UNIVERSITY OF CALIFORNIA SAN DIEGO

Leveraging Data from Large Biorepositories to Study the Genetic Basis of Metabolic Syndrome

A Thesis submitted in partial satisfaction of the requirements for the degree Master of Science

in

Biology

by

Steven Cao

Committee in charge:

Professor Rany Salem, Chair
Professor Steve Briggs, Co-Chair
Professor Gen-Sheng Feng

2019

The Thesis of Steven Cao is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

Co-Chair

_____

Chair

University of California San Diego

2019

Table of Contents

List of Figures

List of Tables

ABSTRACT OF THE THESIS

Leveraging Data from Large Biorepositories to
Study the Genetic Basis of Metabolic Syndrome

by

Steven Cao

Master of Science in Biology

University of California San Diego, 2019

Professor Rany Salem, Chair
Professor Steve Briggs, Co-Chair

Metabolic syndrome (MetS) is an emerging global epidemic of public health importance. MetS is a syndrome characterized by having three of the following conditions: abdominal obesity, hypertension, high blood sugars, abnormal levels of triglycerides and low levels of high-

density lipoprotein (HDL). It is well-established that environmental and lifestyle factors play a major role in the development of MetS, but a full understanding of the genetic variants that are involved in the disease pathogenesis is incomplete. To identify these genetic variants, we have conducted large-scale Genome Wide Association Studies (GWAS) on samples retrieved from the Database for Genotype and Phenotypes (dbGaP), a biomedical repository for individual level genotype and phenotype data. Extensive work has been performed on individual components of MetS, leveraging thousands of samples. In contrast, prior GWAS of MetS have largely utilized modest sample sizes. In this analysis, three different studies with a total of 10,000 MetS cases were used to discover novel genetic variants associated with MetS. For each study, an extensive quality control was performed to filter and harmonize these datasets. Variable harmonization for uniformity and genetic variant imputation for maximizing the number of Single Nucleotide Polymorphisms (SNPs) were also completed prior to the GWAS. After quality control and imputation, GWAS was performed separately on each dataset, and results were combined via meta-analysis. In this analysis, we identified four genome-wide significant variants (rs287, rs964184, rs11076176, rs247616) in the European ancestry subset and two genome-wide significant variants in the African ancestry subset (rs117729532, rs115553887).

Chapter I

Introduction

Metabolic Syndrome Background and Definition

Metabolic syndrome (MetS), also known as syndrome X or insulin resistance syndrome, is characterized by having three out of five of the following conditions: abdominal obesity, hypertension, high blood sugars, and abnormal levels of triglycerides and low levels of high-density lipoprotein (HDL). Individuals with MetS have a higher risk of developing cardiovascular disease (CVD) and Type II Diabetes (T2D). MetS imposes significant costs on the health care system. Each individual component of MetS may require substantial medical treatment and associated costs. As a result, each risk factor of MetS is said to increase an individual's healthcare bill by as much as 1.6-fold, or 2000 dollars per year, and each additional risk factor will increase this rate by another 24% [1]. To address this rapidly pressing issue, the International Diabetes Federation (IDF) established a uniform definition of MetS for clinical practice (Table 1 and Table 2). The IDF definition incorporates an ethnic-specific cutoff for the waist circumference values. The new IDF definition originally supports the fact that central obesity, which is assessed by the measurement of waist circumference, must be present in each case of MetS [2]. In 2009, this criteria was revised, and a joint statement released by International Diabetes Federation Task Force on Epidemiology and Prevention; National Heart, Lung, and Blood Institute; American Heart Association; World Heart Federation; International Atherosclerosis Society; and International Association for the Study of Obesity agreed that there should be no obligatory component, but that IDF's original cutoff points were to be used as a guideline for MetS [3]. In summary, having three out of the five conditions of high blood sugar,

Table 1: International Diabetes Federation (IDF) definition for MetS

According to the new IDF definition, for a person to be defined as having the metabolic syndrome they must have: **Any of the following 3 out of 5 factors. Central Obesity is the fifth factor and is defined in Table 2 with ethnic breakdowns.**	
Raised triglycerides	≥ 150 mg/dL (1.7 mmol/L) **or specific treatment for this lipid abnormality**
Reduced HDL cholesterol	< 40 mg/dL (1.03 mmol/L) in males < 50 mg/dL (1.29 mmol/L) in females **or specific treatment for this lipid abnormality**
Raised blood pressure	systolic BP ≥ 130 or diastolic BP ≥ 85 mm Hg **or treatment of previously diagnosed hypertension**
Raised fasting plasma glucose	(FPG) ≥ 100 mg/dL (5.6 mmol/L), **or previously diagnosed type 2 diabetes** If above 5.6 mmol/L or 100 mg/dL, OGTT is strongly recommended but is not necessary to define presence of the syndrome.

Table 2: Ethnic-Specific Values for Waist Circumference

Country/Ethnic Group		**Waist circumference**
Europids In the USA, the ATP III values (102 cm male; 88 cm female) are likely to continue to be used for clinical purposes.	Male	≥ 94 cm
	Female	≥ 80 cm
South Asians Based on a Chinese, Malay and Asian-Indian population	Male	≥ 90 cm
	Female	≥ 80 cm
Chinese	Male	≥ 90 cm
	Female	≥ 80 cm
Japanese	Male	≥ 90 cm
	Female	≥ 80 cm
Ethnic South and Central Americans	Use South Asian recommendations until more specific data are available	
Sub-Saharan Africans	Use European data until more specific data are available	
Eastern Mediterranean and Middle East (Arab) populations	Use European data until more specific data are available	

blood pressure, waist circumference, triglyceride and low HDL is the current and accepted diagnosis for MetS.

The establishment of MetS as a syndrome has been controversial ever since the original coining of the term "Syndrome X" in Reavan's 1988 Banting Lecture [4]. The validity of the syndrome has been questioned, as has its utility and whether the five components that underlie the MetS diagnosis indeed have biological or genetic interconnections. In addition to the common clustering of these conditions observed in clinical trials, principle components analysis that combine the risk factors of MetS "support the idea that the clustering or correlation among them is heritable and has a genetic basis" [5]. Additional evidence that supports the existence of MetS is the elevated risks of CVD and T2D in patients that fulfill syndrome criteria as opposed to those whom are diagnosed with only a single component. By tying together the interrelated concepts of insulin resistance, dyslipidemia and hypertension, MetS can help researchers better understand the common pathophysiological processes involved in these risk factors [6]. In conclusion, despite controversies in this topic, considering MetS as a specific disease entity is valid and important for research. Doing so promotes research into the genetic susceptibility of the syndrome and potential development of novel treatment approaches for both the syndrome and the individual components.

Metabolic Syndrome Pathogenesis

Metabolic Syndrome has five different components, all of which have a unique contribution to MetS disease pathogenesis, and therefore CVD and T2D risk. The first component is high blood pressure, which is a medical condition in which blood pressure in the arteries is persistently elevated. This condition causes continuous force and damage on the walls

of the blood vessels. Once the walls are damaged, they become susceptible to cholesterol and calcium plaque formation. Undiagnosed plaque build-up can grow with time, eventually leading to blockages in the coronary arteries. These partial blockages are a standard characteristic of coronary artery disease and can cause chest pains known as Angina. In addition to this, high blood pressure also weakens and hardens blood vessels. When these weakened blood vessels rupture, a heart attack occurs. The rupturing of coronary arteries recruits blood clotting factors, known as platelets, that assist in clotting the ruptured vessels, thereby completely blocking the blood flow into the heart. Persistently high blood pressure leads to a cascade of atherogenic symptoms, which advance the development of heart failure and other complications.

The second criterion is high blood sugar in the bloodstream. High blood sugar is primarily the result of insulin resistance, where the cells do not correctly respond to insulin signaling to absorb circulating blood sugar. Elevated cholesterol and fat content in the body, along with habitually high intake of sugars, are thought to increase risk of insulin resistance. A harmful aftermath of having high blood sugars is reduced levels of the powerful vasodilator nitric oxide in blood vessels. The presence of nitric oxide relaxes blood vessels, so a rapid reduction increases blood pressure in the bloodstream, causing a variety of health problems such as coronary heart disease and chronic kidney disease. Furthermore, an individual with a fasting plasma glucose level of 126 milligrams per deciliter or higher is diagnosed with T2D, which is associated with an array of other metabolic disorders.

The third component is low levels of high-density lipoproteins (HDL). HDL are proteins that move fat around in the body. HDL cholesterols are known as the maintenance crew of the body because they scavenge the bloodstream to remove and recycle low-density lipoprotein (LDL), a form of "bad" cholesterol which has the primary function of delivering cholesterol to

the cell. A high-serum LDL can build up in blood vessels, resulting in plaque that blocks blood flow to essential organs of the body. Low levels of HDL are often observed with high levels of LDL, which are known to increase cardiovascular disease risks through plaque formation and build-up.

The last two criteria of MetS are high triglyceride levels and central obesity. Triglycerides are a form of fat either consumed or produced by the liver. The primary function of triglyceride is for energy storage and production. Triglycerides are stored within the adipose tissues of the body, and an excess amount of triglyceride storage causes central obesity because adipose tissues are stored in the abdominal area. Besides its contribution to central obesity, triglycerides, though not reported to directly increase atherosclerosis and cardiovascular disease rates, are important biomarkers for several types of atherogenic lipoproteins, and therefore are associated with CVD [7].

It is no surprise that obesity makes up an important component of MetS, as excessive body weight has long been known to be associated with a range of co-morbidities including cardiovascular diseases, diabetes and even some types of cancer. The biological mechanism that propels central obesity to cardiovascular risks is not fully understood. Adipose tissues have the main role of long-term energy storage in forms of fat. In one potential hypothesis, excess obesity results in phenotypic change in the tissue. This involves a low-grade chronic inflammation created by the expansion of the adipose tissues and other metabolic factors, including "excess fatty acids, hypoxia, and activation of the inflammasome" [8]. This inflammation process recruits an abundance of immune cells, including M1 macrophages and T lymphocytes, as a response to the cellular stress signal. These immune cells then release pro-inflammatory cytokines, which blunt the insulin signaling cascade and cause insulin resistance. From this point on, the onset of

T2D is eminent, and the consequences of having high blood sugar is a higher risk of cardiovascular disease or stroke.

Metabolic Syndrome Risk Factors

Alcohol, smoking, physical activity and female gender are all significant risk factors that influence the occurrence of MetS cases. Alcohol usage has mixed associations with MetS, and has been reported to be both as a risk factor and a protective factor depending on the amount and type of alcohol consumed. Multiple observational studies have reported that a light/moderate alcohol usage have favorable effects, including reduced CVD risks and all-cause mortality. This protective effect is much more significant in wine drinkers than beer and liquor drinkers and nondrinkers. Research suggests that alcohol's effect may differ by age and gender, with the most significant associations in women and individuals age of 70 or lower. In the PRevención con DIeta MEDiterránea (PREDIMED) trial, it was reported that wine drinker's risk of MetS was reduced by 44%, and these individuals also "showed a lower risk of having abnormal [waist circumference], low HDL-C, high [blood pressure], and high fasting plasma glucose levels." [9]. The protective factor of red wine might be due to its concentration of polyphenol, an organic chemical known to increase levels of HDL. Red wine has also been associated with glycemic control and an overall lower insulin resistance when consumed in moderate amounts. On the other hand, high alcohol intake has unfavorable effects in multiple components of MetS. A multiple logistic regression analysis by Kim et al. concluded that "alcohol consumption >5 g/day may contribute to abnormalities of MetS, including high glucose and blood pressure, hypertriglyceridemia, and low HLD cholesterol" [10].

In a meta-analysis of over 56,000 participants, Sun et al found that smokers have a 26% increased rate of Metabolic Syndrome [11]. Life Lines Cohort Study found the same association within both sexes indiscriminately [12]. Smoking may affect MetS risk through two important mechanisms. First, increased cortisol hormone production in smokers trigger a fight-or-flight response from the body, causing a flood of glucose to be released to supply an immediate energy source. The build-up of glucose and lack of insulin to convert these sugars into usable energy increases blood sugar content in the bloodstream. Cortisol release also causes the body to send signals to the brain to eat in order to supply the energy demands of cells. False hunger signals result in weight gain and other metabolic morbidities related to high triglycerides and an abnormal waist circumference.

In terms of gender differences, many studies have shown that risk of developing MetS is significantly greater in females than in males. In a comparative study involving 500 patients, the prevalence of MetS in women and men corresponded to 29% and 23% [13]. Several factors may explain the increased risk of MetS in women vs. men, including pregnancy, oral contraceptive use and menopause. Jain et al.'s work on gender differences of MetS determined that increased waist circumference and hyperglycemia were larger contributors to the MetS criteria in women compared to men [13]. Pregnancy, oral contraceptives and menopause all promote increased central adiposity through fat distribution changes or weight gain.

A healthy routine of physical activity has proven to be one of the most effective treatments for MetS and T2D. Physical activity leads to enhanced energy consumption in the body, reducing the risk of abnormal ectopic storage of excessive energy in non-adipose tissues, such as the liver. Aerobic and resistance type exercises have also been shown to decrease the risk of T2D. In a large prospective study performed by Hemrich et al., incidence of diabetes

decreased by 6% as energy expenditure increased by 500 kcal. per week [14]. Additional evidence for improved glucose homeostasis from physical activity was shown in a randomized control trial that involved a lifestyle intervention of at least 150 minutes of physical activity per week. In a study performed by Diabetes Prevention Program Research Group involving 3,234 non-diabetic patients with elevated fasting glucose, physical activity was found to be more effective than Metformin alone in reducing the onset of T2D [15]. Finally, physical activity has also been shown to reduce the risk of cardiovascular events, potentially through the improvement of lipid profile and anti-inflammatory factors. To study the effects of physical activity on dyslipidemia, Kraus et al. randomly assigned different levels of intensity and duration exercises to 111 subjects [16]. The study reported a decrease in concentration of LDL and an increase in concentration of HDL in the high-intensity high duration subjects. Work by Majka et al. suggests an inverse relationship between inflammatory agents such as high sensitivity C-reactive protein and physical activity [17]. This protective effect is the result of physical activity increasing anti-inflammatory agents such as cytokine inhibitors and TNK receptors. Recall that inflammatory markers blunt the insulin signaling cascade, thereby speeding up the development of T2D. CRP is also associated with inflammation within blood vessel walls, promoting plaque buildup and CVD.

Genetics of Metabolic Syndrome

The underlying genetic complexities of MetS stem from its numerous components, their interactions and interplay with numerous lifestyle and environmental risk factors. Though widely studied, the genetic architecture and origin of MetS is still not completely elucidated. Since MetS is a syndrome comprised of multiple underlying morbidities, two main hypotheses have been proposed to explain the development of MetS. McGarry's Banting Lecture in 2001 supported the

MetS obesity theory, in which obesity-associated metabolic dysfunctions "induce cellular stress that initiates and propagates an inflammatory cycle" [5]. In this hypothesis, accumulation of abdominal and ectopic fat leads to elevated fatty acids that blunt the insulin signaling cascade, therefore promoting the development of insulin resistance. The alternate theory, known as the insulin hypothesis, proposes that obesity is the result of insulin and leptin resistance. Though controversial, many studies have shown that components of insulin resistance, obesity, hypertension and dyslipidemia all have joint genetic contributions to MetS. In order to decode the genetic variability of MetS, we first consider GWAS that highlight genetic variants that are associated with disease endpoints of the individual component of MetS. Then, we can briefly review studies that look for pleiotropic genetic variants, or genes that influence multiple phenotypes.

Type II diabetes has been observed to have both a strong environmental and heritable component. Multiple twin studies have determined the heritability of T2D to be in the range of 20%-80% [18]. *IGF2BP2*, otherwise known as insulin-like growth factor 2 mRNA binding protein 2, is one of the Mendelian genes that contribute to T2D development. This gene codes for a transcription factor that binds to an *IGF2*, or insulin-like growth factor, promoter region, thereby regulating the translational activity of this gene. *IGF2* is a gene known for insulin receptor binding, metabolism and growth. Epigenetic changes to this gene can result in growth restrictions and susceptibility to diabetes. Another well-documented genetic variant is located on the *CDKAL1* gene. In a Finnish study, Stancáková tested the insulin and glucose levels of two groups of participants: one of which had the "C" allele at the rs7754840 SNP location and another who did not have this allele. The result of the study showed that participants having the

"C" allele had lower first-phase insulin release in an IVGTT, or IV Glucose-Tolerance test, and also had a higher glucose area under the curve in an OGTT, or Oral Glucose Tolerance Test [19].

Hypertension is another MetS component with high heritability, with twin studies reports ranging from 30-70% [20]. Most of the SNPs associated with blood pressure can be divided into two categories: the SNPs that predispose an individual to hypertension and the SNPs that influence drugs that are usually used to treat hypertension, such as ACE inhibitors. One of these SNPs (rs4961) on the *ADD1* gene is known to increase the risk of high blood pressure by 1.8 times [21]. Another well-studied SNP (rs5186) is located in the 3' untranslated region of the angiotensin II receptor type 1 gene *AGTR1*. Being homozygote (C:C) for this SNP results in a 7.3-fold increase in hypertension risk [22]. The *AGTR1* gene produces a hormone that is a primary regulator of aldosterone secretion. Aldosterone is a hormone that is responsible for regulating sodium homeostasis, thereby controlling blood pressure and blood volume. Finally, two SNPs (rs1801253 and rs1801252) located in the *ADRB1* gene are associated with affecting beta blocker therapy. In a study done by Johnson et al., the effect of beta blocker monotherapy was compared when given to participants with varying SNP genotypes [23]. Interestingly, individuals who were homozygous carriers of rs1801252 (A;A) and rs1801253 (C;C) logged an average of a fifteen point drop in blood pressure compared to only a less than one point drop in heterozygous carriers of rs1801252 (A;G) and rs1801253 (C;G).

Family studies have reported triglycerides to have a heritability estimate of around 40%, and over 30 single nucleotide polymorphisms have been reported to influence plasma triglyceride levels [24]. Recent large-scale meta-analysis attempts to address whether "common and rare variants in genes whose products are determinants of plasma triglycerides are also associated with clinical cardiovascular endpoints." [25]. A recent paper by Dron et al. pointed to

triglycerides, along with decreased levels of HDL, as being potentially causal for cardiovascular and stroke outcomes [25]. Two of these significant variants, rs7679 on the *PCIF1* gene and rs2624265 located on chromosome 15, result in increased levels of serum triglyceride [26].

The heritability estimates for plasma HDL cholesterol range from 40-60% and well over 30 SNPs have been reported to be associated with influencing HDL levels [27]. A SNP (rs4149268) on the *ABCA1* gene is associated with lower HDL levels [28]. The *ABCA1* gene translates into a protein that acts as a cholesterol efflux pump, which removes excess lipids from the cell. Another SNP (rs2271293) that falls on the *LCAT* gene has a minor allele that is associated with decreased HDL, LDL, and triglycerides [26]. *LCAT* encodes for the extracellular cholesterol esterifying enzyme, which is responsible for the esterification and transportation of cholesterol. A mutation in this gene therefore reduces both HDL and LDL cholesterol levels. Many SNPs that are associated with HDL also relate to both LDL and triglycerides, which reinforces the idea of a close relationship between pathways that are involved in dyslipidemia.

Finally, results from recent genome-wide associations study confirm the numerous genetic interconnections between MetS and obesity, potentially through genes with pleiotropic effects. From family and twin studies, the range of heritability for adiposity phenotypes is 30-70% [5]. Though the biological mechanism for many of these variants remain unclear, a number of these SNPs are located on genes that are associated in the brain and hypothalamus, "suggesting a role for neuronal control in body weight regulation" [5]. An example of this would be the association of the *NRXN3* gene with obesity. The *NRXN3* gene is expressed in the brain and is involved in addiction, reward behavior and synaptic plasticity. Polymorphisms in this gene have been linked to genetic predisposition to drug addiction and obesity. Two polymorphisms (rs1121980 and rs9939609) in the *FTO* gene have been reported to be associated with early onset

obesity. The *FTO* gene, also known as the fat mass and obesity-associated gene, plays a major role in controlling feeding behavior and energy expenditure. In a study by John R. Speakman, researchers discovered that individuals carrying at-risk (A:T) and (A:A) alleles at rs9939609 consumed between 125 and 280 kcal more each day than those carrying the protective (T:T) genotype [29].

MetS has an estimated heritability of around 30% [5]. Human genetic studies of MetS often attempt to find variants and gene loci that relate to one or more phenotypes. These genes that have penetrance on multiple phenotypes are known as having pleiotropic effects. An example of this is the *NR3C1* gene, which has been associated with hypertension, obesity and insulin resistance [30]. The *NR3C1* gene codes for the Glucocorticoid Receptor (GR), which is the receptor that cortisol binds to. GR is expressed in almost every cell of the body and is involved in the regulation of development, metabolism and immune response. Recall that the release of cortisol in smokers causes a rise in blood sugar and a fight-or-flight response, which triggers false hunger signals. Impairment of this cortisol receptor can contribute to the development of MetS in a similar manner, where the binding of cortisol and other glucocorticoids by an affected GR triggers a downstream reaction of risk factors. The *ADIPOQ* gene is another example of a pleiotropic gene, affecting multiple phenotypes such as diabetes, hypertension and dyslipidemia [31]. *ADIPOQ* is well known for producing a protein hormone that is localized in adipose tissues. These hormones are involved in regulating glucose levels and promoting fatty acid breakdown in the cell. Therefore, risk alleles presented on this gene that can potentially alter or impair its functions might have catastrophic consequences. For example, a deregulation of glucose levels in the bloodstream can lead to the onset of diabetes. Also, an

abnormal fatty acid build-up in the adipose tissues increases waist circumference and promotes dyslipidemia.

There have been extensive GWAS efforts to study MetS, but some key limitations exist in these studies. First, many studies suffer from modest sample size (N<2500), which poses the challenge of lacking power to detect significant variants. Another consequence of a small sample size is its inability to detect associations of rare variants. Lin et al., Zhu et al. and Zabeneh et al. have all performed a GWAS on MetS with less than 2,500 MetS case samples (1811, 545 and 2,371, respectively) [47] [48] [49]. Another potential issue with the current MetS GWAS protocol is the definition of control samples. The standard case-control definitions, such as those in the studies listed above, define control samples with having zero to two metabolic abnormality. There have been no studies that differentiate between the benefit of establishing a strict super control definition (zero MetS criteria met) as opposed to the traditional normal control definition. The inclusion of controls with MetS disease components (1 or 2 of 5) may attenuate power of the genetic analyses. In our study, sample size is maximized by leveraging three dbGaP datasets for analysis. After this preliminary analysis, 350,000 European ancestry subjects from UK Biobank are planned to be added to a future GWAS. Finally, to better understand the effectiveness of different control definitions, both super and normal control criteria are used in the analysis.

Research Aims and Goals

This project builds on the availability of large-scale individual level genotype and phenotype data from dbGaP, a biorepository that provides a wealth of research datasets with great potential for data mining and genetic studies. There are two main research goals in the project. The first main goal is to harmonize and quality control a large collection of dbGaP

datasets that have phenotype information of interest in order to prepare for current and future research studies. The second research aim is to use the quality-controlled datasets to study the disease components and pathogenesis of MetS. To do this, we ran multiple GWAS on relevant studies that measure MetS phenotypes to identify genetic variants associated with this disease. The MetS variants are likely to be pleiotropic, showing association with more than one MetS disease phenotypes, and can potentially guide development of new drug targets for disease treatment. In addition, pleiotropic SNPs can illuminate MetS disease pathophysiology by tying together multiple biological processes, which will greatly enhance our understanding of the disease. Many current studies of MetS have GWAS with moderate sample sizes. Moderate sample-sized GWAS often filter out the effects of rare variants because there are not enough case and control subjects that are representative of a particular rare SNP. In my research, I attempt to incorporate a large amount of study subjects in order to capture the effects of these low-frequency variants, with minor allele frequencies down to 1%.

Chapter II

Materials and Methods

Database for Genotype and Phenotype (dbGaP)

Database for Genotype and Phenotype (dbGaP) is a widely accessible biomedical biorepository containing troves of valuable individual-level genotype and phenotype information. These datasets have become a major resource for conducting genetic epidemiology research studies. dbGaP is a biomedical repository managed by the National Center for Biotechnology Information (NCBI), within the National Institute of Health, developed to "archive and distribute the data and results from studies that have investigated the interaction of genotype and phenotype in Humans" [32]. dbGaP facilitates research access by centralizing data from hundreds of studies and standardizing application process to restrict access to researchers with proper credentials. Although biorepositories like dbGaP benefit the scientific community greatly, the available data often cannot be used immediately. Since data is deposited from a variety of different research groups, data quality and its accompanying documentation is heterogeneous. For example, there is no consistent genotyping array technology that is used for the deposited genetic data. Instead, GWAS genetic data is genotyped on two major genotyping platforms (Affymetrix and Illumina) and a multitude of versions of genotyping arrays released over the past decade. Some groups provide raw genotype data, while others provide comprehensively quality-controlled data. Since the organization of key variables and tables are not uniform amongst studies, it can be difficult to extract the information needed for a research study. Therefore, a meticulous quality control of genetic data and data harmonization process must be considered prior to doing any type of analysis on the dbGaP datasets.

Quality Control Pipeline

To accomplish this goal, we undertook a complete and comprehensive review and revision of Dr. Salem's semi-automated quality control pipeline for whole-genome genotyping data. This process entails a thorough review of the code to check for errors, improvements to state-of-the-field computational tools and development of new scripts to increase performance and reduce computational load. The first step in this process was to update the paths of scripts used in the pipeline. Outdated programs were then replaced with faster and more efficient versions. In the pipeline, shell scripts were optimized and any redundant code was removed to increase computational efficiency. A summary of each quality control step after each run was incorporated to speed up manual data recording steps. With these changes, the updated code was five to ten times faster than the original code. Besides updating the quality control pipeline, tools were also developed to work with dbGaP data. To simplify these files, a data dictionary tool was developed to parse the phenotype files given in XML format into a concise table to aid with the variable harmonization process. A data harmonization tool was also developed to rename dbGaP variables and output a concatenated subset of the main data table.

The multi-step QC process results in the application of uniform quality control filters applied to raw genotype data. The quality control pipeline was run in a Linux environment using numerous bash shell scripts. In addition, custom Python, R and Perl were all utilized in the quality control process. Data manipulation of the genotype files was largely performed using PLINK. Other programs used in the pipeline include EIGENSTRAT for Principle Component Analysis and KING for unrelated subjects subsetting [33] [34]. The updated code is briefly described below.

Step 1 is the pre-processing step of the procedure. In this step, phenotype and genotype data retrieved from dbGaP are placed in secure servers. Phenotype files contain phenotypic variables of each subjects, such as weight or smoking information. Genotype files contain whole genome genotyping results for each individual. This file contains a range of 200,000 to 2.5 million variants along with their location and allele information. After converting the files to PLINK format, genotype files go through a duplicates check to remove SNPs or subjects that have duplicate identifiers. Next, the genome build of the genetic variants is determined and variants that do not match this build are removed. A reference genome build is a widely used digital database assembled by scientists as a representative of a specific genome. As sequencing technology advances, the number and lengths of gaps in the reference genome decreases, fostering the advancement to a new genome build. Studies from dbGaP have used reference genome build HG16, HG17, HG18, and HG19, depending on when the subjects are genotyped. Besides labeling the genome build, another goal of this pipeline is to align all genomes to be HG19 build in order to simplify comparison and association tests. Next, case-control status and sex identifiers are checked to ensure correct labeling. HapMap subjects, which are reference samples often included on genotyping runs for comparison purposes, are filtered and removed. These individuals are also filtered and removed. Finally, variants with missing alleles are filtered out of genotyping files. The resulting file from Step 1 is utilized as input for Step 2.

Step 2 of the pipeline is run automatically, with manual review checks at the end of the run. In Step 2, the following quality control filters are applied. First, a sex check that compares reported sex with genotyped sex through using X and Y chromosome variants is applied. Any subjects that differ from their genotyped sex are filtered out in order to avoid potential bias. Next, subjects are divided into their respective self-reported racial/ethnic ancestry. Subject call rates

and variant call rates are also verified during this step. The variant call rates filter removes SNPs that have high rates of missingness (95% SNP call rate required). Similarly, the subject call rates filter removes subjects that do not have sufficient SNP information across all variants (98% subject call rate required). Chromosome-specific subject missingness filters, which removes subjects that are missing more than 5% of SNPs on a particular chromosome, is also implemented. Heterozygosity checks are run after call rate filters. This procedure estimates the total heterozygosity among the variants genotyped across the genome. Extreme low and high heterozygosity values are used to flag subjects with potential DNA quality issues, e.g. sample contamination (extreme high) or low quality DNA (low heterozygosity). Subjects with heterozygosity values more than four standard deviations from the mean value are noted as "extreme" and are filtered out. Figure 1 illustrates how this extreme heterozygosity is determined. After the heterozygosity checks, Rayner's Perl script titled "1000G Imputation preparation and checking" is used to check variant positions and compare allele frequencies against 1000Genomes, a widely utilized reference genome. Rayner's script also updates variant genomic position information to HG19. Lifting over to the HG19 genome build provides uniformity across the datasets and ensures that all SNPs are located on the positive strand. Ambiguous or palindromic variants that cannot be resolved are also removed. Ambiguous SNPs are variants that have a major and minor allele indistinguishable on either strand of the chromosomes (e.g. A/T and C/G). For example, in 1000Genomes, a variant's minor allele is an 'A' with a 40% frequency rate. The alternative allele is a 'T' with a 60% frequency rate. In our study samples, the same variant's allele 'A' is labeled with a 45% frequency rate. Since this frequency is close to both the frequency of 40% and 60%, it would be difficult to tell whether the minor allele is labeled correctly, and since both strands contain A and T ambiguous alleles, the reference

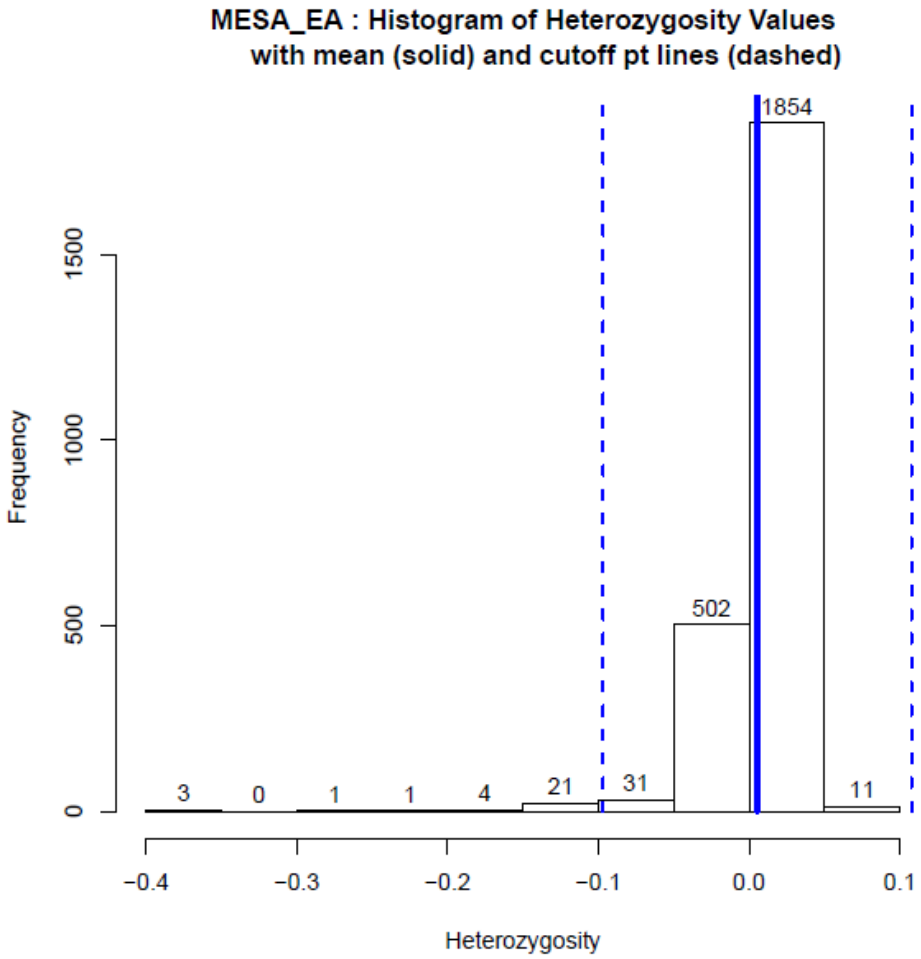

Figure 1: Heterozygosity Plot

This is the heterozygosity plot for the European samples in the MESA cohort. The procedure estimates the total heterozygosity across the variants genotyped in the genome. Values that are extremely high or low are removed. The solid blue line above shows the mean value of the plot and the dashed blue lines denote values four standard deviations from the mean. On the left of the blue line, five samples are above four standard deviations away from the mean, and are therefore removed.

genome does not help with distinguishing between minor and major allele frequency rates. These problematic alleles are removed by the Rayner Perl script.

After the Perl script, an additional check involves checking if variant allele frequencies differ from the 1000Genomes reference by more than 20% for common variants (minor allele frequency greater than 2%) or 5% for rare variants (minor allele frequency less than 2%) is implemented. This step ensures that SNP frequencies do not differ significantly from the already established reference genome. The SNP frequency check is performed with different reference allele frequencies for each ethnic/racial group. At the end of Step 2, an Identity by State (IBS) analysis is ran and a manual check is required. IBS looks at nucleotide differences amongst a group of people and predicts any familial relationship within the group. It is important for individuals who are related to be identified since naive analysis GWAS will result in inflated significant estimates. IBS can also be utilized to identify problematic samples by detecting individuals with many spurious relationships to other individuals in the study. The spurious association is driven by excessive heterozygosity due to sample contamination or DNA degradation. These subjects are removed from the analysis.

In Step 3, a Principle Components Analysis (PCA) is performed on the wide set of variants. PCA is a machine-learning statistical data reduction technique where variables are summarized and simplified into what is known as their most important factors. In our example, every SNP is a variable, or a feature of a subject. A covariance matrix is made where correlations between all combinations of SNPs are calculated. Then, the eigenvalues and eigenvectors of the covariance matrix are calculated. The eigenvector is a vector that defines a direction in which points have the maximum variance when projected onto the vector. In another words, these new eigenvectors provide a new axis that allow us to better visualize the differences amongst the data

All Samples - Known Duplicates & Family subjects removed				
Numer Subjects Removed	Total # pairs (ibs ≥ 5%) (after removal)	Unique Subjects (after removal)	Max IBS pairs/sample (after removal)	Subject ID (removd)
0	57	99	12	0:13149
1	45	86	2	11650:12160
2	43	85	2	11786:11786
3	41	82	1	11121:11973
4	40	80	1	11257:12646
5	39	78	1	0:13129
6	38	76	1	0:10036
7	37	74	1	10862:10862
8	36	72	1	0:10319
9	35	70	1	0:12277
10	34	68	1	0:10163
11	33	66	1	13201:13201
12	32	64	1	13423:13423
13	31	62	1	11724:11724
14	30	60	1	0:10902
15	29	58	1	0:12794
16	28	56	1	0:12980
17	27	54	1	0:12139
18	26	52	1	12264:14285
19	25	50	1	0:13798
20	24	48	1	14454:10004

Figure 2: MESA EA IBS Cutoff Table

This is the resulting table of the Identity by State analysis, which looks for nucleotide differences amongst a group of people to predict relatedness. The example is from the European ancestry samples of the MESA study. The IBS analysis shows that the highlighted subject pair has a spurious association with twelve people. In order to remove this association, one of the subject is removed.

points. Eigenvalues correspond to the percentage of variance of each eigenvector. In the end, we are left with new variables that are constructed as a linear combinations of the initial variables. PCA1 and PCA2 explain a large amount of the variability of the initial points, and are often able to differentiate subjects into groups of continental ancestry. After calculating PCA, Step 3 projects PCA1 and PCA2 values into the HapMap space. The International HapMap Project is an organization aimed to provide a reference haplotype map of the human genome [35]. The HapMap Project contains subjects from eleven population groups, representative of all continental ancestries with complete genome-wide genotype data. PCA is performed and graphed, and study subjects are projected onto the same space and any outliers are manually removed based on their self-reported ancestry. Figure 3 and Figure 4 show an example of the PCA check process. The PCA analysis generates a set of components that explain variability in genetic data and may capture the different allele frequencies that underlie continental ancestry. It is of utmost importance to ensure that the labeled ethnicity (in Step 2) is in accordance with their actual ethnicity. In addition, to address potential admixture (allele frequencies between racial/ethnic groups), another benefit of separating individuals by ancestry is the elimination of covariates. Specific ethnic/racial groups share common characteristics including population history, culture, and diet that can be accounted for if analysis is isolated within the group.

Step 4 of the QC pipeline consists of additional SNP and subject quality control checks. The first check is the Hardy Weinberg Equilibrium (HWE) filter. HWE states that allele frequencies, in the absence of particular evolutionary influences, will remain constant. Though evolutionary factors might be present, we do not expect to see SNP genotypes to vary too much from HWE values and if they do, genotyping error is the most likely cause. Step 4 also checks for SNP missingness prediction by plate information, haplotype block and case control status if

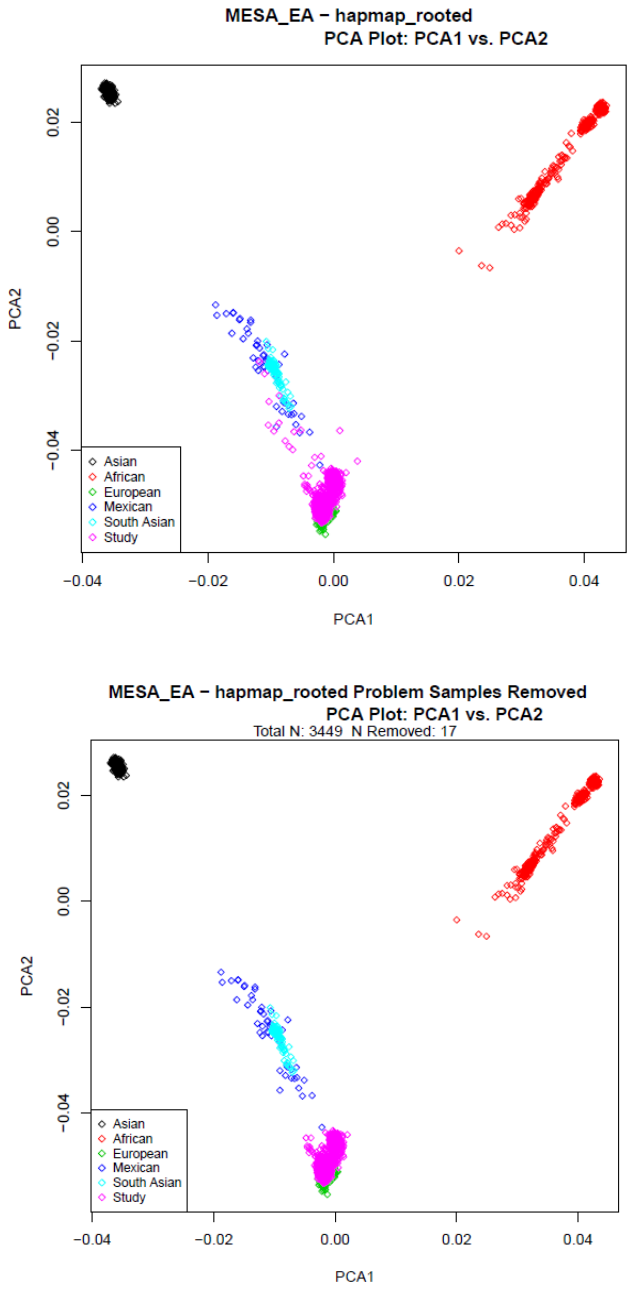

Figure 3: MESA EA PCA Plots Before and After Outlier Removal

Both of these plots are from the MESA European ancestry cohort. In these graphs, PCA1 is mapped against PCA2. PCA1 and PCA2 capture the most variability in the genetic variants and are able to differentiate continental ancestry. PCA's are projected onto HapMap space and HapMap samples are used as a reference to eliminate outliers. On the top graph (before outlier removal), most study samples (pink points) are clustered on the European HapMap samples (green points). Outliers of the main cluster are removed to reduce population admixture. The bottom graph uses the same color and coding scheme, except with outlier subjects removed.

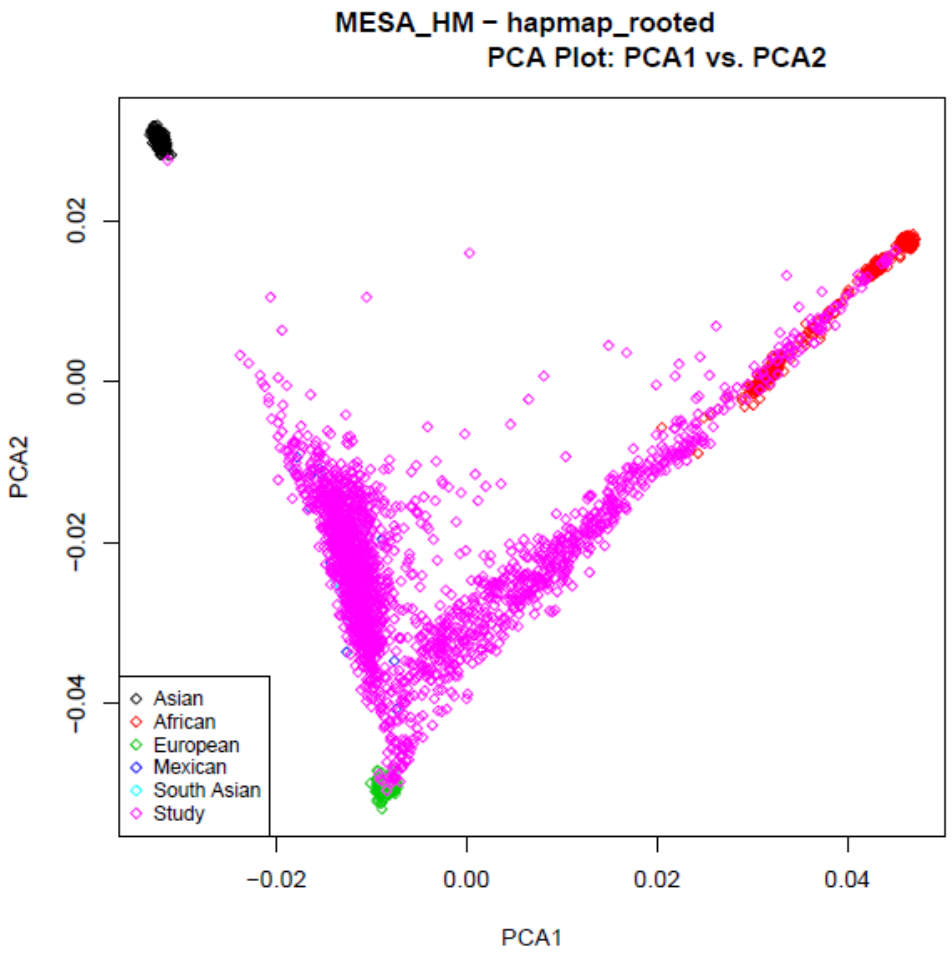

Figure 4: MESA HM PCA Plot

This PCA plot shows the Hispanic subset of the MESA cohort. The extent population admixture of individuals with Hispanic descent is shown in this plot. Genetically, these individuals overlap with multiple ethnic groups. For Hispanic samples like this one, we do not remove any individuals from the study.

available. This missingness check eliminates variants that are missing based on its association with a particular genotyping plate or a particular haplotype block. The last check in Step 4 is the Mendel Test, where individuals with documented family structures are leveraged to check for problematic genotype data. For this check, variants are removed if alleles are not compatible within a family, (e.g. child's genotype is not compatible with parents, indicating a Mendelian error).

Step 5 is the last step for the Quality Control protocol. In Step 5, Principle Components are re-run with the final set of QC'ed subjects for later use as covariates in our GWAS. Final checks and file integrity storage are also performed in this step. Cohorts that have successfully underwent this quality control pipeline are ready for the next genotype imputation.

Imputation

After the meticulous checks in the QC pipeline, variants are guaranteed to be of optimal quality for analysis and genotype imputation. Imputation is a statistical inference method to infer and assign unobserved genotypes. Whole genome genotyping is the process of genotyping SNPs based on a microarray designed to capture a certain predetermined number of variants. Since whole genome genotyping only captures a limited number of SNPs, imputation is used to impute in the missing information of ungenotyped variants. Imputation leverages the linkage disequilibrium (correlation) between genetic variants and haplotypes from a reference population to output probabilistic alleles that fill in for missing variants. Since imputation is computationally intensive, the process is performed on the NIH-funded Michigan Imputation Server [36]. The Michigan Imputation Server provides free imputation services with multiple reference panels and a phasing output option. Overall, imputation increases the power for genetic association studies

by increasing the number of variants and aids in the interpretation of results by including a complete set of variants for analysis.

Variable Harmonization

When working with dbGaP data, variable and data tables come from all sources. Variable harmonization is necessary for case-control selection. For example, lab results are often reported in units that are not standardized across studies. For this project, HDL, LDL, triglycerides, and fasting glucose need to be converted into scientific units, e.g. milligrams per deciliter. Waist circumference is converted into centimeters. Another issue with variables is the naming conventions that are unique to each study. Since all studies use different names for these biomarkers, a uniform naming scheme is implemented across studies. Variable selection is also an issue, as most studies have either multiple variables for a single biomarker or confusing annotations for variables. Longitudinal studies are particularly challenging to work with because of the naming conventions and data tables that often change across the duration of the study. To work with these challenges, R Studios is used to merge tables and rename variables. Variables are also checked for missingness to ensure accurate measurements of all biomarkers. At the end of the harmonization process, all studies have a unique data table with consistent units and uniform variables ready for case-control selection.

Genome Wide Association Study (GWAS)

Genome Wide Association Study (GWAS) is a statistical genetics analysis procedure that performs a hypothesis-free association analysis on hundreds of thousands of genome wide variants for the trait of interest. Essentially, a case-control GWAS compares allele frequencies

that differ significantly between a case group and a control group after covariate and ancestry adjustments. In our MetS case-control GWAS, case subjects are selected based on the criteria in Table 1 and Table 2. For control subjects, we used two definitions: a super control group and a normal control group, running separate GWAS for each case-control set. The super control group consists of individuals with no abnormalities in terms of the five MetS components, while normal controls have two or less criteria of MetS. GWAS is performed using SNPTEST, a program created at the University of Oxford [40].

Meta-Analysis

Meta-analysis is a statistical methodology to combine summary statistics from multiple studies into a single combined summary statistic. GWAS outputs a list of p-values, odds ratios/beta estimates, along with the effect allele and allele frequencies for each variant are required. The summary statistics parameters serve as input values for the meta-analysis. Meta-analysis is performed using the program Metal [41]. Metal combines the effect size and standard errors from multiple cohorts into a single summary statistic that is comparable to performing the analysis with all cohorts combined. This program uses an inverse variance-weighted meta-analysis technique where the directionality of p-values on effect alleles and non-effect alleles matter. This approach takes into account the direction of effect (beta) and weights based on sample size (inverse of standard error) to calculate the combined test statistic.

Chapter III

Results

Studies of Interest

In total, 113 studies and 306,898 subjects have been fully quality controlled and imputed. Table 3 shows a breakdown of the number of cohorts from each ethnicity. These studies, which all originate from dbGaP, are prepared for future GWAS analysis. Within this list of imputed cohorts, three studies of interest were selected for MetS analysis. The first cohort of participants is from the Multi-Ethnic Study of Atherosclerosis, or MESA [37]. MESA is a longitudinal family study that focuses on characteristics of subclinical cardiovascular disease and risk factors associated with it. MESA has 2,231 study samples of European ancestry (EA) and 1,573 study samples of African ancestry (AA) genotyped on the Affymetrix 6.0 array. The study contains 822,855 SNPs from the EA cohort and 824,729 SNPs from the AA cohort after the quality control protocol. The second group of participants is from the Atherosclerosis Risk in Communities, or ARIC study [38]. ARIC is a longitudinal cohort study sponsored by the National Heart, Lung and Blood Institute (NHLBI). This project is a prospective epidemiology study designed to look at the etiology, risk factors and natural history of atherosclerosis. ARIC has 8,401 study samples of European ancestry and 2,678 participants of African ancestry. These subjects were genotyped on the Affymetrix 6.0 array. ARIC contains 696,934 EA variants and 748,870 AA variants after QC. Cardiovascular Health Study (CHS) is the third and final study utilized in this project [39]. CHS is a longitudinal study focused on finding associations of risk factors with CVD and stroke. The study has 3,191 EA study samples and 562 AA study participants genotyped on the Illumina Human CNV370v1 array. This study has 314,705 EA SNPs and 317,863 AA SNPs. All three studies were imputed on the Michigan Imputation Server

Table 3: Quality Control and Imputation Progress

Race Codes	N Cohorts	N
AA: African Ancestry	27	47,130
AS: Asian Ancestry	7	10,325
EA: European Ancestry	59	212,264
HM: Hispanic Mixed Ancestry	18	33,814
SA: South Asian Ancestry	1	1,432
Total	**113**	**306,898**

with the HRC (Version r1.1 2016) imputation panel, which is comprised of 64,940 haplotypes. After imputation, the MESA cohort had 41,626,293 EA SNPs and 41,626,596 AA SNPs. The ARIC cohort had 41,626,437 EA SNPs and 41,626,564 AA SNPs. The CHS cohort had 41,583,825 EA SNPs and 41,583,826 AA SNPs.

GWAS Summary

We performed a GWAS study on 18,636 samples (n= 13,823 of European and n=4813 African ancestry) from three individual studies. Summary statistics for each cohort is reported in detail in Table 4 and Table 5. GWAS was performed stratified by racial groups (European and African American) for two case-control definitions (normal and super controls). The test statistic was filtered based on null values, genomic inflations and low-allele frequency values (effect and non-effect alleles must both have over ten total alleles). A meta-analysis was performed to combine and generate a single test statistic for each variant.

European Ancestry: Cases and Normal Controls

GWAS of the European cases and normal control set of subjects identified four genome-wide significant SNPs, reported in detail in Table 9. Genome-wide significant SNPs have a p-value of less than 5x10-8. The first SNP (rs11076176) is located on the *CETP* gene and has a p-value of 3.05×10^{-8}. The *CETP* gene, also known as cholesteryl ester transfer protein, encodes a plasma protein that facilitates the transport of cholesteryl esters and triglycerides between lipoproteins. The second variant (rs247616) is located on intergenic regions of chromosome 16

Table 4: Demographics, Case Control Breakdown, and Medication Status of Study Samples

	MESA		ARIC		CHS	
	European Ancestry	**African Ancestry**	**European Ancestry**	**African Ancestry**	**European Ancestry**	**African Ancestry**
Subjects after Quality Control (N)	2231	1573	8401	2678	3191	562
Men	1189	842	4084	1031	1267	209
Woman	1042	731	4317	1647	1924	353
Metabolic Syndrome Cases	1045	849	4813	1735	1453	242
Metabolic Syndrome Super Controls	700	304	2701	591	935	117
Metabolic Syndrome Normal Controls	1186	724	3588	942	1738	320
Individuals on Hypoglycemic Medication (%)	4.5%	13.4%	3.8%	14.6%	1.3%	3.9%
Individuals on Anti-Hypertensive Medication (%)	33.5%	50.2%	23.0%	52.1%	36.4%	56.4%
Individuals on Hypolipidemic medication (%)	17.8%	15.6%	3.5%	1.4%	4.4%	5.7%

Table 5: Clinical Characteristic of Study Samples

	Mean + SD					
	MESA		ARIC		CHS	
	European Ancestry	African Ancestry	European Ancestry	African Ancestry	European Ancestry	African Ancestry
Age	62.8 + 10.2	62.3 + 10.1	54.4 + 5.7	53.4 + 5.8	72.4 + 5.4	72.6 + 5.5
Waist Circumference (cm)	98.3 + 14.1	101.2 + 14.6	96.4 + 13.2	99.4 + 15.0	93.3 + 12.8	98.5 + 14.1
Triglycerides (mg/dl)	132.2 + 89.3	105.3 + 71.0	138.2 + 91.5	116.5 + 87.3	140.5 + 74.4	113.0 + 54.3
HDL Cholesterol (mg/dl)	51.8 + 15.7	52.3 + 15.3	50.2 + 16.5	54.7 + 17.2	55.0 + 15.8	58.63 +15.5
Systolic Blood Pressure (mm)	123.5 + 20.1	131.7 + 21.8	118.6 + 16.9	128.7 + 21.2	135.0 + 21.5	140.4 + 21.0
Diastolic Blood Pressure (mm)	70.5 + 9.8	74.5 + 10.3	71.6 + 10.0	80.0 + 12.3	70.1 + 11.8	74.7 + 10.2
Fasting Glucose (mg/dl)	91.5 + 21.6	100.3 + 32.5	105.3 + 31.1	118.7 + 58.3	107.8 + 28.8	117.7 + 53.7

and has a p-value of 5.88×10^{-10}. The third SNP (rs287) lies on the *LPL* gene on chromosome 8 and has a p-value of 1.64×10^{-13}. LPL, also known as lipoprotein lipase, encodes for an enzyme that hydrolyzes triglycerides in lipoproteins. Finally, the last significant SNP (rs964184) is located on the *ZPR1* gene and has a p-value of 5.18×10^{-12}. The *ZPR1* gene encodes for a zinc finger protein. The zinc finger protein is known to interact with survival motor neuron proteins to enhance pre-mRNA splicing. We also generated Quantile-Quantile (QQ) plots to visually check for inflation. The QQ plot is a graphical technique used to verify if a set of test statistics comes from a certain distribution. For GWAS, we expect to see p-values that follow a uniform distribution with an inflated "tail" if true associations exist. The study's p-value is graphed on the y-axis against a uniform distribution of p-values on the x-axis. QQ plots in this study show a slight inflation, with genomic control variable of 1.011 (Figure 5, Pane 2). Manhattan plots in Figure 8 show thirty-six genome-wide suggestive SNPs, or variants that have a reported p-value of 1×10^{-5} or less. The Manhattan plot is a type of scatter plot for GWAS visualization that graphs the negative logarithms of the p-values on the y-axis with the chromosome positions on the x axis. The suggestive SNPs from the Manhattan plots are reported in Table 12.

European Ancestry: Cases and Super Controls

For the European cases and super control set of subjects, QQ plots in Figure 5, Pane 1, show slight inflation with the genomic control variable of 1.014. Like the normal controls, there are also four genome-wide significant SNPs, reported in Table 5. The normal control subset was able to capture the same four genome-wide significant SNP as the super control subset. However, super controls reported more suggestive SNPs than the normal controls. The first genome-wide

significant SNP (rs11076176) that fall on the *CETP* gene has a p-value of 3.91×10^{-8}. The second variant (rs247616) is on an intergenic region and has a p-value of 2.95×10^{-8}. The third SNP (rs287) is on the *LPL* gene and has a p-value of 2.16×10^{-13}. The final SNP (rs964184) is on the *ZPR1* gene and has a p-value of 1.13×10^{-12}. Manhattan plots in Figure 6 show forty-one genome-wide suggestive SNPs, which are reported in Table 10.

African Ancestry: Cases and Normal Controls

Figure 5, Pane 4, shows the QQ plots for the cases and normal control subjects with African ancestry. With the genomic control variable of 0.983, the graph shows a slight deflation. As seen in the Manhattan plots in Figure 9, only one SNP (rs117729532) reached genome-wide significance with the p-value of 1.05×10^{-8}. This SNP is also duplicated as a genome-wide significant variant in the cases and super control subset. Cases and normal controls were able to capture sixty-nine suggestive signals listed in Table 13.

African Ancestry: Cases and Super Controls

Lastly, the QQ plots shown for cases and super controls for subjects of African ancestry are shown in Figure 5, Pane 3. There is a minor inflation with genomic control variable of 1.01. There were two SNPs that reached genome-wide significance, shown in the Manhattan plot in Figure 7. The first SNP (rs117729532), with a p-value of 2.28×10^{-9}, is on chromosome 7 and belongs to a non-coding transcript variant. This SNP resides on an uncharacterized gene known as *LOC107986717*. The second SNP (rs115553887) has a p-value of 4.38×10^{-8} and resides on the *RBM20* gene. The *RBM20* gene codes for a protein that regulates the splicing of the *TTN* gene. There were 150 variants that were genome-wide suggestive, which are listed in Table 11.

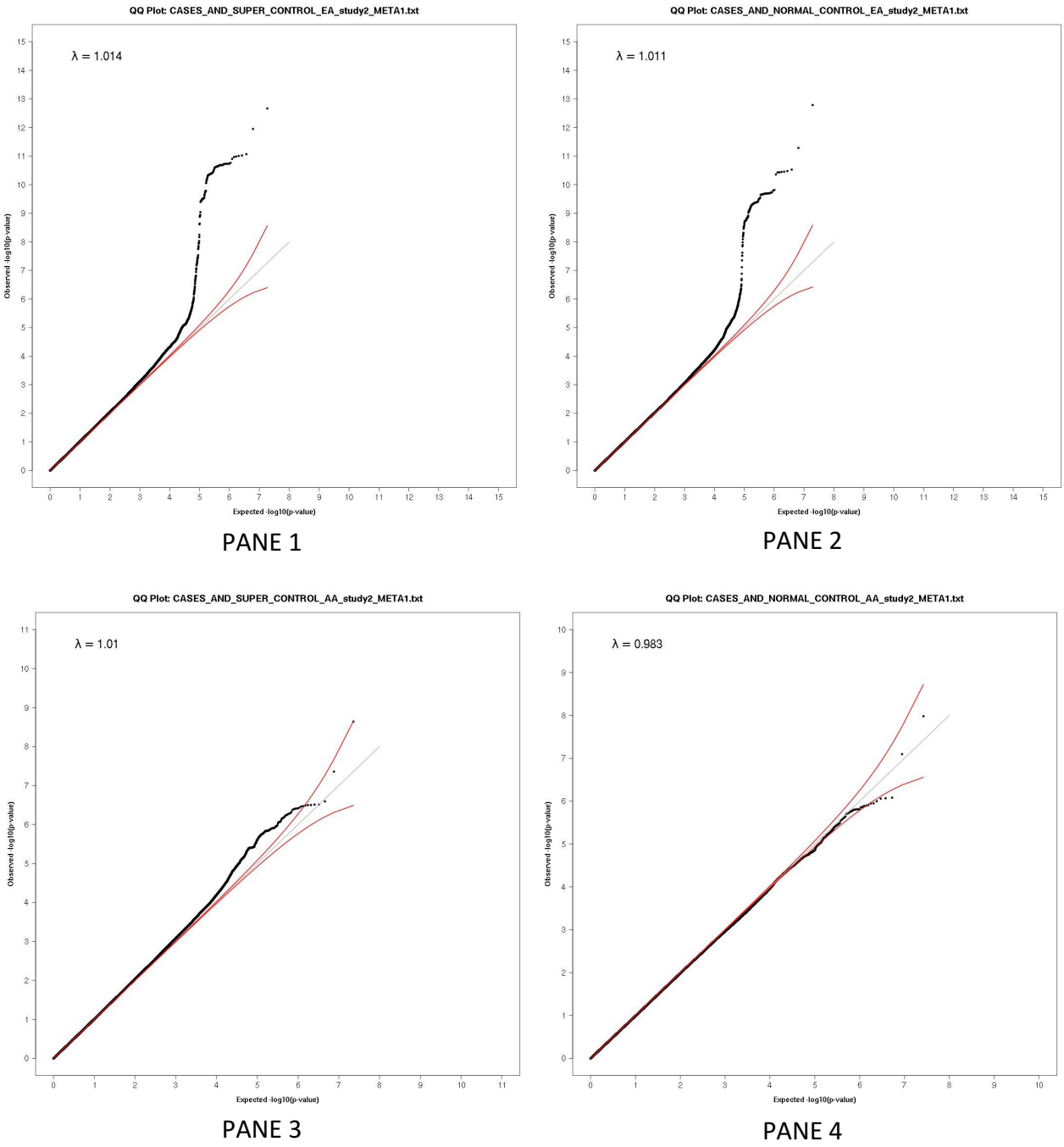

Figure 5: QQ Plots

The QQ plot is a graphical technique that visualizes expected versus observed p-value distributions of a GWAS. Note: p-values on x- and y-axis are reported in negative logarithm scale. Pane 1 shows the QQ plots and genomic control variable in the European cases and super controls subset. Pane 2 shows the plot for the European cases and normal controls subset. Pane 3 shows the plot for the African cases and super controls subset. Pane 4 shows the plot for the African cases and normal controls subset.

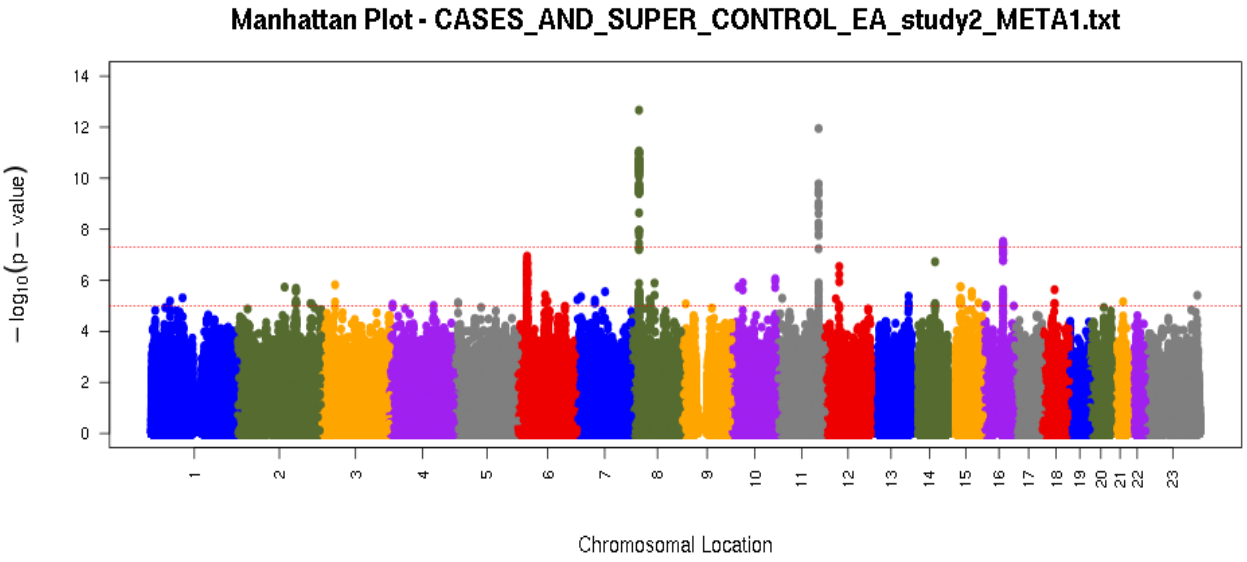

Figure 6: Manhattan Plot: European Ancestry, Cases and Super Controls

A Manhattan Plot is a scatter plot used to display p-values from a GWAS. The x-axis represents the linearized genomic position while the y axis represents the negative logarithm of the p-values. Four points reach genome-wide significance (top threshold line) in this plot: rs287 on chromosome 8, rs964184 on chromosome 11, rs11076176 on chromosome 16 and rs247616 on chromosome 16. These points have a p-value of less than $5x10^{-8}$. Forty-one points reach genome-wide suggestive (bottom threshold line) in this plot. These points have a p-value of less than $1x10^{-5}$.

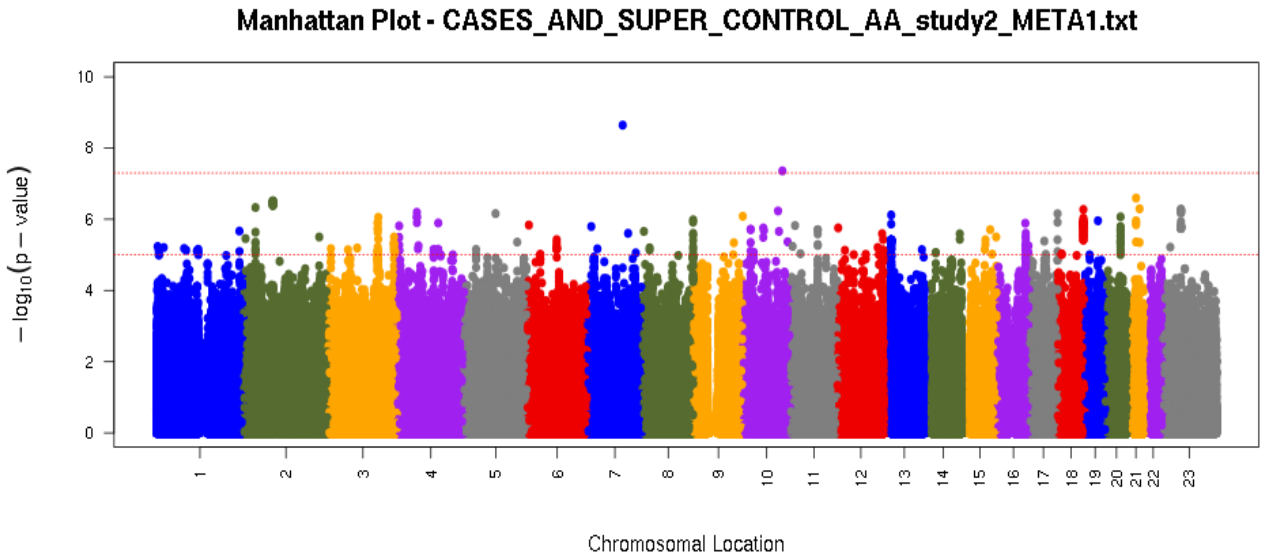

Figure 7: Manhattan Plot: African Ancestry, Cases and Super Controls

Two loci reach genome-wide significance (top threshold line) in this plot: rs117729532 on chromosome 7 and rs115553887 on chromosome 10. These points have a p-value of less than $5x10^{-8}$. 150 points reach genome-wide suggestive (bottom threshold line) in this plot. These points have a p-value of less than $1x10^{-5}$.

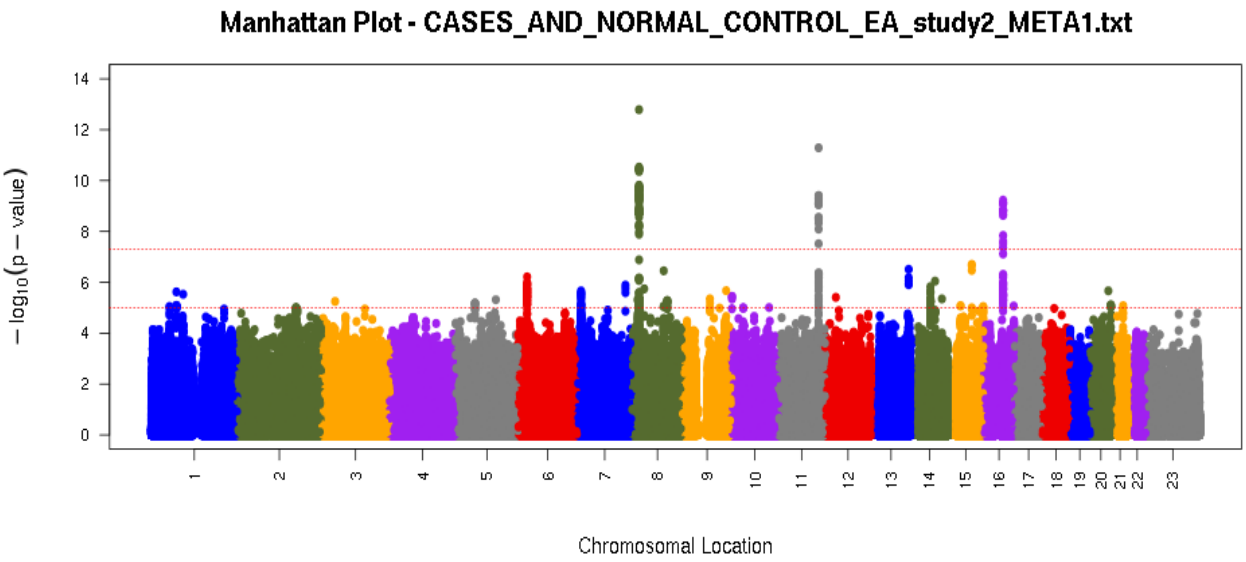

Figure 8: Manhattan Plot: European Ancestry, Cases and Normal Controls

Four loci reach genome-wide significance (top threshold line) in this plot: rs287 on chromosome 8, rs964184 on chromosome 11, rs11076176 on chromosome 16 and rs247616 on chromosome 16. These points have a p-value of less than $5x10^{-8}$. Thirty-six points reach genome-wide suggestive (bottom threshold line) in this plot. These points have a p-value of less than $1x10^{-5}$.

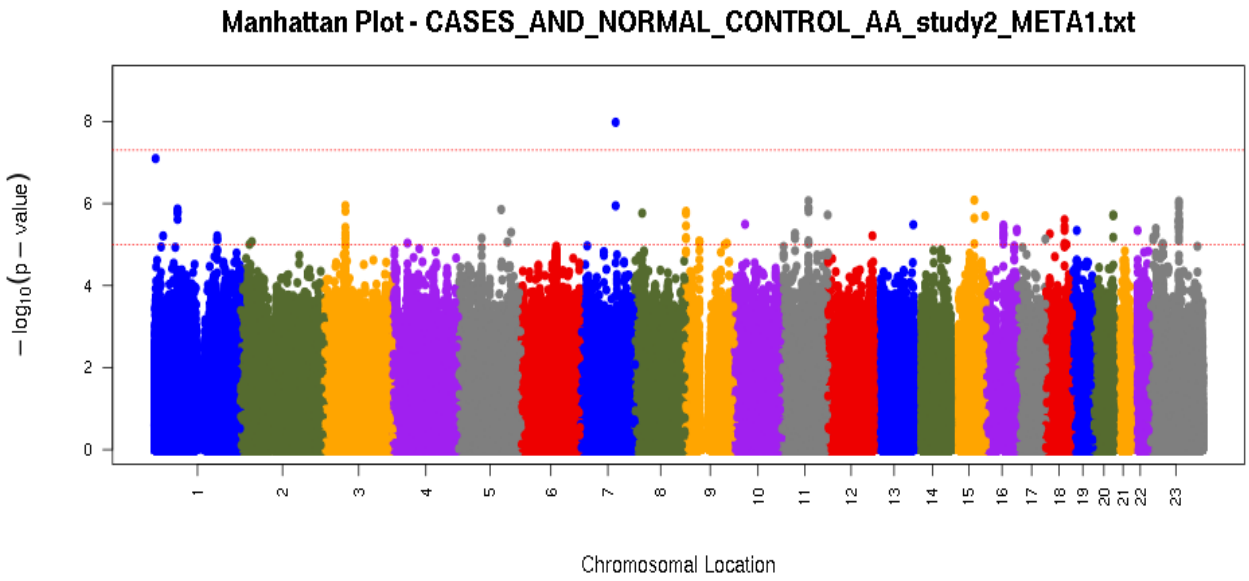

Figure 9: Manhattan Plot: African Ancestry, Cases and Normal Controls

One locus reached genome-wide significance (top threshold line) in this plot: rs117729532 on chromosome 7. This point has a p-value of less than $5x10^{-8}$. Sixty-nine points reach genome-wide suggestive (bottom threshold line) in this plot. These points have a p-value of less than $1x10^{-5}$.

Table 6**:** Genome-Wide Significant Variants for European Cases and Super Controls

CHR	SNP	Gene	EA	Freq EA	Beta	StdErr	P-value	OR
8	rs287	LPL	A	0.748	0.237	0.0323	2.16E-13	1.27
11	rs964184	ZPR1	C	0.8606	-0.2877	0.0404	1.13E-12	0.75
16	rs11076176	CETP	T	0.8348	-0.2131	0.0388	3.91E-08	0.81
16	rs247616	Intergenic	T	0.3236	-0.1659	0.0299	2.95E-08	0.85

Table 7: Genome-Wide Significant Variants for European Cases and Normal Controls

CHR	SNP	Gene	EA	Freq EA	Beta	StdErr	P-value	OR
8	rs287	LPL	A	0.7455	0.2108	0.0286	1.64E-13	1.23
11	rs964184	ZPR1	C	0.8616	-0.2477	0.0359	5.18E-12	0.78
16	rs11076176	CETP	T	0.8355	-0.1914	0.0346	3.05E-08	0.83
16	rs247616	Intergenic	T	0.3339	-0.1639	0.0265	5.88E-10	0.85

Table 8: Genome-Wide Significant Variants for African Cases and Super Controls

CHR	SNP	Gene	EA	Freq EA	Beta	StdErr	P-value	OR
7	rs117729532	LOC107986717	A	0.9734	2.3672	0.3961	2.28E-09	10.67
10	rs115553887	RBM20	A	0.0189	-2.1996	0.4018	4.38E-08	0.11

Table 9: Genome-Wide Significant Variants for African Cases and Normal Controls

CHR	SNP	Gene	EA	Freq EA	Beta	StdErr	P-value	OR
7	rs117729532	LOC10798671	A	0.9709	1.1885	0.2077	1.05E-08	3.28

Table 10: Genome-Wide Suggestive Variants for European Cases and Super Controls

CHR	SNP	Gene	EA	Freq EA	Beta	StdErr	P-value	OR
1	rs72662344	Intergenic	C	0.2618	-0.1495	0.0331	6.47E-06	0.86
1	rs564021270	Intergenic	A	0.0046	-1.2447	0.2723	4.84E-06	0.29
2	rs142316818	Intergenic	A	0.005	-1.2963	0.2718	1.85E-06	0.27
2	rs2740590	G6PC2	A	0.158	0.1856	0.0391	2.05E-06	1.2
2	rs145888022	Intergenic	A	0.0104	-0.7185	0.161	8.12E-06	0.49
2	rs142610152	VWC2L	T	0.0098	-0.9244	0.2075	8.36E-06	0.4
3	rs142562516	Intergenic	C	0.0089	-1.0535	0.219	1.51E-06	0.35
4	rs186595162	Intergenic	T	0.0044	-1.2392	0.2801	9.66E-06	0.29
4	rs73085842	Intergenic	A	0.8929	-0.2026	0.0455	8.63E-06	0.82
5	rs274646	Intergenic	A	0.2461	-0.1459	0.0326	7.57E-06	0.86
6	rs6942072	Intergenic	A	0.3136	-0.1583	0.0299	1.14E-07	0.85
6	rs1778258	Intergenic	T	0.1414	0.1863	0.0403	3.77E-06	1.2
6	rs117485832	CEP162, MRAP2	T	0.0044	-1.4313	0.3175	6.53E-06	0.24
7	rs77118159	PHF14	A	0.9817	0.5188	0.113	4.40E-06	1.68
7	rs17712441	Intergenic	C	0.0334	-0.3705	0.0818	5.95E-06	0.69
7	rs138277611	DNAAF5	A	0.0114	-1.0872	0.2397	5.76E-06	0.34
7	rs1208090	Intergenic	A	0.5484	0.1415	0.0302	2.80E-06	1.15
8	rs287	LPL	A	0.748	0.237	0.0323	2.16E-13	1.27
8	rs80283638	Intergenic	T	0.9719	0.3994	0.0879	5.58E-06	1.49
8	rs2885164	Intergenic	A	0.0386	0.3264	0.0728	7.31E-06	1.39
8	rs148800856	LINC01289	T	0.0116	-0.7456	0.154	1.29E-06	0.47
9	rs139751296	PTPRD	T	0.0143	-0.6621	0.1486	8.33E-06	0.52
10	rs4021528	Intergenic	T	0.3125	0.1573	0.032	8.70E-07	1.17
10	rs12765306	PLXDC2	T	0.0292	-0.4092	0.0858	1.83E-06	0.66
10	rs118171930	Intergenic	T	0.977	0.5027	0.1036	1.23E-06	1.65
11	rs117182743	GALNT18	A	0.0192	-0.4906	0.1075	5.03E-06	0.61
11	rs964184	ZPR1	C	0.8606	-0.2877	0.0404	1.13E-12	0.75
11	rs7925256	PAFAH1B2	T	0.0902	0.2223	0.0494	6.95E-06	1.25
12	rs7925256	DENND5B	A	0.8322	0.1787	0.0393	5.31E-06	1.2
12	rs145305667	Intergenic	T	0.992	1.2207	0.2379	2.87E-07	3.39
13	rs76345073	Intergenic	T	0.0286	-0.4276	0.093	4.22E-06	0.65
14	rs8004913	GALNT16	T	0.8536	-0.2066	0.0396	1.88E-07	0.81
15	rs11073147	Intergenic	A	0.455	-0.1321	0.0276	1.77E-06	0.88
15	rs58293302	LOC105370873	A	0.2016	-0.1773	0.0379	2.86E-06	0.84
15	rs117222771	NTRK3	A	0.9738	0.4309	0.0963	7.64E-06	1.54
16	rs58293302	Intergenic	T	0.3236	-0.1659	0.0299	2.95E-08	0.85
16	rs11076176	CETP	T	0.8348	-0.2131	0.0388	3.91E-08	0.81
16	rs6416820	RBFOX1	T	0.1117	-0.2002	0.0452	9.44E-06	0.82
18	rs113185854	Intergenic	T	0.015	-0.6019	0.1275	2.34E-06	0.55
21	rs77084313	Intergenic	A	0.1059	-0.2073	0.0461	6.92E-06	0.81

Table 11: Genome-Wide Suggestive Variants for African Cases and Super Controls

CHR	SNP	Gene	EA	Freq EA	Beta	StdErr	P-value	OR
1	rs116612854	Intergenic	T	0.0172	-1.3698	0.306	7.60E-06	0.25
1	rs114058333	Intergenic	T	0.0145	-1.4003	0.3138	8.13E-06	0.25
1	rs59797519	Intergenic	T	0.9788	1.0199	0.2288	8.27E-06	2.77
1	rs72985317	Intergenic	T	0.0214	-1.0221	0.2274	6.98E-06	0.36
1	rs113675722	Intergenic	A	0.0207	-1.0598	0.2398	9.90E-06	0.35
1	rs188888281	Intergenic	A	0.0134	-1.4898	0.3297	6.22E-06	0.23
1	rs12038715	RYR2	T	0.2309	-0.3661	0.0773	2.15E-06	0.69
1	rs80130145	RYR2	T	0.0151	-1.3471	0.3018	8.07E-06	0.26
1	rs10909951	MEGF6	T	0.2137	0.381	0.0841	5.84E-06	1.46
1	rs45510693	TNFRSF25	C	0.0173	-1.3875	0.3139	9.85E-06	0.25
1	rs202049535	PLEKHG5	T	0.9835	1.4392	0.3215	7.56E-06	4.22
1	rs1686341	Intergenic	A	0.5217	0.2551	0.0566	6.63E-06	1.29
1	rs58469176	SSX2IP	A	0.0152	-1.4022	0.3129	7.44E-06	0.25
2	rs80353771	PECR	T	0.0916	-0.4923	0.1056	3.16E-06	0.61
2	rs2098566	LINC01317	T	0.224	-0.3365	0.0668	4.69E-07	0.71
2	rs115261557	Intergenic	T	0.9714	0.9596	0.2068	3.49E-06	2.61
2	rs113998642	Intergenic	A	0.0299	-1.2497	0.2469	4.16E-07	0.29
2	rs76699442	Intergenic	T	0.9701	1.2516	0.247	4.04E-07	3.5
2	rs191383538	Intergenic	T	0.0299	-1.2535	0.2471	3.92E-07	0.29
2	rs142388955	Intergenic	A	0.9701	1.2541	0.2471	3.88E-07	3.5
2	rs115767024	Intergenic	C	0.0299	-1.2547	0.2472	3.84E-07	0.29
2	rs115764561	Intergenic	A	0.0299	-1.2611	0.2474	3.43E-07	0.28
2	rs116132520	Intergenic	T	0.0299	-1.2619	0.2474	3.37E-07	0.28
2	rs142723182	Intergenic	A	0.0299	-1.2634	0.2473	3.25E-07	0.28
2	rs76160666	Intergenic	A	0.0299	-1.264	0.2472	3.17E-07	0.28
2	rs116619331	Intergenic	A	0.0299	-1.2641	0.2472	3.16E-07	0.28
2	rs116475445	Intergenic	T	0.9701	1.2643	0.247	3.09E-07	3.54
2	rs114898961	Intergenic	T	0.0299	-1.2643	0.247	3.06E-07	0.28
2	rs115812434	Intergenic	T	0.9699	1.2526	0.2463	3.67E-07	3.5
3	rs73869641	RASA2	A	0.0378	-0.7569	0.1691	7.60E-06	0.47
3	rs73869652	RASA2	A	0.047	-0.7363	0.1518	1.24E-06	0.48
3	rs56233244	RASA2	T	0.9659	0.8136	0.1824	8.21E-06	2.26
3	rs4437106	Intergenic	T	0.5862	-0.2786	0.0567	8.83E-07	0.76
3	rs35262984	LPP	A	0.2096	0.3368	0.0723	3.16E-06	1.4
3	rs73888515	OSTN	A	0.0259	-1.0223	0.229	8.03E-06	0.36
3	rs112883499	ERC2	T	0.934	0.5201	0.1159	7.17E-06	1.68
3	rs116663894	GRM7	T	0.8429	-0.3567	0.0792	6.70E-06	0.7
3	rs114179158	LINC02008	A	0.9046	0.4721	0.1046	6.40E-06	1.6
4	rs4648057	NFKB1	T	0.047	-0.752	0.169	8.58E-06	0.47
4	rs4648109	NFKB1	C	0.0411	-0.816	0.1816	6.99E-06	0.44
4	rs77605772	ELOVL6	C	0.9824	1.3839	0.3088	7.42E-06	3.99

Table 11: Genome-Wide Suggestive Variants for African Cases and Super Controls (cont'd, 2 of 4)

CHR	SNP	Gene	EA	Freq EA	Beta	StdErr	P-value	OR
4	rs77987747	Intergenic	T	0.9803	1.6855	0.3479	1.27E-06	5.4
4	rs17006188	FGF2	T	0.0309	-0.8462	0.1915	9.92E-06	0.43
4	rs72966662	Intergenic	T	0.0315	-0.7721	0.1748	9.99E-06	0.46
4	rs4689939	Intergenic	T	0.1377	0.3886	0.0809	1.54E-06	1.47
4	rs17827152	Intergenic	A	0.0701	-0.6132	0.1231	6.39E-07	0.54
4	rs563732848	Intergenic	A	0.9772	1.222	0.2715	6.79E-06	3.39
4	rs73199427	Intergenic	A	0.0335	-0.7585	0.1669	5.53E-06	0.47
5	rs1363414	LOC105378237	A	0.7376	-0.2905	0.0633	4.39E-06	0.75
5	rs10041838	ADAMTS12	T	0.9521	0.6215	0.1406	9.83E-06	1.86
5	rs12332388	ADAMTS12	A	0.9517	0.6287	0.1399	7.02E-06	1.88
5	rs9968625	LINC01339	T	0.0825	-0.5305	0.1069	6.94E-07	0.59
6	rs115183512	MTCH1	C	0.9721	0.8885	0.2008	9.69E-06	2.43
6	rs149076725	Intergenic	T	0.9819	1.4461	0.3002	1.46E-06	4.25
6	rs1936820	ME1	T	0.1767	-0.3347	0.0724	3.75E-06	0.72
6	rs191630648	ME1	A	0.8369	0.3442	0.0751	4.53E-06	1.41
7	rs117729532	Intergenic	A	0.9734	2.3672	0.3961	2.28E-09	10.67
7	rs76836910	CTTNBP2, LOC105375469	T	0.0139	-1.7437	0.3703	2.49E-06	0.17
7	rs62433165	none	A	0.0153	-1.8758	0.3909	1.60E-06	0.15
7	rs78672236	HIPK2	A	0.1057	0.4272	0.0961	8.80E-06	1.53
7	rs12537629	CHN2	A	0.0174	-1.2833	0.2852	6.79E-06	0.28
8	rs116248526	TOP1MT	T	0.0221	-1.0575	0.2246	2.49E-06	0.35
8	rs115216114	TOP1MT	C	0.9747	0.916	0.2065	9.22E-06	2.5
8	rs185167231	RHPN1	A	0.0183	-1.4143	0.2896	1.04E-06	0.24
8	rs60457759	RHPN1	T	0.9805	1.3669	0.2816	1.21E-06	3.92
8	rs192827394	RHPN1	T	0.0149	-1.4565	0.3213	5.82E-06	0.23
8	rs138549292	RHPN1	T	0.0168	-1.5066	0.3218	2.84E-06	0.22
8	rs149793922	ZC3H3	T	0.0125	-1.8866	0.4088	3.93E-06	0.15
8	rs75551077	Intergenic	C	0.1314	-0.3741	0.0829	6.34E-06	0.69
8	rs62474695	CSMD1	A	0.0166	-1.7243	0.3642	2.20E-06	0.18
9	rs4135186	TXN	T	0.0154	-1.8939	0.4287	9.96E-06	0.15
9	rs113812899	Intergenic	A	0.9323	0.5743	0.1253	4.58E-06	1.78
9	rs184434626	PNPLA7	A	0.0142	-2.3601	0.4788	8.25E-07	0.09
10	rs74156619	HPSE2	T	0.9045	0.5043	0.1009	5.86E-07	1.66
10	rs6584411	Intergenic	A	0.3963	0.2673	0.0565	2.20E-06	1.31
10	rs115553887	RBM20	A	0.0189	-2.1996	0.4018	4.38E-08	0.11
10	rs14606	ADAM12	C	0.9641	0.8899	0.1936	4.30E-06	2.43
10	rs191806961	Intergenic	A	0.0163	-1.3683	0.2983	4.50E-06	0.25
10	rs528494547	Intergenic	A	0.0259	-0.9343	0.2091	7.86E-06	0.39
10	rs186850483	Intergenic	T	0.022	-1.2356	0.2597	1.95E-06	0.29
10	rs78548751	Intergenic	T	0.9788	1.2385	0.2601	1.91E-06	3.45

Table 11: Genome-Wide Suggestive Variants for African Cases and Super Controls (cont'd, 3 of 4)

CHR	SNP	Gene	EA	Freq EA	Beta	StdErr	P-value	OR
10	rs2505128	JCAD, LOC101929256	A	0.0437	-0.6631	0.1489	8.48E-06	0.52
10	rs141327348	Intergenic	A	0.9777	1.0492	0.2195	1.76E-06	2.86
10	rs60804070	Intergenic	A	0.0229	-1.0209	0.2193	3.22E-06	0.36
10	rs72791504	Intergenic	A	0.0167	-2.0584	0.4337	2.08E-06	0.13
11	rs3816360	ARNTL	T	0.4666	-0.2641	0.0549	1.50E-06	0.77
11	rs142441721	Intergenic	A	0.0135	-1.5495	0.3495	9.27E-06	0.21
11	rs112614723	Intergenic	A	0.9838	2.0166	0.4448	5.79E-06	7.51
11	rs193019036	Intergenic	A	0.9596	0.7165	0.1524	2.59E-06	2.05
11	rs11237567	Intergenic	C	0.9603	0.7259	0.1527	1.98E-06	2.07
11	rs11237568	Intergenic	A	0.0398	-0.7227	0.152	1.97E-06	0.49
11	rs116356381	Intergenic	A	0.0372	-0.7061	0.1549	5.15E-06	0.49
11	rs10219155	Intergenic	A	0.0371	-0.7045	0.1549	5.44E-06	0.49
12	rs36203374	TBX3, LOC105370000	A	0.9459	0.6589	0.1458	6.20E-06	1.93
12	rs11616110	KSR2	A	0.6823	0.2652	0.06	9.81E-06	1.3
12	rs57763252	LINC02376	A	0.1077	0.4625	0.0983	2.55E-06	1.59
12	rs12826896	Intergenic	A	0.9765	1.0443	0.2257	3.72E-06	2.84
12	rs10848575	ADIPOR2	A	0.6588	-0.2792	0.0584	1.75E-06	0.76
12	rs80233242	SLCO1B1	A	0.9818	1.4914	0.3329	7.46E-06	4.44
12	rs138302352	Intergenic	T	0.0221	-1.0143	0.2295	9.90E-06	0.36
12	rs140528976	PTPRQ	A	0.0206	-1.1834	0.2674	9.66E-06	0.31
13	rs114535481	FAM155A	C	0.0202	-1.2709	0.2831	7.12E-06	0.28
13	rs78673614	Intergenic	A	0.2248	-0.3386	0.0685	7.61E-07	0.71
14	rs193050452	Intergenic	C	0.0365	-0.8066	0.1742	3.66E-06	0.45
14	rs77811189	Intergenic	A	0.037	-0.8167	0.1737	2.57E-06	0.44
14	rs116771708	Intergenic	A	0.0222	-1.1085	0.2494	8.79E-06	0.33
14	rs116243311	Intergenic	T	0.9779	1.1095	0.2494	8.62E-06	3.03
15	rs146103401	Intergenic	A	0.9851	1.4626	0.3257	7.08E-06	4.32
15	rs113365583	SMAD3	T	0.0206	-1.1663	0.2557	5.07E-06	0.31
15	rs111902897	SMAD3	T	0.9799	1.1921	0.2578	3.77E-06	3.29
15	rs74395789	CEMIP	T	0.0135	-2.3903	0.5024	1.96E-06	0.09
15	rs145596139	Intergenic	T	0.9825	1.5787	0.3566	9.55E-06	4.85
15	rs114035855	Intergenic	A	0.0148	-1.4071	0.3019	3.15E-06	0.24
16	rs148171727	WWOX	T	0.9416	0.6663	0.1377	1.29E-06	1.95
16	rs9923417	MAF	T	0.0419	-0.7448	0.1637	5.35E-06	0.47
16	rs75735313	MAF	A	0.0417	-0.7378	0.1643	7.09E-06	0.48
16	rs7202443	LOC102724084	A	0.6467	-0.2756	0.0585	2.48E-06	0.76
16	rs4889144	LOC102724084	A	0.7147	0.2796	0.0629	8.80E-06	1.32
16	rs57154787	LOC105369246	T	0.8835	0.4333	0.0973	8.49E-06	1.54
16	rs58421241	LOC105369246	A	0.116	-0.443	0.0975	5.57E-06	0.64

Table 11: Genome-Wide Suggestive Variants for African Cases and Super Controls (cont'd, 4 of 4)

CHR	SNP	Gene	EA	Freq EA	Beta	StdErr	P-value	OR
17	rs28497626	DCAKD	A	0.8813	0.4032	0.0875	4.12E-06	1.5
17	rs145149232	ZNF652	A	0.9725	0.9217	0.2083	9.62E-06	2.51
17	rs140603527	METRNL	A	0.9755	1.3851	0.3104	8.11E-06	4
17	rs145502883	METRNL	A	0.0362	-1.2345	0.2543	1.21E-06	0.29
17	rs147171466	Intergenic	A	0.0348	-1.1439	0.2474	3.77E-06	0.32
17	rs111915469	Intergenic	T	0.9664	1.3306	0.2682	7.00E-07	3.78
18	rs145989159	Intergenic	A	0.9572	0.7387	0.1667	9.30E-06	2.09
18	rs142959114	Intergenic	A	0.0143	-1.4392	0.3256	9.86E-06	0.24
18	rs35042968	LOC339298	A	0.9826	1.3583	0.2709	5.32E-07	3.89
19	rs7249029	CACNA1A	A	0.0517	-0.6395	0.1448	9.97E-06	0.53
19	rs139297424	ZNF420	A	0.0182	-1.2278	0.252	1.10E-06	0.29
20	rs73909514	LINC01430	T	0.9705	1.0755	0.2331	3.97E-06	2.93
20	rs73909522	SERINC3	T	0.985	1.356	0.3068	9.89E-06	3.88
20	rs191592101	PKIG	T	0.0161	-1.3606	0.3003	5.90E-06	0.26
20	rs61208911	PKIG	A	0.0163	-1.3821	0.3021	4.76E-06	0.25
20	rs76224531	PKIG	A	0.9842	1.425	0.3044	2.84E-06	4.16
20	rs73909541	PKIG	A	0.0157	-1.3967	0.303	4.04E-06	0.25
20	rs73909542	PKIG	T	0.0166	-1.3281	0.2988	8.81E-06	0.26
20	rs73909543	PKIG	T	0.0163	-1.3838	0.3023	4.69E-06	0.25
20	rs73909544	PKIG	A	0.0153	-1.4228	0.3048	3.05E-06	0.24
20	rs73909545	PKIG	C	0.0158	-1.4496	0.3065	2.24E-06	0.23
20	rs59434057	PKIG	A	0.0158	-1.4572	0.3058	1.89E-06	0.23
20	rs192774638	PKIG	T	0.9839	1.4289	0.3008	2.03E-06	4.17
20	rs147389339	PKIG	A	0.0157	-1.4466	0.3015	1.60E-06	0.24
20	rs187300129	ADA	T	0.0141	-1.4129	0.3147	7.14E-06	0.24
20	rs73909565	ADA	A	0.9836	1.492	0.3031	8.56E-07	4.45
21	rs9979730	Intergenic	A	0.9816	1.8305	0.3552	2.55E-07	6.24
21	rs183747783	LOC105372787	T	0.9758	1.1606	0.253	4.50E-06	3.19
21	rs192507448	Intergenic	A	0.017	-1.5219	0.303	5.09E-07	0.22

Table 12: Genome-Wide Suggestive Variants for European Cases and Normal Controls

CHR	SNP	Gene	EA	Freq EA	Beta	StdErr	P-value	OR
1	rs112046192	ZFYVE9	T	0.119	-0.1839	0.0414	8.93E-06	0.83
1	rs6699744	LOC105378797	A	0.3625	-0.1229	0.0261	2.39E-06	0.88
1	rs11165720	TGFBR3	A	0.0916	0.2002	0.0428	2.96E-06	1.22
2	rs485094	ABCB11	A	0.3495	-0.1146	0.0259	9.63E-06	0.89
3	rs14256251	Intergenic	C	0.009	-0.8968	0.1976	5.65E-06	0.41
5	rs725151	Intergenic	T	0.2811	-0.1269	0.0278	4.84E-06	0.88
5	rs9687846	C5orf67	A	0.1939	0.1414	0.0314	6.62E-06	1.15
6	rs9358901	Intergenic	T	0.669	0.13	0.0261	6.12E-07	1.14
7	rs9358901	PHF14	A	0.0877	-0.2066	0.0436	2.17E-06	0.81
7	rs41733	TBXAS1	T	0.9849	0.5097	0.1053	1.29E-06	1.66
8	rs78824412	Intergenic	T	0.0645	-0.2441	0.0536	5.23E-06	0.78
8	rs7825304	TUSC3	A	0.5453	-0.1128	0.0247	4.87E-06	0.89
8	rs287	LPL	A	0.7455	0.2108	0.0286	1.64E-13	1.23
8	rs189491454	Intergenic	T	0.9958	1.3229	0.2771	1.81E-06	3.75
8	rs2068449	LINC00534	T	0.9346	-0.2573	0.0505	3.54E-07	0.77
9	rs571166782	RC3H2	T	0.0103	0.7932	0.1673	2.13E-06	2.21
9	rs1339254	PCSK5	A	0.696	0.1228	0.0268	4.47E-06	1.13
10	rs11193189	SORCS1	T	0.0344	0.3239	0.0733	9.82E-06	1.38
10	rs11597169	ADARB2	A	0.4917	-0.1168	0.0252	3.59E-06	0.89
11	rs964184	ZPR1	C	0.8616	-0.2477	0.0359	5.18E-12	0.78
11	rs7925256	PAFAH1B2	T	0.0897	0.1953	0.0439	8.77E-06	1.22
12	rs71455663	DENND5B	A	0.8313	0.1598	0.0346	3.91E-06	1.17
13	rs76345073	Intergenic	T	0.0295	-0.4113	0.0804	3.09E-07	0.66
14	rs10135742	Intergenic	C	0.018	0.4732	0.0983	1.48E-06	1.61
14	rs8004913	GALNT16	T	0.8538	-0.1734	0.0353	9.01E-07	0.84
14	rs148743949	Intergenic	T	0.0111	0.6488	0.1414	4.49E-06	1.91
15	rs724541	PCSK6	C	0.3747	0.1132	0.0255	8.95E-06	1.12
15	rs11073147	Intergenic	A	0.4558	-0.1092	0.0245	8.37E-06	0.9
15	rs58293302	LOC105370873	A	0.203	-0.1736	0.0334	1.98E-07	0.84
16	rs247616	Intergenic	T	0.3339	-0.1639	0.0265	5.88E-10	0.85
16	rs11076176	CETP	T	0.8355	-0.1914	0.0346	3.05E-08	0.83
16	rs2019697	JPH3	A	0.1151	-0.1725	0.0388	8.52E-06	0.84
20	rs116974458	DOK5	A	0.0492	0.3054	0.0645	2.16E-06	1.36
20	rs34048310	CDH4	T	0.08	-0.2108	0.0472	7.80E-06	0.81
21	rs77448271	Intergenic	C	0.9623	0.3066	0.0688	8.27E-06	1.36

Table 13: Genome-Wide Suggestive Variants for African Cases and Normal Controls

CHR	SNP	Gene	EA	Freq EA	Beta	StdErr	P-value	OR
1	rs115921440	Intergenic	A	0.0181	-1.08	0.24	6.13E-06	0.34
1	rs75683367	Intergenic	C	0.9819	1.07	0.24	7.49E-06	2.92
1	rs114888382	Intergenic	A	0.982	1.07	0.24	7.68E-06	2.92
1	rs116202393	Intergenic	A	0.0174	-1.05	0.23	6.16E-06	0.35
1	rs10909951	MEGF6	T	0.2089	0.36	0.07	8.00E-08	1.43
1	rs77105791	Intergenic	C	0.1012	-0.35	0.07	1.36E-06	0.7
2	rs112984974	DTNB	T	0.0876	0.37	0.08	9.90E-06	1.44
2	rs111522539	TTC27	T	0.9659	0.58	0.13	8.51E-06	1.78
3	rs9849539	FHIT	A	0.5622	-0.24	0.05	5.17E-06	0.78
3	rs6771732	FHIT	T	0.7183	-0.24	0.05	1.12E-06	0.78
3	rs17063440	FHIT	T	0.8938	0.33	0.07	7.08E-06	1.39
4	rs78117133	Intergenic	T	0.9672	-0.58	0.13	9.23E-06	0.56
5	rs337093	Intergenic	A	0.1414	0.32	0.07	1.40E-06	1.37
5	rs4151699	PCDHGA1	C	0.0216	-0.76	0.17	8.66E-06	0.47
5	rs294966	LINC01933	T	0.7085	-0.22	0.05	5.03E-06	0.8
5	rs80243942	MAST4	A	0.0934	0.35	0.08	6.96E-06	1.42
7	rs117729532	Intergenic	A	0.9709	1.19	0.21	1.05E-08	3.28
8	rs10096633	Intergenic	T	0.4222	-0.21	0.04	1.72E-06	0.81
9	rs62577134	Intergenic	C	0.614	0.2	0.05	9.24E-06	1.23
9	rs10972570	DOCK8	C	0.572	-0.22	0.05	1.55E-06	0.81
9	rs1273673	CNTNAP3	A	0.9278	0.48	0.11	9.20E-06	1.62
9	rs2774154	CNTNAP3	T	0.0712	-0.48	0.11	8.24E-06	0.62
9	rs9407144	CNTNAP3	A	0.9297	0.49	0.11	8.53E-06	1.62
9	rs2480985	CNTNAP3	A	0.9302	0.49	0.11	8.44E-06	1.63
10	rs12254439	Intergenic	A	0.899	-0.35	0.08	3.20E-06	0.7
11	rs647837	JAM3	A	0.8251	0.28	0.06	1.90E-06	1.32
11	rs140764785	Intergenic	A	0.0258	-0.71	0.16	6.98E-06	0.49
11	rs140764785	Intergenic	A	0.0256	-0.72	0.16	5.32E-06	0.49
11	11:39461416	Intergenic	A	0.9744	0.72	0.16	5.63E-06	2.05
11	rs185285210	Intergenic	T	0.9752	0.71	0.16	7.07E-06	2.04
11	rs114768081	NARS2	A	0.9423	0.47	0.1	1.58E-06	1.59
11	rs111716430	Intergenic	A	0.0588	-0.47	0.1	8.57E-07	0.62
11	rs11237554	NARS2	T	0.9425	0.47	0.1	1.30E-06	1.6
11	rs116558162	NARS2	A	0.0527	-0.49	0.1	1.27E-06	0.61
11	rs193019036	Intergenic	A	0.9575	0.5	0.11	9.40E-06	1.65
11	rs11237567	Intergenic	C	0.9581	0.5	0.11	8.13E-06	1.66
11	rs11237568	Intergenic	A	0.0421	-0.5	0.11	9.29E-06	0.61
12	rs149395900	Intergenic	A	0.0183	-1.02	0.23	6.14E-06	0.36
13	rs9577723	Intergenic	A	0.6404	0.21	0.05	3.29E-06	1.24
15	rs113365583	SMAD3	T	0.0221	-0.76	0.16	2.28E-06	0.47
15	rs111902897	SMAD3	T	0.9781	0.8	0.16	8.30E-07	2.22

Table 13: Genome-Wide Suggestive Variants for African Cases and Normal Controls (cont'd, 2 of 2)

CHR	SNP	Gene	EA	Freq EA	Beta	StdErr	P-value	OR
15	rs112275953	SMAD3	C	0.0216	-0.79	0.18	9.59E-06	0.45
15	rs4965328	Intergenic	A	0.7111	-0.24	0.05	2.00E-06	0.79
16	rs2908872	Intergenic	C	0.0717	-0.39	0.09	6.97E-06	0.68
16	rs3095659	Intergenic	T	0.9283	0.4	0.09	3.67E-06	1.5
16	rs2908887	Intergenic	T	0.0717	-0.4	0.09	3.84E-06	0.67
16	rs2908885	Intergenic	C	0.0717	-0.4	0.09	3.97E-06	0.67
16	rs147862874	Intergenic	A	0.0718	-0.39	0.09	5.53E-06	0.67
16	rs147862874	Intergenic	C	0.072	-0.4	0.09	5.05E-06	0.67
16	rs58096694	Intergenic	T	0.9278	0.38	0.09	9.59E-06	1.46
16	rs78373567	Intergenic	A	0.9288	0.39	0.09	5.59E-06	1.48
16	rs116642730	Intergenic	T	0.0712	-0.39	0.09	5.95E-06	0.68
16	rs60856989	Intergenic	C	0.9288	0.39	0.09	6.11E-06	1.48
16	rs16946828	Intergenic	T	0.0713	-0.4	0.09	3.34E-06	0.67
16	rs113904604	Intergenic	A	0.0714	-0.4	0.09	5.32E-06	0.67
16	rs16946830	Intergenic	A	0.0712	-0.39	0.09	6.19E-06	0.67
16	rs57154787	LOC105369246	T	0.8835	0.34	0.07	4.25E-06	1.41
16	rs58421241	LOC105369246	A	0.1156	-0.34	0.08	4.83E-06	0.71
17	rs530570910	CSNK1D	A	0.0424	-0.63	0.14	7.40E-06	0.53
18	rs142959114	Intergenic	A	0.0149	-1.23	0.27	5.48E-06	0.29
18	rs139568633	WDR7	T	0.025	-0.67	0.15	9.31E-06	0.51
18	rs116833222	WDR7	A	0.9735	0.68	0.15	3.62E-06	1.97
18	rs7237665	WDR7	A	0.0273	-0.68	0.14	2.48E-06	0.51
18	rs115684400	WDR7	T	0.9731	0.67	0.15	4.54E-06	1.95
18	rs306218	Intergenic	T	0.2434	-0.24	0.05	9.97E-06	0.79
19	rs8100892	ZNF69	T	0.5858	-0.21	0.05	4.55E-06	0.81
20	rs6027272	C20orf197	T	0.6164	-0.23	0.05	1.90E-06	0.8
22	rs186190745	ATP6V1E1	A	0.0359	-0.57	0.12	4.53E-06	0.57

Chapter IV

Discussion

European Ancestry

Both super and normal controls captured four identical genome-wide significant signals. The first signal (rs287) is located on the *LPL* gene, also known as lipoprotein lipase. In the STAMPEED Consortium, six different SNPs located on the *LPL* gene were associated with HDL, waist circumference, blood pressure, MetS, triglyceride and glucose levels [42]. The *LPL* gene encodes a protein called lipoprotein lipase that is found in adipose tissues. The primary function of lipoprotein lipase is to break down triglycerides for the body to use as a source of energy. Mutations that change the expression levels of this gene can halt this process, causing a buildup of triglycerides in the blood stream and in some cases, hyperlipidemia. The rs287 variant found in our GWAS analysis is located on the intron portion of the gene. Even though introns are spliced out during translation, intronic mutations can still affect gene expression when the sequence falls on a splicing site or enhancer region. Therefore, an impaired expression of *LPL* caused by rs283 can potentially lead to one or multiple components of MetS.

The second genome-wide significant SNP (rs964184) is located on the three prime un-translated region of the *ZPR1* gene. This specific variant has been cited for its association with plasma lipid levels along with CVD. In Kati Kristiansson et al.'s GWAS study in four Finnish cohorts, rs964184 is found to be associated with MetS status in all four studies [43]. *ZPR1* is a regulatory protein for signal transduction and cell proliferation. The *ZPR1* gene is in proximity with the apolipoprotein *APOA5* locus, which is known for playing an important role in regulation of plasma triglyceride and HDL. Alleles in variants of *ZPR1* "may alter the metabolism of triglycerides, HDL cholesterol or glucose through the interaction with *APOA5*" [44].

The third significant SNP (rs11076176) is an intron variant located on the *CETP* gene. *CETP*, or cholesteryl ester transfer protein, has a main function of transferring cholesteryl esters and triglycerides between the lipoproteins. In a multi-ethnic analysis of lipid-associated loci, Musunuru et al. tested around 50,000 polymorphisms in 25,000 European Americans and found two *CETP* polymorphisms that are associated with HDL [45]. Although the exact mechanism is unknown, a defect in this gene is known to cause Hyperalphalipoproteinemia, a condition characterized by increased HDL levels.

The last significant variant (rs247616) is located in the intergenic region of chromosome 16. Though not directly located on a gene, this variant is upstream of the *CETP* gene and have been associated with HDL cholesterol levels. The mechanism of this upstream regulation has yet to be established, but Suhy et al. proposed a model based on their transcription factor binding predictions [46]. In this model, rs247616 makes up the enhancer region of the *CETP* gene, and highly conserved transcription factors *YBX1* and *CEBPA* are unable to bind to the minor allele of this variant, thereby causing a change in *CETP* expression.

African Ancestry

The first genome-wide significant signal was shared by both normal controls and super controls. This SNP (rs117729532) is a 2KB upstream variant of an uncharacterized gene known as *LOC107986717*. More information regarding this uncharacterized gene is needed to establish any relationship between the upstream variant and MetS disease endpoints. The second variant was only present in the cases versus super control subset. This SNP (rs115553887) is an intronic variant located on the *RBM20* gene. The *RBM20* gene encodes a protein that binds RNA and regulates splicing. Mutations to this gene can cause Familial Dilated Cardiomyopathy, a genetic

form of heart disease where heart muscles of one chamber become thin and weakened, causing the open area of the chamber to become enlarged. *RBM20* is responsible for creating a protein that splices *TTN*, which is a gene that provides instruction for making titin. Titin provides flexibility and stability of muscles cells, including cardiomyocytes. Though not directly related to MetS, this variant relates to similar key cardiovascular endpoints as Metabolic Syndrome. More fine mapping of this variant must be done to observe the true associations between this variant and MetS.

Results Comparison

For European ancestry samples, both super and normal control's QQ plots (Figure 5) show a slight inflation. Distributions of the GWAS p-values should follow a uniform distribution until the end where truly significant SNPs inflate p-values. In our case, a small inflation can be a sign of the residue effects of population stratification in our samples. For variant discovery, super controls and normal controls both tagged four genome-wide significant SNPs. For suggestive SNPs, super controls were able to find a few more variants compared to the normal controls. This might suggest a modest advantage of using the super control criteria.

Compared to the QQ plots for subjects of European ancestry, the African ancestry plots in Figure 5 show lower genomic control variables, with the cases and normal controls having a deflated value below the line of identity. The reduction in power in these studies can be attributed to the smaller sample sizes. Cases and super controls resulted in two genome-wide significant SNPs as opposed to the one SNP found in cases and normal control. As for suggestive SNPs, cases and super controls once again had more results than cases and normal controls, with 150 and sixty-nine variants respectively. The strict criteria of super controls might decrease

sample size, but overall, it does make a stronger comparison between the case and control subsets. This might potentially be the reason why cases and super controls have more success in finding significant signals.

Conclusion

In summary, the dbGaP biorepository is an invaluable resource that contains raw phenotype and genotypes information suitable for all types of genetic studies. Unfortunately, the resource is underutilized due to issues with how the data is made available to researchers. With proper quality control of genotype files and processing of phenotype files, it is possible to leverage these data for high quality research and analyses. In order to explore the pathogenesis and genetic variants of MetS, three dbGaP studies of MESA, ARIC and CHS were used to perform a GWAS. Two sets of criteria were used for control selection in order to explore the best threshold for this process. For the results, super controls performed better in terms of variant finding power than the widely-used normal controls. Four genome-wide significant SNPs (rs287, rs964184, rs11076176, rs247616) were detected in the European subset after the meta-analysis. Two genome-wide significant SNPs, (rs117729532, rs115553887), were found in the African subset.

Future Directions

There are several future directions and objectives that can be used to extend this project. First, it is important to enlarge our sample size as much as possible to increase variant detection power. Besides adding on more studies from dbGaP, subjects from the biorepository UK Biobank will also be harmonized and incorporated into future GWAS runs. In addition to adding

samples, it would be interesting to rerun tests on male and female specific subsets in order to find variants specifically significant for a gender. The incorporation of super control versus normal control showed that super controls were superior in detecting significant variants. Another idea for future tests is to run super cases, which are subjects with all five components of MetS.

Next, a variety of post-GWAS tools can be run to increase the interpretability of our results. After assigning a nearest gene to significant variants, a gene set enrichment and pathway analysis can be run to see if a differentially expressed set of genes are associated with a certain biological pathway or molecular function. Web Gestalt is a web tool that finds associations with disease phenotypes given a list of genes. Depict is another web tool that highlights enriched pathways and predicts most likely causal genes. Finally, a network analysis can be run to observe how key components of these related biological pathways interact.

Another future objective is to make dbGaP datasets and genotypic data more accessible to fellow researchers. One of the objectives in Salem Lab is to release the quality control pipeline as a web tool to resolve the convoluted issue of QCing datasets. It would also be informative to release a paper regarding the navigation of dbGaP datasets, website and data tables to provide tips and tools for harmonizing these studies.

References

1. Boudreau, D. M., Malone, D. C., Raebel, M. A., Fishman, P. A., Nichols, G. A., Feldstein, A. C., Boscoe, A. N., Ben-Joseph, R. H., Magid, D. J., & Okamoto, L. J. (2009). Health care utilization and costs by metabolic syndrome risk factors. *Metabolic Syndrome and Related Disorders*, *7*(4), 305–314. doi: 10.1089/met.2008.0070

2. International Diabetes Federation (2006). Consensus statements. Retrieved from https://www.idf.org/e-library/consensus-statements/60-idfconsensus-worldwide definitionof-the-metabolic-syndrome.html

3. Alberti, K. G., Eckel, R. H., Grundy, S. M., Zimmet, P. Z., Cleeman, J. I., Donato, K. A., Fruchart, J. C., James, W. P., Loria, C. M., Smith, S. C., Jr., International Diabetes Federation Task Force on Epidemiology and Prevention, National Heart, Lung, and Blood Institute, American Heart Association, World Heart Federation, International Atherosclerosis Society, & International Association for the Study of Obesity. (2009). Harmonizing the metabolic syndrome. *Circulation*, *120*(16),1640–1645. doi: 10.1161/CIRCULATIONAHA.109.192644

4. Reaven, G. M. (1988). Banting lecture 1988. Role of insulin resistance in human disease. *Diabetes*, *37*(12), 1595–1607. doi: 10.2337/diabetes.37.12.1595

5. Monda, K. L., North, K. E., Hunt, S. C., Rao, D. C., Province, M. A., Arnett, D. K., & Kraja, A. T. (2010). The genetics of obesity and the metabolic syndrome. *Endocrine, Metabolic and Immune Disorders - Drug Targets*, *10*(2), 86–108. doi: 10.2174/187153010791213100

6. Huang, P. L. (2009). A comprehensive definition for metabolic syndrome. *Disease Models & Mechanisms*, *2*(5-6), 231–237. doi: 10.1242/dmm.001180

7. Talayero, B. G., & Sacks, F. M. (2011). The role of triglycerides in atherosclerosis. *Current Cardiology Reports*, *13*(6), 544–552. doi: 10.1007/s11886-011-0220-3

8. Pereira, S. S., & Alvarez-Leite, J. I. (2014). Low-grade inflammation, obesity, and diabetes. *Current Obesity Reports*, *3*(4), 422–431. doi: 10.1007/s13679-014-0124-9

9. Pérez-Martínez, P., Mikhailidis, D. P., Athyros, V. G., Bullo, M., Couture, P., Covas, M. I., de Koning, L., Delgado-Lista, J., Díaz-López, A., Drevon, C. A., Estruch, R., Esposito, K., Fitó, M., Garaulet, M., Giugliano, D., García-Ríos, A., Katsiki, N., Kolovou, G., Lamarche, B., Maiorino, M. I., Mena-Sánchez, G., Muñoz-Garach, A., Nikolic, D., Ordovás, J. M., Pérez-Jiménez, F., Rizzo, M., Salas-Salvadó, J., Schröder, H., Tinahones, F. J., de la Torre, R., van Ommen, B., Wopereis, S., Ros, E., & López-Miranda, J. (2017). Lifestyle recommendations for the prevention and management of metabolic syndrome: An international panel recommendation. *Nutrition Reviews*, *75*(5), 307–326. doi: 10.1093/nutrit/nux014

10. Kim, S. K., Hong, S., Chung, J., & Cho, K. B. (2017). Association between alcohol consumption and metabolic syndrome in a community-based cohort of Korean adults. *Medical Science Monitor*, *23*, 2104–2110. doi: 10.12659/msm.901309

11. Sun, K., Liu, J., & Ning, G. (2012). Active smoking and risk of metabolic syndrome: A meta-analysis of prospective studies. *PLoS ONE*, *7*(10). doi: 10.1371/journal.pone.0047791

12. Slagter, S. N., van Vliet-Ostaptchouk, J. V., Vonk, J. M., Boezen, H. M., Dullaart, R. P., Kobold, A. C., Feskens, E. J., van Beek, A. P., van der Klauw, M. M., & Wolffenbuttel, B. H. (2014). Combined effects of smoking and alcohol on metabolic syndrome: The LifeLines Cohort study. *PLoS ONE*, *9*(4). doi: 10.1371/journal.pone.0096406

13. Beigh, S. H., & Jain, S. (2012). Prevalence of metabolic syndrome and gender differences. *Bioinformation*, *8*(13), 613–616. doi: 10.6026/97320630008613

14. Helmrich, S. P., Ragland, D. R., Leung, R. W., & Paffenbarger, R. S. (1991). Physical activity and reduced occurrence of non-insulin-dependent diabetes mellitus. *The New England Journal of Medicine*, *325*(3), 147–152. doi: 10.1056/nejm199107183250302

15. Diabetes Prevention Program Research Group. (2002). Reduction in the incidence of type 2 diabetes with lifestyle intervention or Metformin. The New England Journal of Medicine, 346(6), 393–403. doi: 10.1056/NEJMoa012512

16. Kraus, W. E., Houmard, J. A., Duscha, B. D., Knetzger, K. J., Wharton, M. B., McCartney, J. S., Bales, C. W., Henes, S., Samsa, G. P., Otvos, J. D., Kulkarni, K. R., & Slentz, C. A. (2002). Effects of the amount and intensity of exercise on plasma lipoproteins. *The New England Journal of Medicine*, *347*, 1483–1492. doi: 10.1056/NEJMoa020194

17. Majka, D. S., Chang, R. W., Vu, T. T., Palmas, W., Geffken, D. F., Ouyang, P., Ni, H., & Liu, K. (2008). Physical activity and high-sensitivity C-reactive protein. *American Journal of Preventive Medicine*, *36*(1), 56–62. doi: 10.1016/j.amepre.2008.09.031

18. Ali, O. (2013). Genetics of type 2 diabetes. *World Journal of Diabetes*, *4*(4), 114–123. doi: 10.4239/wjd.v4.i4.114

19. Stančáková, A., Pihlajamäki, J., Kuusisto, J., Stefan, N., Fritsche, A., Häring, H., Andreozzi, F., Succurro, E., Sesti, G., Boesgaard, T. W., Hansen, T., Pedersen, O., Jansson, P. A., Hammarstedt, A., Smith, U., Laakso, M., & EUGENE2 Consortium (2008). Single nucleotide polymorphism rs7754840 of CDKAL1Is associated with impaired insulin secretion in nondiabetic offspring of type 2 diabetic subjects and in a large sample of men with normal glucose tolerance. *The Journal of Clinical Endocrinology & Metabolism*, *93*(5), 1924–1930. doi: 10.1210/jc.2007-2218

20. Doris, P. A. (2010). The genetics of blood pressure and hypertension: The role of rare variation. *Cardiovascular Therapeutics*, *29*(1), 37–45. doi: 10.1111/j.1755-5922.2010.00246.x

21. Cusi, D., Barlassina, C., Azzani, T., Casari, G., Citterio, L., Devoto, M., Glorioso, N., Lanzani, C., Manunta, P., Righetti, M., Rivera, R., Stella, P., Troffa, C., Zagato, L., & Bianchi, G. (1997). Polymorphisms of α-adducin and salt sensitivity in patients with essential hypertension. *The Lancet*, *349*(9062), 1353–1357. doi: 10.1016/s0140 6736(97)01029-5

22. Bonnardeaux, A., Davies, E., Jeunemaitre, X., Féry, I., Charru, A., Clauser, E., Tiret, L., Cambien, F., Corvol, P., & Soubrier, F. (1994). Angiotensin II type 1 receptor gene polymorphisms in human essential hypertension. *Hypertension*, *24*(1), 63–69. doi: 10.1161/01.hyp.24.1.63

23. Johnson, J. A., Zineh, I., Puckett, B. J., McGorray, S. P., Yarandi, H. N., & Pauly, D. F. (2003). Beta 1-adrenergic receptor polymorphisms and antihypertensive response to metoprolol. *Clinical Pharmacology & Therapeutics*, *74*(1), 44–52. doi: 10.1016/s0009-9236(03)00068-7

24. Kullo, I. J., de Andrade, M., Boerwinkle, E., Mcconnell, J. P., Kardia, S. L. R., & Turner, S. T. (2005). Pleiotropic genetic effects contribute to the correlation between HDL cholesterol, triglycerides, and LDL particle size in hypertensive sibships. *American Journal of Hypertension*, *18*(1), 99–103. doi: 10.1016/j.amjhyper.2004.09.002

25. Dron, J. S., & Hegele, R. A. (2017). Genetics of triglycerides and the risk of atherosclerosis. *Current Atherosclerosis Reports*, *19*(7). doi: 10.1007/s11883-017-0667-9

26. 23andMe. (2008, December 8). New papers from Nature Genetics yield a bounty of cholesterol SNPs. Retrieved from https://blog.23andme.com/news/new-papers-from-nature-genetics-yield-a-bounty-of-cholesterol-snps/

27. Weissglas-Volkov, D., & Pajukanta, P. (2010). Genetic causes of high and low serum HDL-cholesterol. *Journal of Lipid Research*, *51*(8), 2032–2057. doi: 10.1194/jlr.r004739

28. Willer, C. J., Sanna, S., Jackson, A. U., Scuteri, A., Bonnycastle, L. L., Clarke, R., Heath, S. C., Timpson, N. J., Najjar, S. S., Stringham, H. M., Strait, J., Duren, W. L., Maschio, A., Busonero, F., Mulas, A., Albai, G., Swift, A. J., Morken, M. A., Narisu, N., Bennett, D., Parish, S., Shen, H., Galan, P., Meneton, P., Hercberg, S., Zelenika, D., Chen, W., Li, Y., Scott, L. J., Scheet, P. A., Sundvall, J., Watanabe, R. M., Nagaraja, R., Ebrahim, S., Lawlor, D. A., Ben-Shlomo, Y., Davey-Smith, G., Shuldiner, A. R., Collins, R., Bergman, R. N., Uda, M., Tuomilehto, J., Cao, A., Collins, F. S., Lakatta, E., Lathrop, G. M., Boehnke, M., Schlessinger, D., Mohlke, K. L., & Abecasis, G. R. (2008). Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nature Genetics*, *40*(2), 161–169. doi: 10.1038/ng.76

29. Speakman, J. R., Rance, K. A., & Johnstone, A. M. (2008). Polymorphisms of the FTO gene are associated with variation in energy intake, but not energy expenditure. *Obesity*, *16*(8), 1961–1965. doi: 10.1038/oby.2008.318

30. Rosmond, R., & Holm, G. (2008). A 5-year follow-up study of 3 polymorphisms in the human glucocorticoid receptor gene in relation to obesity, hypertension, and diabetes. *Journal of the CardioMetabolic Syndrome*, *3*(3), 132–135. doi: 10.1111/j.1559-4572.2008.00008.x

31. Kondo, H., Shimomura, I., Matsukawa, Y., Kumada, M., Takahashi, M., Matsuda, M., Ouchi, N., Kihara, S., Kawamoto, T., Sumitsuji, S., Funahashi, T., & Matsuzawa, Y. (2002). Association of adiponectin mutation with type 2 diabetes: A candidate gene for the insulin resistance syndrome. *Diabetes*, *51*(7), 2325–2328. doi: 10.2337/diabetes.51.7.2325

32. dbGaP/database of Genotypes and Phenotypes/ National Center for Biotechnology Information, National Library of Medicine (NCBI/NLM)/https://www.ncbi.nlm.nih.gov/gap

33. Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904–909. doi: 10.1038/ng1847

34. Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, *26*(22), 2867–2873. doi: 10.1093/bioinformatics/btq559

35. Thorisson, G. A., Smith, A. V., Krishnan, L., & Stein, L. D. (2005). The International HapMap Project web site. *Genome Research*, *15*(11), 1592–1593. doi: 10.1101/gr.4413105

36. Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M., Schlessinger, D., Stambolian, D., Loh, P. R., Iacono, W. G., Swaroop, A., Scott, L. J., Cucca, F., Kronenberg, F., Boehnke, M., Abecasis, G. R., & Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nature Genetics*, *48*(10), 1284–1287. doi: 10.1038/ng.3656

37. The data/analyses presented in the current project are based on the use of study data downloaded from the dbGaP web site, under phs000209.v13.p3/https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000209.v13.p3

38. The data/analyses presented in the current project are based on the use of study data downloaded from the dbGaP web site, under phs000280.v5.p1/https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000280.v5.p1

39. The data/analyses presented in the current project are based on the use of study data downloaded from the dbGaP web site, under phs000287.v6.p1/https://www.ncbi.nlm.nih.gov/projects/gap/cgi bin/study.cgi?study_id=phs000287.v6.p1

40. Marchini, J., Howie, B., Myers, S., Mcvean, G., & Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, *39*(7), 906–913. doi: 10.1038/ng2088

41. Willer, C. J., Li, Y., & Abecasis, G. R. (2010). METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, *26*(17), 2190–2191. doi: 10.1093/bioinformatics/btq340

42. Kraja, A. T., Vaidya, D., Pankow, J. S., Goodarzi, M. O., Assimes, T. L., Kullo, I. J.,Sovio, U., Mathias, R. A., Sun, Y. V., Franceschini, N., Absher, D., Li, G., Zhang, Q., Feitosa, M. F., Glazer, N. L., Haritunians, T., Hartikainen, A. L., Knowles, J. W., North, K. E., Iribarren, C., Kral, B., Yanek, L., O'Reilly, P. F., McCarthy, M. I., Jaquish, C., Couper, D. J., Chakravarti, A., Psaty, B. M., Becker, L. C., Province, M. A., Boerwinkle, E., Quertermous, T., Palotie, L., Jarvelin, M. R., Becker, D. M., Kardia, S. L., Rotter, J. I., Chen, Y. D., & Borecki, I. B. (2011). A bivariate genome-wide approach to metabolic syndrome: STAMPEED consortium. *Diabetes*, *60*(4), 1329–1339. doi: 10.2337/db10 1011

43. Kristiansson, K., Perola, M., Tikkanen, E., Kettunen, J., Surakka, I., Havulinna, A. S., Stancáková, A., Barnes, C., Widen, E., Kajantie, E., Eriksson, J. G., Viikari, J., Kähönen, M., Lehtimäki, T., Raitakari, O. T., Hartikainen, A. L., Ruokonen, A., Pouta, A., Jula, A., Kangas, A. J., Soininen, P., Ala-Korpela, M., Männistö, S., Jousilahti, P., Bonnycastle, L. L., Järvelin, M. R., Kuusisto, J., Collins, F. S., Laakso, M., Hurles, M. E., Palotie, A., Peltonen, L., Ripatti, S., & Salomaa, V. (2012). Genome-wide screen for metabolic syndrome susceptibility loci reveals strong lipid gene contribution but no evidence for common genetic basis for clustering of metabolic syndrome traits. *Circulation: Cardiovascular Genetics*, *5*(2), 242–249. doi: 10.1161/circgenetics.111.961482

44. Ueyama, C., Horibe, H., Yamase, Y., Fujimaki, T., Oguri, M., Kato, K., Arai, M., Watanabe, S., Murohara, T., & Yamada, Y. (2015). Association of FURIN and ZPR1 polymorphisms with metabolic syndrome. *Biomedical Reports*, *3*(5), 641–647. doi: 10.3892/br.2015.484

45. Musunuru, K., Romaine, S. P. R., Lettre, G., Wilson, J. G., Volcik, K. A., Tsai, M. Y., Taylor, H. A., Jr., Schreiner, P. J., Rotter, J. I., Rich, S. S., Redline, S., Psaty, B. M., Papanicolaou, G. J., Ordovas, J. M., Liu, K., Krauss, R. M., Glazer, N. L., Gabriel, S. B., Fornage, M., Cupples, L. A., Buxbaum, S. G., Boerwinkle, E., Ballantyne, C. M.,

Kathiresan, S., & Rader, D. J. (2012). Multi-ethnic analysis of lipid-associated loci: The NHLBI CARe Project. *PLoS ONE*, *7*(5). doi: 10.1371/journal.pone.0036473

46. Suhy, A., Hartmann, K., Papp, A. C., Wang, D., & Sadee, W. (2015). Regulation of CETP expression by upstream polymorphisms. *Pharmacogenetics and Genomics*, *25*(8), 394–401. doi: 10.1097/fpc.0000000000000151

47. Lin, E., Kuo, P., Liu, Y., Yang, A. C., & Tsai, S. (2017). Detection of susceptibility loci on APOA5 and COLEC12 associated with metabolic syndrome using a genome-wide association study in a Taiwanese population. *Oncotarget*, *8*(55). doi: 10.18632/oncotarget.20967

48. Zabaneh, D., & Balding, D. J. (2010). A genome-wide association study of the metabolic syndrome in Indian Asian men. *PLoS ONE*, *5*(8). doi: 10.1371/journal.pone.0011961

49. Zhu, Y., Zhang, D., Zhou, D., Li, Z., Li, Z., Fang, L., Yang, M., Shan, Z., Li, H., Chen, J., Zhou, X., Ye, W., Yu, S. Li, H., Cai, L. Liu, C., Zhang, J., Wang, L., Lai, Y., Ruan, L., Sun, Z., Zhang, S., Wang, H., Liu, Y., Xu, Y., Ling, J., Xu, C., Zhang, Y., Lv, D., Yuan, Z., Zhang, J., Zhang, Y., Shi, Y., & Lai, M. (2017). Susceptibility loci for metabolic syndrome and metabolic components identified in Han Chinese: A multi- stage genome- wide association study. *Journal of Cellular and Molecular Medicine*, *21*(6), 1106–1116. doi: 10.1111/jcmm.13042