for physical activity may have a significant genetic component thus skewing the results. A targeted approach examining genes that have been implicated to have a biological significance or are in known pathways that may influence diet or physical activity may be a more productive method of performing gene by diet interactions. Future studies in other populations should aim to include environmental data to allow for better characterization of the populations.

This study explored variants associated with lipoprotein levels; despite the success of identifying susceptibility loci there is still much to be learned about each of the

loci implicated. In many cases, it is unclear which gene(s) are being affected by the associated variants (example *MMAB/MVK* locus). In addition, we do not know the role of many associated genes in disease biology due to their unknown function or un-linked pathway to lipoprotein biology (example *GALNT2* **Chapter V**).

Chapter V: Transcriptional activity of SNPS at the *GALNT2* locus associated with human high-density lipoprotein cholesterol levels.

The main goal of this study was to identify the functional SNP(s) that accounts for the HDL-C association signal at the *GALNT2* locus. While the functional SNPs at most of the loci associated with lipoprotein levels have yet to be identified, the *GALNT2* locus was of particular interest because of its relatively localized association signal in intron one of *GALNT2* (~ 15 kb), the small number of variants within that signal, and the limited number of genes located near the signal. SNPs within *GALNT2* have strong evidence of association with HDL-C and triglycerides in several GWA studies [19,43]. However, there is no evidence of association in the CLHNS population. Although this locus was chosen independently of the lipoprotein association study in **Chapter IV**, the methods utilized in this study can be used to investigate the functional SNPs within other associated loci.

In order to identify the functional variant, we first needed to ensure that we had the most inclusive set of SNPs within the association signal. Performing resequencing of exon 2 and dense resequencing of the association region and exons allowed us to identify most of the common variants in this region. However, some of the amplicons for the dense resequencing failed, thus, we were limited to ~80% coverage. Unfortunately, the

failed regions were within some of the most interesting predicted functionally important regions. We further supplemented the resequencing results with data from the HapMap and the 1000 genomes project (http://browser.1000genomes.org), which at the time of analysis had thoroughly resequenced ~45 individuals. We are confident that we have identified most of the common variants at this locus. However, we will be able to better access the entire common genetic variation when more data becomes available from the 1000 genomes project.

We used computational and experimental data of open chromatin and histone modifications [44,45] in the human hepatocellular carcinoma cell line (HepG2) to identify regions that are highly suggestive of regulatory function to prioritize SNPs for testing. These methods have evolved over the last two years to include large genome wide scale analyses of multiple tissue types. We identified two sets of SNPs representing the best *a priori* regulatory candidates at this locus. We tested haplotypes in two distinct genomic regions (Region 1: 780 bp, Region 2: 565 bp) for ability to drive or enhance luciferase transcription in HepG2 cells.

We show that the SNPs in Haplotype 1 containing alleles associated with increased HDL-C levels, the 21 bp deletion of rs6143660, the T allele of rs2144300, and the A allele of rs4846913 of Region 1 consistently cause a significant increase in transcription compared to Haplotype 2 and Haplotype 3 in both a minimal promoter vector and a basic vector in the forward and reverse orientation. These exciting findings suggest that one or all three of these variants are strong functional drivers in this region. Haplotype 2 of Region 1 with the A allele of rs1555290 shows an increase in transcription compared to Haplotype 3 that contains the C allele of rs1555290 in the

forward direction of both vectors, but does not show an increase in the reverse direction suggesting that this SNP may have a modest effect. However, the alleles that differentiate Haplotype 1 are consistent in both orientations and in both vectors and they are strongly associated with HDL-C levels. rs1555290 has an r^2 of .22 with the index association SNP (rs4846914). Additionally rs1555290 was not specifically tested for association but is in r^2 of .88 with a SNP that was tested, rs2296065, which has a P value of 7.1×10^{-6} [46]. In unpublished data from Teslovich *et al*. (Abstract American Society of Human Genetics 2009) of a meta-analysis of 100,000 individuals, rs2296065 is no longer among the most significantly associated SNPs and is sifted out in the "noise". These reasons provide strong support that one of these three variants are likely the functional variants in the *GALNT2* association signal and that rs1555290 may only be moderately contributing if at all. We did not observe allele-specific effects of transcription for Region 2 suggesting the variants contained in this region do not have a role in regulation. The luciferase data as well as the open chromatin data suggest that Region 1 may contain an enhancer or promoter.

Functional experimental approaches are very problematic in that the causal alleles that are being identified in the association signals are likely to have very subtle and potentially immeasurable molecular and cellular effects. Within the human body, these alleles may have long-term effects that cannot be identified in artificial settings. Additionally, the effects shown in this chapter do not by themselves prove a causal role in increased HDL-C. However, we can speculate that one or all four alleles may play a role in regulating transcription.

Notably data from Edmondson *et al.* (Abstract American Society of Human Genetics 2009), contradicts the direction of our results. Edmondson overexpressed mouse and human *GALNT2* in mouse hepatocytes, which led to a significant decrease in HDL-C by approximately 20% at 28 days. Additionally the knockdown endogenous *Galnt2* lead to a 37% increase in HDL-C. These results suggest that *GALNT2* is the casual gene at this locus and that changes in hepatic *GALNT2* expression are associated with inverse changes in HDL-C levels. One hypothesis that could explain this inconsistency is that the variants identified in Region 1 are acting in some way to down regulate expression of GALNT2. This possibility could occur if the variants were acting on a non-coding RNA that in turn regulated *GALNT2* transcription or were competing for regulatory proteins with the promoter of *GALNT2* causing a decrease in transcription. As we move into the future we may be able to elucidate these irregularities in functional data and pinpoint the exact variant(s) that is responsible for the association signal.

SIGNIFICANCE AND FUTURE DIRECTIONS

These studies presented in this dissertation show how the field of human genetics has evolved over the past five years. Predictably it will continue to evolve; the appearance of methods that allow for increased genetic coverage and the accompanying data will allow for more detailed exploration, but also present new challenges for analysis and interpretation.

Public health significance and personalized medicine

Identifying genetic variants that influence susceptibility to risk factors of CVD such as obesity-related traits and lipoproteins and understanding how these genes interact with modifiable environmental factors has important public health significance. Such information can produce additional targeting criteria for disease intervention and prevention.

CVD is an increasing public health burden. By 2030, an estimated 23.6 million people worldwide will die from CVDs, with the largest increase in the number of deaths occurring in the South-East Asia Region. Increasingly, low to middle income countries are becoming disproportionally affected by CVDs with ~80% of all CVD deaths taking place in these countries. CVDs are a multidimensional problem caused by a complex relationship of environmental, social, economic, and behavioral factors, acting on a background of genetic susceptibility. In order to make strides in reducing disease, we must focus on a wide range of research including genetic, molecular, environmental, clinical, and epidemiologic studies.

Ultimately, researchers should seek to translate the results of GWA studies to predict an individual's risks of disease based upon his or her genetic and environmental information. Doctors have long focused on modifiable environmental factors such as diet, exercise, and smoking, but the use of genetic information to guide medical decisions is on the horizon. Promises of personalized medicine have been touted over the last few years; however, we must first fully understand the biological mechanisms of a disease before the reality of this can occur.

Future directions

We are now in a new era of human genetics. We are moving away from traditional linkage and candidate gene studies; we are in the prime of GWA studies and are progressing towards re-sequencing to identify variants associated with disease.

The goal of GWA studies is to identify novel loci across the genome that are associated with disease. In the past few years, these studies have identified hundreds of variants associated with quantitative traits and common complex disease. Many new genes are now implicated which were not previously known to be involved in disease such as *GALNT2* leading researchers to learn more about the pathophysiology of disease. There is still much work to be done in clarifying the role of these genes where the function is not yet known.

Currently, many population studies are focusing on samples with European descent. As observed in **Chapter III** and **Chapter IV**, certain genes identified in the European populations are replicated in the CLHNS samples. However, not all of the results from studies in European populations can be generalized to global populations. Non-European populations, specifically those of African and Asian ancestry, have different genetic structures, disease incidence, and cultural and environmental characteristics and must be independently studied in order to be fully understood. Non-European populations may be useful in identifying variants that may be important in other populations because of ancestry-related differences present in the allele frequencies and the pattern of LD [47,48].

The field of human genetics is progressing towards using larger and larger studies such as a meta-analysis, which can examine multiple data sets and overcome the limitations of power that can compromise single studies. An example of a large meta-

analysis is the Global Lipids Genetics Consortium which includes >100,000 individuals of European ancestry (Teslovich *et al.*, Abstract American Society of Human Genetics 2009). It is valuable to consider including a wider range of ethnic groups to support the discovery of additional susceptibility variants. Additionally, we must concentrate on creating studies that include specific definitions of traits and environment to best eliminate heterogeneity that may cloud the results. Many of the loci identified and that will be identified in GWA studies will only explain a small proportion of the individual differences in disease predisposition underscoring the complexity of these traits

One reason for the relatively limited proportion of phenotypic variation ascribed is because current GWA studies have only explored part of the genome variation represented by common SNPs. This method does not comprehensively evaluate structural variants such as copy number polymorphisms and rare alleles. Soon technological and analytical advances will allow us to more thoroughly analyze these variations. These types of genetic variation can be identified through deep resequencing; the identification of these variants is likely to increase the heritability that can be explained as we potentially identify more than one independent causal variant at a locus.

However, even after we scour all the variation in the genome the entirety of the genetic variants that contribute to the interindividual variation of a disease or a trait will probably never be identified. There are almost certainly hundreds or thousands of variants within each of us that may be modulated by environment and interactions with other variants that contribute minutely to disease. These variants will be harder and harder to identify as human geneticists classify the majority of variants with the strongest

genetic effects and begin probing to variants that may affect fewer individuals and have a minor effect.

In order to understand the true contribution of a locus we must progress towards a biological understanding of its mechanisms. The main output of the current research is association signals, which can rarely pinpoint a causal variant(s) or even the causal gene(s). This is the next roadblock to translating research to application. What we will glean from these experimental advancements will drive new treatments, biomarkers, and prevention.

REFERENCES

1. Encode Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306: 636-640.

2. International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851-861.

3. Rankinen T, Zuberi A, Chagnon YC, Weisnagel SJ, Argyropoulos G, et al. (2006) The human obesity gene map: the 2005 update. Obesity (Silver Spring) 14: 529-644.

4. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, et al. (2007) A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science 316: 889-894.

5. Dina C, Meyre D, Gallina S, Durand E, Korner A, et al. (2007) Variation in FTO contributes to childhood obesity and severe adult obesity. Nat Genet 39: 724-726.

6. Chambers JC, Elliott P, Zabaneh D, Zhang W, Li Y, et al. (2008) Common genetic variation near MC4R is associated with waist circumference and insulin resistance. Nat Genet 40: 716-718.

7. Loos RJ, Lindgren CM, Li S, Wheeler E, Zhao JH, et al. (2008) Common variants near MC4R are associated with fat mass, weight and risk of obesity. Nat Genet 40: 768-775.

8. Cho YS, Go MJ, Kim YJ, Heo JY, Oh JH, et al. (2009) A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. Nat Genet 41: 527-534.

9. Meyre D, Delplanque J, Chevre JC, Lecoeur C, Lobbens S, et al. (2009) Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. Nat Genet 41: 157-159.

10. Soranzo N, Rivadeneira F, Chinappen-Horsley U, Malkina I, Richards JB, et al. (2009) Meta-analysis of genome-wide scans for human adult stature identifies novel loci and associations with measures of skeletal frame size. PLoS Genet 5: e1000445.

11. Thorleifsson G, Walters GB, Gudbjartsson DF, Steinthorsdottir V, Sulem P, et al. (2009) Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. Nat Genet 41: 18-24.

12. Willer CJ, Speliotes EK, Loos RJ, Li S, Lindgren CM, et al. (2009) Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. Nat Genet 41: 25-34.

13. Heard-Costa NL, Zillikens MC, Monda KL, Johansson A, Harris TB, et al. (2009) NRXN3 is a novel locus for waist circumference: a genome-wide association study from the CHARGE Consortium. PLoS Genet 5: e1000539.

14. Lindgren CM, Heid IM, Randall JC, Lamina C, Steinthorsdottir V, et al. (2009) Genome-wide association scan meta-analysis identifies three loci influencing adiposity and fat distribution. PLoS Genet 5: e1000508.

15. Gerken T, Girard CA, Tung YC, Webby CJ, Saudek V, et al. (2007) The obesity-associated FTO gene encodes a 2-oxoglutarate-dependent nucleic acid demethylase. Science 318: 1469-1472.

16. Jia G, Yang CG, Yang S, Jian X, Yi C, et al. (2008) Oxidative demethylation of 3-methylthymine and 3-methyluracil in single-stranded DNA and RNA by mouse and human FTO. FEBS Lett 582: 3313-3319.

17. Fredriksson R, Hagglund M, Olszewski PK, Stephansson O, Jacobsson JA, et al. (2008) The obesity gene, FTO, is of ancient origin, up-regulated during food deprivation and expressed in neurons of feeding-related nuclei of the brain. Endocrinology 149: 2062-2071.

18. Fischer J, Koch L, Emmerling C, Vierkotten J, Peters T, et al. (2009) Inactivation of the Fto gene protects from obesity. Nature 458: 894-898.

19. Kathiresan S, Melander O, Guiducci C, Surti A, Burtt NP, et al. (2008) Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. Nat Genet 40: 189-197.

20. Liu YJ, Liu XG, Wang L, Dina C, Yan H, et al. (2008) Genome-wide association scans identified CTNNBL1 as a novel gene for obesity. Hum Mol Genet 17: 1803-1813.

21. Lyon HN, Emilsson V, Hinney A, Heid IM, Lasky-Su J, et al. (2007) The association of a SNP upstream of INSIG2 with body mass index is reproduced in several but not all cohorts. PLoS Genet 3: e61.

22. Hotta K, Nakamura M, Nakata Y, Matsuo T, Kamohara S, et al. (2008) INSIG2 gene rs7566605 polymorphism is associated with severe obesity in Japanese. J Hum Genet 53: 857-862.

23. Zhang J, Lin R, Wang F, Lu M, Lin RY, et al. (2008) A common polymorphism is associated with body mass index in Uyghur population. Diabetes Res Clin Pract 81: e11-13.

24. Orkunoglu-Suer FE, Gordish-Dressman H, Clarkson PM, Thompson PD, Angelopoulos TJ, et al. (2008) INSIG2 gene polymorphism is associated with increased subcutaneous fat in women and poor response to resistance training in men. BMC Med Genet 9: 117.

25. Hall DH, Rahman T, Avery PJ, Keavney B (2006) INSIG-2 promoter polymorphism and obesity related phenotypes: association study in 1428 members of 248 families. BMC Med Genet 7: 83.

26. Dina C, Meyre D, Samson C, Tichet J, Marre M, et al. (2007) Comment on "A common genetic variant is associated with adult and childhood obesity". Science (New York, NY) 315: 187; author reply 187.

27. Loos RJ, Barroso I, O'Rahilly S, Wareham NJ (2007) Comment on "A common genetic variant is associated with adult and childhood obesity". Science (New York, NY) 315: 187; author reply 187.

28. Rosskopf D, Bornhorst A, Rimmbach C, Schwahn C, Kayser A, et al. (2007) Comment on "A common genetic variant is associated with adult and childhood obesity". Science (New York, NY) 315: 187; author reply 187.

29. Kumar J, Sunkishala RR, Karthikeyan G, Sengupta S (2007) The common genetic variant upstream of INSIG2 gene is not associated with obesity in Indian population. Clin Genet 71: 415-418.

30. Smith AJ, Cooper JA, Li LK, Humphries SE (2007) INSIG2 gene polymorphism is not associated with obesity in Caucasian, Afro-Caribbean and Indian subjects. Int J Obes (Lond) 31: 1753-1755.

31. Kuzuya M, Ando F, Iguchi A, Shimokata H (2007) No association between rs7566605 variant and being overweight in Japanese. Obesity (Silver Spring) 15: 2531-2534.

32. Tabara Y, Kawamoto R, Osawa H, Nakura J, Makino H, et al. (2008) No association between INSIG2 Gene rs7566605 polymorphism and being overweight in Japanese population. Obesity (Silver Spring) 16: 211-215.

33. Andreasen CH, Mogensen MS, Borch-Johnsen K, Sandbaek A, Lauritzen T, et al. (2008) Non-replication of genome-wide based associations between common variants in INSIG2 and PFKP and obesity in studies of 18,014 Danes. PLoS One 3: e2872.

34. Boes E, Kollerits B, Heid IM, Hunt SC, Pichler M, et al. (2008) INSIG2 polymorphism is neither associated with BMI nor with phenotypes of lipoprotein metabolism. Obesity (Silver Spring) 16: 827-833.

35. Oki K, Yamane K, Kamei N, Asao T, Awaya T, et al. (2009) The single nucleotide polymorphism upstream of insulin-induced gene 2 (INSIG2) is associated with the prevalence of hypercholesterolaemia, but not with obesity, in Japanese American women. Br J Nutr 101: 322-327.

36. Wiedmann S, Neureuther K, Stark K, Reinhard W, Kallmunzer B, et al. (2009) Lack of association between a common polymorphism near the INSIG2 gene and BMI,

myocardial infarction, and cardiovascular risk factors. Obesity (Silver Spring) 17: 1390-1395.

37. Vimaleswaran KS, Franks PW, Brage S, Sardinha LB, Andersen LB, et al. (2009) Absence of association between the INSIG2 gene polymorphism (rs7566605) and obesity in the European Youth Heart Study (EYHS). Obesity (Silver Spring) 17: 1453-1457.

38. Croteau-Chonka DC, Marvelle AF, Lange EM, Lee N, Adair LS, Lange LA, Mohlke KL. (unpublished) Genome-wide association study of four anthropometric traits in a cohort of Filipino women from the Cebu Longitudinal Health and Nutrition Survey.

39. Huang L, Li Y, Singleton AB, Hardy JA, Abecasis G, et al. (2009) Genotype-imputation accuracy across worldwide human populations. Am J Hum Genet 84: 235-250.

40. Yeh TY, Beiswenger KK, Li P, Bolin KE, Lee RM, et al. (2009) Hypermetabolism, hyperphagia, and reduced adiposity in tankyrase-deficient mice. Diabetes.

41. Ohtani K, Suzuki Y, Eda S, Kawai T, Kase T, et al. (2001) The membrane-type collectin CL-P1 is a scavenger receptor on vascular endothelial cells. J Biol Chem 276: 44222-44228.

42. Miller WM, Nori-Janosz KE, Lillystone M, Yanez J, McCullough PA (2005) Obesity and lipids. Curr Cardiol Rep 7: 465-470.

43. Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, et al. (2008) Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet 40: 161-169.

44. Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD (2007) FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res 17: 877-885.

45. Crawford GE, Holt IE, Mullikin JC, Tai D, Blakesley R, et al. (2004) Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. Proc Natl Acad Sci U S A 101: 992-997.

46. Kathiresan S, Willer CJ, Peloso GM, Demissie S, Musunuru K, et al. (2009) Common variants at 30 loci contribute to polygenic dyslipidemia. Nat Genet 41: 56-65.

47. Unoki H, Takahashi A, Kawaguchi T, Hara K, Horikoshi M, et al. (2008) SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. Nat Genet 40: 1098-1102.

48. Yasuda K, Miyake K, Horikawa Y, Hara K, Osawa H, et al. (2008) Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus. Nat Genet.