GENETIC ASSOCIATION STUDIES: APPLICATION IN THE INVESTIGATION OF BIOMARKERS RELATED TO CARDIOVASCULAR DISEASES AND STUDY DESIGN

Jin Li

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

Chapel Hill
2015

Approved by:

Ethan M. Lange

Leslie A. Lange

Yun Li

Karen L. Mohlke

Monte S. Willis

© 2015
Jin Li
ALL RIGHTS RESERVED

ABSTRACT

JIN LI: Genetic association studies: application in the investigation of biomarkers related to cardiovascular diseases and study design
(Under the direction of Ethan M. Lange)

Cardiovascular disease (CVD) is the No. 1 cause of death in the United States, killing about 610,000 people every year. Biomarkers are important tools to identify vulnerable individuals at high risk of CVD. Investigation of the genetic architecture for biomarkers and other risk factors related to CVD is of critical importance in the prevention and treatment of CVD.

For my first chapter, I conducted genome-wide admixture and association studies for iron-related traits in 2347 African Americans (AAs) participants from the Jackson Heart Study (JHS). I identified, for the first time, a second independent genome-wide significant signal in the *TF* region associated with total iron binding capacity levels. I also identified a novel functional missense variant in the *G6PD-GAB3* region significantly associated with ferritin levels. Both results were replicated in a second AA cohort with iron measures.

For my second chapter, I conducted genome-wide admixture and association studies, and gene-based exome-wide association studies of rare variants, to identify variants or genes, harboring a high burden of rare functional variants, associated with lipoprotein(a) [Lp(a)] cholesterol levels in 2895 AAs participating in the JHS. I observed significant evidence for association between Lp(a) and both local ancestry and hundreds variants spanning ~10Mb the

LPA gene region on chromosome 6q. Of note, the region containing associated variants became much narrower, centered over the *LPA* gene (<1Mb), after adjusting for local ancestry. I also observed a single significant non-synonymous SNP in *APOE* and a high burden of coding variants in *LPA* and *APOE* significantly associated with Lp(a) levels

For my third chapter, I investigated the genetic association of four candidate variants with blood pressure and tested the modifying effects of environmental factors in 7,319 Chinese adults from the China Nutrition and Health Survey (CHNS). I observed that rs1458038 exhibited a significant genotype-by-BMI interaction affecting blood pressure measures, with the strongest variant effects in those with the highest BMI.

Finally, for my last chapter, I described a multistage GWAS study design that uses selective phenotyping to increase power for studies with existing genome-wide genotypic data and to-be-measured quantitative phenotypes that are under a sample-size constraint. The approach uses simulated annealing to identify the optimal subset of subjects to be phenotyped in Stage 2 of a two-stage GWAS. I showed that our approach has greater statistical power than the conventional approach of randomly selecting a subset of subjects for phenotyping. We demonstrate the gains in power for both directly genotyped and imputed genetic variants.

Together, these studies further our understanding of the genetic architecture of risk factors for CVD, suggest some candidates for future genetic and molecular studies, and also shed some light on the study design of future large-scale genetic association studies where the cost constraints will be determined by the expense of measuring new biomarkers in studies that have existing genetic data.

ACKNOWLEDGEMENTS

I am grateful to all the people who gave me guidance, help, and support during my doctoral studies. I would like to thank my advisor Dr. Ethan Lange for being such a helpful, wise, and considerate mentor: giving me lots of trust and opportunities for me to play a major role in several studies, allowing me lots of flexibility so that I could arrange research, study, and life in a balanced way and achieve my master degree in statistics at the same time, and being extremely patient to me whatever questions I ask him to explain. In addition, Ethan is always optimistic no matter what frustrating situations we are in, and he never say a harsh word to me but always encourage me instead when anything went wrong, which greatly built up my confidence in myself as well as in conducting research.

I would also appreciate the help from the other members of my dissertation committee: thanks to Dr. Leslie Lange for giving me opportunities to participate in many large cohort studies and introducing me to other senior investigators in our field; thanks to Dr. Yun Li for her statistical assistance and suggestions; thanks to Dr. Karen Mohlke for her guidance when I was rotating in her lab and her suggestions when I am looking for jobs; thanks to Dr. Monte Willis for his contribution to my work with his profound expertise in the pathology of cardiovascular diseases. I will always be grateful for what they have contributed to my scientific career.

Next, I would like to acknowledge people in and outside of our lab during my time at UNC. I would like to thank Dr. Yunfei Wang, Yurong Lu, and Dr. Jaclyn Ellis, former members

of the Lange Labs, for getting me familiar with the lab work when I first joined the lab four years ago, and for helping to create a solid foundation for my research later on. I would like to thank my fellow friends in the Curriculum of Bioinformatics and Computational Biology, especially Qing Duan and Guosheng Zhang, for their valuable help with my course work and research projects. I would also like to thank other post doctorate colleagues, especially Dr. Zheng Xu for his tremendous help with my questions in statistics and Dr. Ying Wu for her valuable advice on several projects. In addition, I would acknowledge the contributions made by our colleagues in Jackson Heart Study, China Health and Nutrition Survey, CHARGE, etc.

Finally, I would like to thank my parents in China for encouraging me to chase my dreams in the U.S. They kept on reminding me to stay healthy and happy, and gave me valuable advice and comfort whenever I suffered through times of difficulty or uncertainty.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

1000G: 1000 Genomes Project

AA: African Americans

APOE: apolipoprotein E

ATP2B1: ATPase, Ca++ transporting, plasma membrane 1

BMI: body mass index

CEU: Utah residents with ancestry from northern and western Europe

CHB: Han Chinese from Beijing, China

CHNS: China Health and Nutrition Survey

CNV: copy number variation

CRP: C-reactive protein

CSK: c-src tyrosine kinase

CVD: cardiovascular disease

CYP17A1: cytochrome P450, family 17, subfamily A, polypeptide 1

DBP: diastolic blood pressure

DNA: deoxyribonucleic acid

EA: European Americans

EAF: effect allele frequency

FAMHES: Fangchenggang Area Male Health and Examination Survey

FGF5: fibroblast growth factor 5

G6PD: glucose-6-phosphate dehydrogenase

GAB3: GRB2-associated binding protein 3

GWAS: genome-wide association study

GWAI: genome-wide association gene-gene interaction

GWEI: genome-wide environmental interaction

HANDLS: Healthy Aging in Neighborhoods of Diversity across the Life Span study

HapMap: International Haplotype Map Project

HDL: high-density lipoprotein

HWE: Hardy-Weinberg equilibrium

JHS: Jackson Heart Study

JPT: Japanese from Tokyo, Japan

LLARRMA: LASSO local automatic regularization resample model averaging

LD: linkage disequilibrium

LDL: low-density lipoprotein

LOD score: logarithm of the odds score

LOF: loss of function

Lp(a): lipoprotein(a)

LPA: lipoprotein, Lp(a)

MAF: minor allele frequency

MCV: mean corpuscular volume

NGS: next-generation sequencing

PC: principal component

PolyPhen: Polymorphism Phenotyping

Q-Q plots: quantile-quantile plots

RBC: red blood cell

RDW: red cell distribution width

RMIP: a resample model inclusion probability

SA: simulated annealing

SAT: transferrin saturation

SBP: systolic blood pressure

SD: standard deviation

SE: standard error of the mean

SIFT: Sorting Intolerant from Tolerant

SKAT: Sequence kernel association tests

SKAT-O: SKAT optimal test

SNP: single nucleotide polymorphism

T2D: type 2 diabetes

TC: total cholesterol

TG: triglycerides

TF: transferrin

TIBC: total iron binding capacity

UNC-CH: University of North Carolina, Chapel Hill

UTR-3: the three prime untranslated region

WES: whole-exome sequencing

WGS: whole-genome sequencing

YRI: Yorubans from Ibadan, Nigeria

CHAPTER I: INTRODUCTION

Cardiovascular diseases

Cardiovascular disease (CVD) is the No. 1 cause of death in the United States, killing about **610,000 people** every year (1). It is the leading cause of death for both American men and women, and for people of most ethnicities in the United States, including African Americans, Hispanics, and whites. The 2011 overall rate of death attributable to CVD was 229.6 per 100 000 Americans, 275.7 for males and 192.3 for females, 271.9 for white males, 352.4 for black males, 188.1 for white females, and 248.6 for black females (2). CVD claims more lives than all forms of cancer combined. On the basis of 2011 death rate data, approximately 155 000 Americans who died of CVD in 2011 were <65 years of age, and 34% of deaths attributable to CVD occurred before the age of 75 years, which is younger than the current average life expectancy of 78.7 years. Worldwide, the Global Burden of Disease study estimated that in 2001, 12.45 million of >56 million total worldwide deaths were caused by CVD, and the number of deaths caused by CVD increased to 17 million in 2008 (3, 4). High blood pressure, high LDL cholesterol, and smoking are key risk factors for heart disease. Several other important risk factors include diabetes, obesity, poor diet, high stress, physical inactivity, and excessive alcohol use (5). Despite these environmental factors, genetic factors are also significant contributors to CVD. The familial clustering of CVD and its heritability are well established, with the heritability of CVD estimated at about 40% (6).

Biomarkers

Biomarkers are defined as measurable and quantifiable biological parameters that are objectively measured and evaluated as indicators for disease trait (risk factor or risk marker), disease state (preclinical or clinical), or disease rate (progression) (7). Biomarkers are important tools to identify vulnerable individuals at high risk of CVD, and to diagnose disease conditions promptly and accurately. Traditional risk factors (cigarette smoking, diabetes, hyperlipidaemia and hypertension) are observed in only a subset of individuals who develop CVD. Up to 20% of patients have no traditional risk factors, and 40% have only one (8). Thus, identifying novel risk markers for CVD has significant potential to improve the selection of individuals for preventative strategies. Furthermore, biomarkers are typically more proximal to gene products than disease outcomes, and thus, they can serve as a surrogate for end points of CVD and help better understand the genetic contribution to CVD (9). Classical CVD biomarkers included lipid profiles, inflammation factors, coagulation factors, cytokines, etc (7). In this dissertation, biomarkers that were specifically investigated included iron-related measures and lipoprotein(a).

Iron-related measures

Iron is critical to an array of metabolic functions, such as oxygen transport and oxidative phosphorylation. Normally, small daily losses of iron in the feces and through menstruation are balanced by its regulated intestinal absorption and its recovery from heme after phagocytosis of senescent red blood cells (10). Controversies have existed for some time regarding the association between iron status and CVD. A recent review suggested that in the reference range, iron status has a neutral effect, while extreme conditions of iron deficiency, as well as of iron overload, are associated with modestly increased CVD risk, although with different proposed mechanisms (11). A study showed that trivalent iron (FeIII) initiates a hydroxyl radical-catalyzed conversion of fibrinogen into a fibrin-like polymer (parafibrin) that is remarkably resistant to the proteolytic dissolution and thus promotes its intravascular deposition (12). Another possible mechanism underlying the association between iron overload and CVD is that of oxidative stress (13). Ferrous iron catalyzes a variety of free-radical oxidative reactions which generate reactive oxygen species (ROS). ROS may seriously damage cellular integrity and contribute directly to plaque disruption and thrombosis. A study on a Finnish cohort showed that serum ferritin was associated with most of the measured oxysterols, independently of major confounders, and also associated with an increased risk of myocardial infarction, independent of major cardiovascular risk factors (14, 15). On the other hand, iron deficiency can also lead to increased CVD risk. Since iron can inhibit thrombopoiesis, iron deficiency can lead to thrombocytosis due to lack of inhibition of thrombopoiesis (16). In addition, iron deficiency can cause anemia. Anemia increases work load of the heart due to increased heart rate and stroke volume, and this may cause ischemia, myocardial cell death and heart failure (17).

Laboratory tests typically used to assess iron transport and storage include serum iron, total iron binding capacity (TIBC), transferrin saturation (SAT), and serum ferritin. Genetic factors could affect iron metabolism through gastrointestinal absorption, transport, tissue uptake, storage, or remobilization from tissue stores, and thus could have a large impact on the variation of these iron-related measures (18, 19). Previous genetic analyses of iron metabolism have identified several variants associated with markers of iron status, including *HFE* variants with serum ferritin, transferrin, and SAT (20), *TF* variants with serum ferritin levels (20), *TMPRSS6* variants with hemoglobin, serum iron and SAT (21, 22), *PCSK7* variants with soluble transferrin receptor

(23), *TFR2* variants with hemochromatosis type 3 (24), and a locus on chromosome 2p14, tagged by SNP rs2698530, with iron deficiency (25).

Lipoprotein(a)

Lipoprotein (a) [Lp(a)] is a low-density lipoprotein (LDL)-like particle that consists of an apolipoprotein(a) covalently linked to apolipoprotein B100 by a disulfide bond (26). Lp(a) is an independent risk factor for CVD (27, 28). The association of elevated Lp(a) and coronary heart disease (CHD) has been reviewed recently in a 2009 meta-analysis (20). In the 24 cohort studies, the risk ratio for CHD was 1.13 (95% CI 1.09-1.18) after adjustment for lipids and other conventional risk factors. Lp(a) excess is observed in 18.6% patients with premature CHD (includes 12.7% with no other dyslipidemias) (29). Elevated Lp(a) predicts 15-year CVD outcomes and improves CVD risk prediction, and the hazard ratio for incident CVD was 1.37 per 1-SD higher Lp(a) level (30). Lp(a) is believed to both promote atherosclerosis and enhance thrombosis by multiple mechanisms. Like other lipoproteins, Lp(a) is bound by VLDL receptors in atherosclerotic macrophage and is detected in larger amounts in tissue from culprit lesions in patients with unstable coronary artery disease (31). More importantly, Lp(a) plays a role in enhancing thrombotic events through its anti-fibrolytic activity. The final vascular obstruction that occurs in atherosclerosis in myocardial infarction and stroke is due to sudden thrombosis at the sight of a narrowed artery. Since Lp(a) contains a region (Kringle IV) that mimics plasminogen, it has been proposed to prevent the inhibition of clot lysis (32), leading to an increased susceptibility to CVD. Other studies have found that Lp(a) may enhance atherosclerosis by the formation of circulating immune complexes with IgG antibodies specific for Chlamydia pneumonia (33), although the specific mechanisms of this have not been reported.

Genetic factors have a large impact on the variation of Lp(a) levels and approximately 70-90% of the total variance of Lp(a) can be attributed to variation within the *LPA* locus across worldwide populations(34, 35). Genetic variants in *LPA* gene are strongly associated with both an increased level of Lp(a) and an increased risk of coronary disease(36). In the *LPA* locus, a copy-number variation (CNV) which encodes a Kringle(IV) type 2 domain accounts for approximately half of variance explained by *LPA* locus(37, 38). Recent genome-wide association studies (GWAS) in subjects of European descent have identified multiple polymorphisms spanning 12.5 Mb on chromosome 6q26-27, which includes *LPA*, that are significantly associated with Lp(a) levels independent of each other and of the Kringle IV size polymorphism in *LPA* ($p<5x10^{-8}$) (37). A candidate gene study on multi-ethnic populations suggested both SNPs at 6q26-27 and the Kringle IV CNV were genomic determinants of Lp(a) level, and the proportion of total variance explained by each determinant differ across ethnic groups (39).

Hypertension

Hypertension is a conventional and important risk factor for CVD (40). Globally, complications of hypertension are responsible for 9.4 million deaths every year (41). Hypertension is responsible for at least 45% of deaths due to heart disease, and 51% of deaths due to stroke. In 2008, worldwide, approximately 40% of adults aged 25 and above had been diagnosed with hypertension; the number of people with the condition rose from 600 million in 1980 to 1 billion in 2008 (42). Overall, low- and middle- income countries have a higher prevalence and larger number of hypertension patients who are undiagnosed, untreated and have their blood pressure uncontrolled compared to high-income countries (42). Hypertension adversely impacts heart and blood vessels in various ways (43, 44). An increased pressure in blood vessels will cause the heart to work harder in order to pump blood, thus, hypertension can lead to a heart attack, an enlargement of the heart and heart failure (43). A consistent high blood pressure will result in weak spots on the blood vessels, making them more vulnerable to clog and burst, which leads to strokes when blood leaks out into the brain (45).

It is well established that hypertension is determined by both genetic factors and environmental factors, including tobacco use, salt and alcohol intake, body mass index (BMI), and physical activity, and their complex interactions (46) (47). Blood pressure has long been established as an inheritable trait (48), and an estimated 30-60% of blood pressure variation is explained by genetic factors (49). Longitudinal data of Framingham Heart Study suggested that 57% and 56% of inter-individual variability in systolic (SBP) and diastolic blood pressure (DBP), respectively, were due to genetic factors (50) in Caucasians. Data from Nigerian families suggest heritabilities of 40% and 36% for SBP and DBP in Africans (51), and data from the Chinese population suggest heritabilities of 31% and 32% for SBP and DBP, respectively, in Asians (52). Through GWAS meta-analyses in recent years, numerous loci have now been identified to be associated with blood pressure variation in different populations.

Genetic association studies

It is well established that genetic variation, including single nucleotide polymorphisms (SNPs), insertion/deletion polymorphisms, variable number tandem repeats, microsatellites, etc, could lead to important biological changes in protein production and/or structure, which provides biochemical basis for much of the diversity in the physiological characteristics of individuals, and also susceptibilities to various diseases and other disorders. Genetic association studies aim to compare the frequencies of the alleles or genotypes at each genomic site of interest between populations of cases and controls to determine whether genotype at the site is associated with susceptibility to certain diseases. A higher frequency of the less-common variant allele in cases is taken as evidence that the allele or genotype is associated with increased risk of disease (53). For quantitative traits, we assess whether genotype is associated with higher/lower levels of the trait in a cohort of individuals. Genetic association studies include many different study designs ranging from family studies to studies on tens of thousands of unrelated participants. Studies of quantitative traits can include either randomly selected individuals or highly selected individuals selected for having extreme phenotype values.

Candidate gene association studies

Candidate gene association studies have been a powerful approach to discover genetic variants associated with complex traits. Candidate gene studies are relatively cheap and quick to perform, and are focused on the selection of genes that, presumably, have some relevance in the mechanism of the disease being investigated. The selection of candidate genes often comes from prior knowledge about gene function (54). To date, candidate genes have been confirmed for many different diseases and traits (55). However, many published candidate gene association studies have been limited by small sample sizes, reducing statistical power, and inadequate correction for multiple testing, where the multiple test correction for a study only reflects the number of variants tested in the reported gene and not all variants tested across all previous genes considered but not reported. Thus, historically, few candidate gene studies have achieved robust results that could be replicated by other investigators. Despite its decreasing popularity, candidate gene studies are still being used as a complementary approach for GWAS, particularly for replication analyses. Further, because GWAS commercial arrays have often been agnostic with respect to variant function when selecting variants for inclusion and preference is given to linkage disequilibrium tagging variants, the coverage of biologically functional variants is often poor in GWAS studies. The poor coverage of functional variants on many GWAS panels has led to the development of gene-centric arrays that facilitate the investigation of a large number of variants, including uncommon functional variants, in genes involved in relevant biological pathways (56, 57), As a result, recent candidate gene association studies have achieved success in identifying genetic associations that are often replicated in independent samples. One successful example is the identification of blood pressure-related SNPs using the HumanCVD BeadChip, which includes approximately 50,000 SNPs from 2000 genes known to be associated with CVD-related traits, based on a discovery-stage sample of 25118 subjects and a replication-stage sample of 59349 subjects (57).

Genome-wide association studies

With the emergence of high-throughput genotyping techniques, array-based GWAS has become a widely used approach to identify genetic variants associated with common complex diseases. GWAS is a comprehensive discovery-driven approach to systematically test SNPs across the genome for association with dichotomous or continuous traits. By taking advantage of linkage disequilibrium (LD) and genotype imputation technology to capture most of genome's variation, a group of tag SNPs can be genotyped and used to evaluate the majority of genome-wide common genetic variation. Since the first GWAS reported in 2005 (58), more than 1500 GWAS have been conducted leading to the successful identification of susceptibility loci for a wide range of human complex diseases. In order to increase statistical power to detect some associations with variants having modest effects, consortia, with the purpose of conducting GWAS meta-analyses in very large samples, have been formed (59). Despite the huge success it has achieved, GWAS still has some limitations. Since GWAS evaluates a large number of individual SNPs ($>10^6$), the threshold of statistical significance is usually set to be very stringent ($P<5x10^{-8}$) in order to avoid false-positive associations, thus some variants with modest effects are not easy to be detected unless large-scale GWAS meta-analyses are conducted. More importantly, the majority of loci identified by GWAS are in the intergenic or intronic region of the genome, and the function of these variants are largely unknown. Post hoc investigations need to be conducted in order to learn whether these variants of unknown function have effects on mRNA expression or any other physiological activities on a molecular, cellular, or systematic level. Although a large number of variants were identified by GWAS, they collectively explain a limited proportion of the total variation of most of the traits being studied, leaving a large proportion of heritability unexplained. Therefore, new approaches are being developed to help explain the missing heritability.

Exome sequencing and Exome Beadchip

It has been hypothesized that much of the missing heritability that GWAS fails to explain may reside in low-frequency or rare variants (60). The successful identification of rare variants that have a large influence on lipid traits, found by sequencing extremes of the population distribution, is one example of the large role that rare variants can play on complex diseases (61). While early sequencing studies typically only sequenced a limited number of genes, the emergence of next-generation sequencing (NGS) technology has now enabled investigators to deeply sequence large stretches of DNA, whole exomes, and even entire genomes in large population-based studies (62). The National Heart, Lung, and Blood Institute launched a whole-exome sequencing project based on >6500 individuals to identify low-frequency and rare variants associated with heart, lung, and blood disorders (63). Exome sequencing has led to the discovery of thousands of rare functional variants that have now been included on genotyping arrays such as the Illumina HumanExome BeadChip. The Illumina HumanExome BeadChip is a genotyping array containing 247,870 variants discovered through exome sequencing in ~12,000 individuals, with ~75% of the variants with a MAF<0.5%. The main content of the chip comprises protein-altering variants (nonsynonymous coding, splice-site and stop gain or loss codons) seen at least three times in a single study and in at least two studies (64). So far, it has enabled identification of several rare functional variants associated with fasting glucose, insulin processing, and type 2 diabetes susceptibility (65) (66).

Gene-gene and Gene-environment interactions

It is widely believed that human complex traits are influenced by the interaction between genetic and environmental (G-E) factors. It is also widely held that such interactions will help explain some of the missing heritability of GWAS and provide better insight into pathway mechanisms for complex diseases (67). Quantifying G-E interactions may help develop improved predictive models, either for disease onset or for response to treatment. Improved predictive models of disease risk can enable preventive strategies, particularly when the risk factors include modifiable environmental exposures, and can advance individualized medicine by assessing a patient’s chance of responding to a particular treatment regimen (68). To date, most interaction studies are conducted after main effects have been identified, as current methods for detecting interactions on a genome-wide scale suffer from a lack of power due to the high cost of multiple test correction (69). However, scanning for the main effects might miss important genetic variants specific to subgroups of the population. In fact, interactions with opposite effects in 2 different exposure groups (cross-interactions) will not show a main effect and will, therefore, not be identified using standard approaches (67). Therefore, there is a need for the further development of methods to identify gene-environment interactions in the context of genome-wide association studies.

Genetic admixture mapping

While early GWAS were performed mostly on Caucasians, nowadays more and more GWAS studies are being conducted in other populations, including admixed populations. The fact that many of the significant variants discovered in Caucasians are also found to reach genome-wide significance in other populations suggests that the physiologic effects of many common variants may be generalized across populations with diverse genetic backgrounds (70). Admixed populations are populations that have variable levels of inherited ancestry from more than one ancestral population. Publically available genotype data (e.g. from HapMap) from contributing parental ancestral populations to the admixed population allows us to probabilistically infer the ancestral origin of alleles in a given individual at a given marker location. For example, for AAs, with available genetic data we can accurately infer at any given location the probability that the participant inherited 0, 1 or 2 copies of an African derived chromosome using HapMap CEU (Northern European) and YRI (Nigerian) haplotype data. Genetic admixture mapping is an approach to investigate whether genomic factors specific to one contributing ancestral population disproportionately influences certain traits in admixed population (71, 72). Since traditional admixture mapping has typically a substantially lower mapping resolution than association analysis (73), due to the often long stretches of ancestral haplotypes kept intact from the relatively recent admixture events, and the increasing availability of high-density genetic data that includes variants specific to one ancestral population or another, the analysis of admixed populations is now moving from admixture mapping towards SNP association.

In admixed populations, population stratification (the confounding between background ancestry and individual variant allele frequencies on traits with ancestry-variable prevalence) can lead to false associations. In order to control for this population stratification, several methods have been developed in the past decade, including genomic control (74), structured association (75), and principal components-based methods (76). The general idea of these methods is to use markers across the genome to capture the global population structure within the study subjects. These methods may be ineffective in removing the effect of population stratification if it is induced by natural selection that occurs only at certain genomic regions. Thus, for admixed populations, controlling for local ancestry, in addition to global ancestry, provides an additional layer of protection from reporting false associations.

Multistage association studies

Historically, the cost of genotyping 100 000's of variants using commercial genotyping arrays has limited the available sample sizes for GWAS studies. To remedy these costs, multistage GWAS studies were often utilized, where a subset of participants were included on the GWAS commercial array (Stage 1) and the remaining samples were genotyped on considerably smaller, and cheaper, arrays (Stage 2) containing a subset of variants demonstrating evidence for association in Stage 1. Data from both stages are combined for the overlapping set of variants to assess the overall significance of any discoveries. It has been shown, under careful design, that the two-stage study has similar statistical power to conventional single-stage studies but at a fraction of the cost.

As of today, hundreds of thousands of subjects with available GWAS data exist and much of the focus is now on measuring and analyzing novel biomarkers in cohorts with these existing GWAS data. As a result, the burden of expense is now shifting from the cost of genotyping to the cost of additional phenotyping, especially when phenotyping involves mRNA abundance data obtained from microarray or RNA-seq experiments, novel blood-based biomarkers or complex physiological and behavioral trait scale variables. Often it is not feasible, due to cost constraints, to phenotype all subjects with available genetic data. Selective phenotyping is a way of capturing the benefits of large population sizes without the need to carry out large-scale phenotyping (77). It utilizes available genetic information to select subsets of individuals to be phenotyped in a manner to increase overall power for the study. Methods to maximize power using selective phenotyping have been developed for a long time (78-81). The common rationale behind these methods is to identify the subset of subjects who are as genetically dissimilar as possible with respect to distributions of the genotype data for the specific markers of interest.

Selective phenotyping designs in genetic studies have been proposed for one-stage genetic association studies where there is an established set of variants of interest for study. Unfortunately, for most novel traits the identity of the “interesting” variants is unknown. GWAS studies are largely discovery-based studies. A reasonable approach to increase statistical power under fixed cost constraints is to randomly phenotype a subset of subjects and test across all variants (Stage 1) in order to identify a subset of “interesting” variants that can be targeted for selective phenotyping (Stage 2) in the remaining participants with genetic data. The results from

Stages 1 and 2 are then combined to assess significance of all variants (genotyped or imputed), where power has been enriched, by using selective phenotyping, for the subset of variants demonstrating strong evidence for association in Stage 1.

This dissertation consists of four chapters. The first two chapters aimed to identify the genetic variants associated with levels of iron-related measures and Lp(a) in a population of African Americans (AAs) from JHS using genome-wide admixture mapping, GWAS, and rare-coding-variant analyses using Exome Beadchip data. The third chapter aimed to assess the interaction between genetic factors and BMI on affecting blood pressure levels in a population- and household-based cohort study from China using a targeted candidate gene association study. The fourth chapter aimed to describe a two-stage GWAS using selective phenotyping for cohorts with existing genetic data who are subject to budgetary constraints when measuring new traits of interest.

.

CHAPTER II: GENOME-WIDE ADMIXUTRE AND ASSOCIATION STUDY OF SERUM IRON, FERRITIN, TRANSFERRIN SATURATION, AND TOTAL IRON BINDING CAPACITY IN AFRICAN AMERICANS [1]

Introduction

Iron is critical to an array of metabolic functions, such as oxygen transport and oxidative phosphorylation. Normally, small daily losses of iron in the feces and through menstruation are balanced by its regulated intestinal absorption and its recovery from heme after phagocytosis of senescent red blood cells (10). Iron deficiency can cause anemia, while iron overload may lead to increased risk for cardiovascular disease including cardiomyopathy, diabetes mellitus, arthritis, and liver disease (18). Laboratory tests typically used to assess iron transport and storage included serum iron, total iron binding capacity (TIBC), transferrin saturation (SAT), and serum ferritin. Ferritin, the predominant iron storage protein, reflects the cumulative iron stores in the bone marrow and tissues. Transferrin functions in iron transport, and the concentration of transferrin is proportional to the total iron binding capacity (TIBC) in serum. Transferrin saturation (SAT), calculated as (serum iron/TIBC) × 100, is affected by the rate of iron absorption in the small bowel as well as the sufficiency of tissue iron stores. Genetic factors could affect iron metabolism through gastrointestinal absorption, transport, tissue uptake, storage, or remobilization from tissue stores, and thus could have a large impact on the variation of these iron-related measures (18, 19).

[1] A version of this work was previously published as Li J *et al. Hum Mol Genet*. 2015 Jan 15; 24(2): 572-81. Epub 2014 Sep 15.

Genome-wide association studies (GWAS) have identified seven loci associated with these measures in subjects of European descent ($p<5x10^{-8}$) (20, 82-84), including the variants in or near *TF-SRPRB, SLC17A1, HFE, HIST1H2BJ, SIK3, PCSK7, TMPRSS6*. Here we present a genome-wide admixture and association study of serum iron, TIBC, SAT, and ferritin among African Americans (AA) enrolled in the JHS and HANDLS cohorts.

Materials and methods

Study Subjects

Discovery Stage – The Jackson Heart Study (JHS) is a longitudinal, population-based cohort designed to identify risk factors for the development of CVD, T2D, obesity, chronic kidney disease and stroke in more than 5000 AAs from the Jackson, metropolitan area(85). The design, recruitment and initial characterization of this study have been described previously (86). The JHS participants for the current study included 1012 AA males and 1335 AA females with available iron-related measures and genome-wide genotype data.

Replication Stage - The Healthy Aging in Neighborhoods of Diversity across the Life Span study (HANDLS) is a community-based, longitudinal epidemiological study that aims to examine the influences of race and socioeconomic status on the development of age-related diseases in African and European Americans from the city of Baltimore. The study consists of 2200 AAs and 1522 European Americans aged 30-64. The design, recruitment, and initial characterization of this study have been described previously (87). The HANDLS participants for the current study included 329 AAs with iron measures and genome-wide genotype data.

JHS and HANDLS participants provided written informed consent. The study protocols and consent forms for these studies were approved by the responsible research ethics committees and institutional research boards.

Phenotypes

All phenotype measures came from blood samples that were collected from fasting blood during the baseline examination, which occurred during 2000-2004 for JHS and 2004-2009 for HANDLS. Iron measures included total levels of serum iron (*μg*/dL), serum ferritin (*n*g/mL), TIBC (*μg*/dL) and SAT (%). Serum iron in JHS participants was measured by the FerroZine colorimetric assay (Roche), standardized to NIST traceable iron standards and calibrated against control sera from the manufacturer. TIBC was determined by colorimetric (FerroZine) measurement of iron that remains unbound after addition of a known amount of iron to the serum. The assay was standardized and calibrated as for serum iron. Ferritin was measured by an immunoturbidimetric assay (Roche) based on agglutination of anti-ferritin-latex conjugates, standardized with human spleen ferritin and calibrated against a standard protein solution provided by the manufacturer. For HANDLS, serum iron and TIBC were measured using standard clinical laboratory spectrophotometric assays. Serum ferritin was measured using chemiluminescence assays. For both studies, SAT was calculated as: (serum iron/TIBC)*100%.

Participants were excluded if they were taking iron supplements or not fasting at time of blood draw and if they had known chronic infectious or inflammatory disease, or residual cancer. Additional exclusions included hematocrit <35%, hemoglobin <11 g/dl, mean red cell volume >100 fL, white blood cell count >11,000 /mm^3, platelet count >400,000 /mm^3, C-reactive protein > 3 standard deviations above the mean, or transferrin saturation < 15% (indicates iron deficiency likely due to blood loss).

Genotyping and Imputation

A total of 3030 JHS participants were genotyped using the Affymetrix Genome-Wide Human SNP Array 6.0. 874,712 SNPs, with a call rate greater than 0.95, minor allele frequency (MAF) greater than 0.01, and genotype distributions consistent with Hardy-Weinberg equilibrium (HWE $p > 1x10^{-5}$) were included for further analysis. Following pre-phasing using MACH 1.0.18 software (88), thirty-eight million SNPs, excluding SNPs monomorphic in CEU/YRI, were imputed using minimac (89) based on 1000 Genome Project phase I reference samples (Nov 2010, Version 3). Analyses were limited to the ~17 million imputed SNPs with estimated imputation quality of r^2 greater than 0.3.

A total of 1024 HANDLS participants were genotyped using the Illumina 1M SNP array, including 329 AAs with iron measures. SNPs with HWE $p > 1x10^{-7}$, MAF > 0.01, and call rate > 95% were included for further analysis. 2,939,993 SNPs were imputed using MACH (88) and Minimac (89) software based on combined reference haplotype data from HapMap Phase 2 CEU+YRI samples that includes monomorphic SNPs in either of the two constituent populations (release 22, build 36.3). Chromosome X variants were imputed based on 1000 Genomes Project EUR+AFR+AMR+ASN reference samples. Only index variants demonstrating significant evidence for association in JHS ($p<5x10^{-8}$) were subsequently analyzed for the relevant iron phenotype in HANDLS.

Statistical Analyses

To assess the impact of genetic admixture on iron measures within the AA population, we first estimated the genome wide average of African ancestry for each JHS participant ("global ancestry"). We used the software ADMIXTURE (90) with K=2 clusters and tested, using linear models implemented in R, whether this estimated global ancestry proportion was associated with each of our iron measures after covariate adjustments for age, sex, and BMI. Values of serum ferritin, total iron and SAT were natural log-transformed to achieve approximate normality of residuals.

ANCESTRYMAP (72) was used to estimate local ancestry (probabilities of whether an individual has 0, 1, or 2 alleles of European ancestry) at 738,831 autosomal SNPs across the genome, for each participant in JHS, as previously described (91). In brief, local ancestry was inferred using a hidden Markov model based on the genotypes from a panel of densely spaced markers differentiated in frequency between African and European populations. To assess whether there were any regions where local ancestry was associated with iron-related measures, we performed admixture mapping across the whole genome by regressing each of our iron measures on the local ancestry estimates at each SNP location, including covariate adjustment for age, sex, BMI and estimated global ancestry. The conventionally reported LOD score, defined as the log, base 10, ratio of the maximum likelihood of the data under a local-ancestry-associated model divided by the likelihood of the data under the null model (with no local ancestry predictor), was computed at each SNP location. For regions showing association of increased African ancestry with higher levels of iron measures, the LOD scores were assigned positive values, and for regions showing association of increased African ancestry with lower level of iron measures, the LOD scores were assigned negative values. The LOD scores were plotted across the genome, and a LOD score of 5 was assumed to be the threshold of statistical significance (72).

Individual genotyped and imputed SNPs were tested for association using multivariable linear regression models in PLINK (92) and MACH2QTL v.1.08 (88), respectively; adjusted for age, sex, BMI and 10 principal components that we constructed using the software EIGENSOFT

(93) to model background ancestry. A second level of covariate adjustment for log-ferritin additionally included self-reported menopausal status, which was available on a large subset of JHS subjects and has been shown previously to be a strong predictor of serum ferritin levels (94, 95). We assumed an additive mode of inheritance and reported β coefficients representing the estimated change in the raw or transformed trait value, associated with each additional copy of the effect allele. For chromosome X SNPs, hemizygous males were modeled so that males with the minor allele had the same value as females homozygous for the minor allele. We used a significance threshold of p =5x10^{-8} to maintain an overall type 1 error rate of ~5% for each phenotype.

Manhattan plots were made to illustrate the association results across the genome and quantile-quantile (Q-Q) plots were made to assess any systematic inflation of the regression test statistics across the genome. In regions demonstrating significant evidence for association, we examined multivariable regression models that included the genotype data of the most strongly associated SNP as a covariate to assess whether there was any evidence for multiple independently associated SNPs in a particular region. If a second signal also reached genome-wide significance after conditioning on the top variant, multivariable regression models were repeated to include the genotypes of both SNPs as covariates. The relevant SNP-SNP interactions were also tested. Region-specific ("locus-zoom") plots were made to show the magnitude of association between all SNPs and the relevant iron phenotype as well as the LD between each SNP in the region and the most strongly associated SNP. Finally, to control for possible confounding between SNP genotype and local ancestry in any observed iron trait-SNP associations, we identified the genetic position of the most strongly associated SNP, selected the local ancestry estimate at the location closest to that SNP (either the SNP itself if genotyped or the closest genotyped SNP), and examined multivariable regression models as described above, but now including estimated local ancestry proportion as an additional covariate.

SNPs that reached genome-wide significance in JHS were selected for testing in the replication study of HANDLS. At each associated region, only the single SNP with the most significant P value (index SNP) was selected to avoid over adjustment for multiple testing. For the *TF* region, where conditional analyses revealed a second independently associated SNP with TIBC, the second independent SNP was also included in the replication analyses. A GWAS result was considered replicated if the effect in the replication was in the same direction as in the discovery stage, and if the association in the replication stage was statistically significant after Bonferroni correction adjusting for the number of SNPs tested.

Ethnic differences in iron related measures have also been observed between subjects of European and African descent (96). Thus, we compared our association results in JHS with established variants from GWASs in populations of European descent in order to assess the importance of these same variants in an AA population. For each prior GWAS-established SNP, we identified and tested all genotyped or imputed proxy SNPs in JHS that were estimated to be in high LD ($r^2 > 0.8$ in CEU based on 1000 Genomes data) with the GWAS index SNPs for association with the reported iron phenotype.

Results

Descriptive statistics for the JHS and HANDLS participants included in this study are detailed in **Table 2.1**. The correlations between these four phenotypes in JHS are shown in **Table 2.2**.

Admixture Analyses

A higher level of estimated overall average African ancestry was significantly associated ($p<0.05$) with lower levels of TIBC, iron and SAT (**Table 2.3**). The estimated proportions of overall average African ancestry obtained from ADMIXTURE were highly correlated with the first principal component from EIGENSOFT (correlation = 0.998). No individual region had local ancestry significantly associated with the iron measures. Three regions (*DZIP1L* on chr 3, *TRDN* on chr 6, and *TMCO5B-RYR3* on chr 15) had a LOD score >3 for TIBC, one region (*DEFB129-DEFB132* on chr 20) had a LOD score >3 for iron. The plots of the LOD scores across the whole genome are shown in **Figure 2.1.**

Summary GWAS Results in JHS

One-hundred-fifty-seven SNPs reached genome-wide significance ($p<5x10^{-8}$) for TIBC (153 SNPs on chromosome 3, 3 SNPs on chromosome 6, and 1 SNP on chromosome 16); five SNPs were significant for ferritin (all on chromosome X). Top SNPs that reached genome-wide significance in JHS was shown in **Table 2.4**. No SNPs reached genome-wide significance for serum iron or SAT. Top results for all four traits are listed in **Tables 2.5-2.8**. Manhattan plots and Q-Q plots for the four traits are shown in **Figure 2.2 and Figure 2.3**. Q-Q plots revealed no substantial evidence for inflated results, due to population stratification, residual relatedness among subjects, or experimental outliers.

TIBC GWAS Results on Chromosome 3

All top SNPs on chromosome 3 clustered within a region spanning less than 150 kb containing 3 genes: *TOPBP1*, *TF,* and *SRPRB* (**Figure 2.4**). The strongest signal (rs8177253,

$p=1.8\times10^{-47}$, MAF=0.24) mapped to the *TF* gene, which encodes transferrin. Rs8177253 was also nominally associated with SAT ($p=3.0x10^{-7}$)**.** Thirty-six *TF* region SNPs were genome-wide significant for TIBC levels after conditional analysis including rs8177253 as a covariate (**Table 2.9**). The top SNP in the conditional analysis was rs9872999 ($p=5.4x10^{-20}$, MAF=0.38), which was not significant ($p=1.0x10^{-6}$) prior to the adjustment for rs8177253 (**Table 2.9**; **Figure 2.4**). There was no evidence of an interaction between the rs8177253 and rs9872999 (p=0.11). No SNPs remained genome-wide significant after covariate adjustment for both rs8177253 and rs9872999, though a large number of SNPs remained nominally significant. SNP rs8177253 and rs9872999 together explained an estimated 11.2% of the total variance of TIBC after accounting for age, gender, BMI, and the first 10 PCs.

Higher local African ancestry in the *TF* region was nominally associated with lower TIBC levels (p=0.0012). The association between TIBC and rs8177253 remained robust after the adjustment for local African ancestry at rs8177253. When stratified by the estimated local number of European versus African chromosomes, the rs8177253-TIBC association was present among 1579 AA who were predicted to carry two African chromosomes ($\beta= 19.64 \pm 1.70$; $p=8.5x10^{-30}$) as well as the 768 AA who were predicted to carry at least one European chromosome ($\beta = 19.68 \pm 2.26$; $p=1.8x10^{-17}$). The rs9872999-TIBC association also remained significant after adjustment for both rs8177253 and local African ancestry at the genotyped marker nearest its location (data not shown).

TIBC GWAS Results on Chromosome 6

The three genome-wide significant SNPs on chromosome 6 mapped near the *HDGFL1* gene (**Figure 2.5**). The strongest signal (rs115923437, $p=1.1x10^{-8}$) mapped ~100Kb distal to

HDGFL1, which is a gene that encodes hepatoma derived growth factor-like 1 and is associated with glycosylated hemoglobin level, and ~3.5Mb proximal to the known iron gene *HFE*. After conditioning on rs115923437, the remaining SNPs within 1Mb of the SNP no longer showed strong evidence for association (all p>1x10^{-4}). SNP rs115923437 explained 1.3% of the total variance of TIBC after accounting for the covariates.

Local African ancestry in the *HDGFL1* region also showed a nominally significant association with TIBC, though this time higher local African ancestry was associated with higher TIBC levels (p=0.0083). The associations between TIBC and rs115923437 remained significant after the adjustment for local African ancestry at the genotyped marker nearest its location. When stratified by the estimated local number of European versus African chromosomes, the rs115923437-TIBC association was present among the 1569 AA who were predicted to carry two African chromosomes (β = -14.11 ± 2.95; p=2.0x10^{-6}), as well as the 778 AA who were predicted to carry at least one European chromosome (β = -15.42 ± 5.78; p=0.0070).

TIBC GWAS results on chromosome 16

A single SNP (rs16951289, p=2.0x10^{-8}) on chromosome 16, an intronic variant in uncharacterized gene *LOC102467146* that is ~150Kb distal to the *MAF* gene (**Figure 2.6**), reached genome-wide significance. *MAF* encodes v-maf musculoaponeurotic fibrosarcoma oncogene homolog. Local African ancestry in this region was not associated with TIBC (p=0.97) and did not modify the evidence for association with rs16951289. Rs16951289 explained 1.2% of the total variance of TIBC.

Ferritin GWAS results on chromosome X

Five SNPs on chromosome X near the *GAB3* gene reached genome-wide significance for association with ferritin levels (**Figure 2.7**). The strongest signal (rs141555380, p=1.1x10^{-8}), mapped to the UTR-3 region of the *GAB3* gene. *GAB3* encodes GRB2-associated binding protein 3, which is involved in several growth factor and cytokine signaling pathways. After conditioning on rs141555380, the remaining SNPs within 1Mb of rs141555380 no longer showed significant evidence for association (all p>1x10^{-4}). The effect estimates for carriers of the rs141555380 minor allele in stratified analysis were similar for hemizygous males and homozygous females (hemizygous males: β = 0.17±0.04, p = 5.05 x 10^{-6}; homozygous females: β = 0.14±0.05, p = 0.002). SNP rs141555380 explained 1.2% of the total variance of ferritin after accounting for age, gender, BMI and the first 10 PCs.

Menopause status-Adjusted analyses of association with ferritin

Association analysis for serum ferritin was conducted by including menopausal status as an additional covariate. Menopausal status was significantly associated with serum ferritin levels p= 5.3x10^{-17}, however, the effect size and significance of SNP-ferritin associations did not change considerably after adjusting for menopausal status. The top SNP rs141555380 (p=1.1x10^{-8}) continued to be the most significant SNP after adjusting for menopause status (p=1.4x10^{-8}). The effect sizes and P-values of top SNPs (p<10^{-7}) before and after adjusting for menopause status are shown in **Table 2.10** This result is consistent with previous findings that the effects of variation in menstrual blood loss, although significant, were small when compared to the genetic effects that influence the iron reserves (94).

Replication Results in HANDLS

Four regions contained SNPs that reached genome-wide significance in GWASs in JHS, including one region, the chromosome 3 *TF* region, which contained a second significant SNP (rs9872999) after covariate adjustment for the top SNP (rs8177253). Five SNPs (rs8177253, rs9872999, chromosome 6 top SNP rs115923437 and chromosome 16 top SNP rs16951289 for TIBC; chromosome X top SNP rs141555380 for ferritin) were selected and tested for association in HANDLS to determine whether the associations could be replicated (**Table 2.4**). Rs9872999 was tested before and after adjustment for rs8177253. Both associations in the *TF* region with TIBC were replicated ($p=1.1x10^{-7}$ for rs8177253; p=0.0012 for rs9872999 before adjusting for rs8177253)**.** As with JHS, the association between rs9872999 and TIBC became more significant after adjusting for the primary signal at rs8177253 ($p=6.2x10^{-6}$). The association between the *GAB3* region SNP rs141555380 and ferritin also replicated ($p=5.7x10^{-3}$). The associations between TIBC and SNPs rs115923437 and rs16951289 did not replicate in HANDLS; rs16951289 was nominally significant (p=0.04), but the direction of the effect went in the opposite direction to that observed in JHS.

Prior European GWAS-established signals at $p<5x10^{-8}$ that are replicated in JHS at p<0.05

GWAS have reported five regions to contain SNPs to be significantly associated ($p<5x10^{-8}$) with at least one of the following iron-related traits: iron, ferritin, SAT and transferrin, including *TF-SRPRB* (rs3811647, rs1830084), *SLC17A1* (rs17342717*), HFE* (rs1799945 [H63D] and rs1800562 [C282Y]*), HIST1H2BJ* (rs13194491), and *TMPRSS6* (rs855791 [V736A] and rs4820268*)* in subjects of European descent (**Table 2.11**). Iron, ferritin, and SAT are directly reported in the current study while TIBC is proportional to transferrin [TIBC (μmol/L) = 25.1 × transferrin (g/L)]. All five regions contained a SNP, either the index SNP in the prior report or a

SNP in strong linkage disequilibrium (defined as estimated $r^2>0.8$ with the index SNP based on CEU 1000 Genomes subjects) with the index SNP, that was nominally associated with at least one corresponding iron measure in JHS at p<0.05. For *HFE* rs1800562, only the associations with ferritin and transferrin were replicated, but not the associations with iron and SAT. For *HIST1H2BJ* rs13194491 and *TMPRSS6* rs855791, the associations with the index SNPs were not replicated, but nearby SNPs, which are estimated to be in high LD with index SNPs in European populations, did reach nominal significance in JHS, suggesting a narrowing of the candidate regions for the causal variants if the causal variants are the same for the two populations.

Discussion

We conducted genome-wide admixture and association studies for iron-related phenotypes including serum iron, serum ferritin, SAT, and TIBC in 2347 AAs participating in the Jackson Heart Study. We observed significant associations between SNPs around *TF,* a well-established region on chromosome 3, and two novel regions: near *HDGFL1* on chromosome 6 and *MAF* on chromosome 16, and TIBC levels. Conditional analyses revealed a second significant SNP associated with TIBC in the *TF* region that was independent of the top SNP from the unconditional analyses. We also observed significant associations between SNPs around *GAB3,* a novel region on chromosome X, and ferritin levels. The two independent associations for TIBC at *TF* and the association for ferritin at *GAB3* were successfully replicated in HANDLS. The associations between the JHS index SNPs near *HDGFL1* and *MAF* were not replicated. The available sample size in HANDLS (N=329) was considerably smaller than for JHS. If we assume the index variants at *HDGFL1* and *MAF* each explain an estimated 1.3% of the total variance of TIBC, then we only had power of 0.34 to observe these associations at p<0.01 in HANDLS.

Thus, larger replication studies will be necessary to make more decisive statements regarding the overall evidence in these other regions. We did not observe any significant SNP associations for the other iron measures, though we were able to replicate associations from previous studies based on subjects of European descent using a less stringent significance threshold ($p<0.05$).

The estimated average (“global”) proportion of African ancestry was significantly associated with lower levels of TIBC, serum iron, and SAT - which are entirely consistent with previous findings reporting lower levels of these same measures, on average, in AAs compared to European Americans (96, 97). Subjects with higher levels of global African ancestry were observed to have higher levels of ferritin, also consistent with prior reported differences between individuals of European and African ancestry (96, 97), although this result was not statistically significant. These results implicate novel genetic risk factors in AAs and underscore the importance of studying this population for genetic risk factors that uniquely/disproportionately impact them. Local ancestry was not significantly associated with any iron measures, though a couple regions containing our GWAS top results were nominally significant.

Variants in and around *TF* have been observed to be associated with serum ferritin levels (20), transferrin levels (20, 83, 84), serum transferrin saturation (22), and serum levels of carbohydrate-deficient transferrin (CDT) (98) in subjects of European descent. In JHS, the top SNP at *TF*, rs8177253, was associated with TIBC ($p=1.8x10^{-47}$) and nominally associated with SAT ($p=3.0x10^{-7}$). SNP rs8177253 is located in an intronic region of *TF*, and is in high LD ($r^2=1$, D’=1 in 1000 Genomes CEU) with GWAS index SNP rs3811647 previously reported to be associated with transferrin in European Americans (rs3811647 is also associated with TIBC [$p=9.7x10^{-35}$] and nominally associated with SAT [$p=2.2x10^{-6}$] in JHS). After conditioning on rs8177253, all SNPs, including rs3811647, previously reported to be significantly associated with TIBC became non-significant in JHS.

Our study is the first study to report a second significant independently associated SNP in the *TF* region for TIBC. Interestingly, this second signal (index SNP rs9872999 which maps to an intergenic region approximately 10Kb proximal to *TF*) only became significant at the genome-wide level after conditioning on our top TF region SNP rs8177253. Rs8177253 and rs9872999 are in modest LD in 1000 YRI Genomes subjects (r^2=0.07,D'=0.58) and in stronger LD in 1000 Genomes CEU subjects (r^2=0.27,D'=0.71). The allele associated with an increase in TIBC for rs8177253 is preferentially on the same haplotype with the allele associated with a decrease in TIBC for rs9872999. Thus, the mean effects for rs9872999 are shrunk towards the null when not factoring in genotype for rs8177253. There is no evidence for an interaction between these SNPs on TIBC levels, thus the results indicate a second independent signal in the *TF* region. It is unclear if this second signal is specific to African Americans, as results from conditional analyses in other populations have not been reported. Conditional analyses would likely be especially important in populations of European descent given the stronger LD between the two SNPs in this population. The minor allele frequency for rs9872999 is 0.39 in 1000 Genomes YRI subjects and 0.50 in 1000 Genomes CEU subjects.

Variants in *HDGFL1* have been reported previously to be associated with glycosylated hemoglobin levels ($p=2.4x10^{-5}$) in European type 1 diabetic subjects (99) and nominally associated ($p<0.05$) with levels of VLDL, LDL, Apolipoprotein C, HDL, and carotid artery disease (100-102). In recent years there has been considerable interest in the possibility that excessive tissue iron stores may contribute to the pathogenesis of both diabetes and ischemic heart disease (19). *MAF*, which encodes v-maf musculoaponeurotic fibrosarcoma oncogene homolog, appears to be important in early development. Mutations in *MAF* have been reported to co-segregate with cerulean congenital cataracts (103) and juvenile-onset pulverulent cataract (104) in human pedigrees. We can find no evidence in the literature suggesting a direct connection between *MAF* and iron metabolism. However, there is some evidence of pleiotropy for iron metabolism and cataracts, namely hereditary hyperferritinemia cataract syndrome (HHCS), which is an inherited syndrome caused by a mutation within the L-ferritin gene and characterized by early-onset cataracts and elevated serum ferritin (105).

A cluster of SNPs near *GAB3* on chromosome X were significantly associated with ferritin (top SNP rs141555380, MAF=0.14, p=1.1x10^{-8}) and nominally associated with SAT (rs141555380, p=0.037). Although no prior studies observed any connection between this gene and iron metabolism, another gene ~0.2 Mb upstream of *GAB3*, *G6PD*, plays a critical role in iron metabolism. *G6PD* deficiency may cause acute hemolysis or severe chronic non-spherocytic hemolytic anemia. Increases in serum ferritin levels have been observed in *G6PD*-deficient patients (106, 107), which is possibly due to both a shortened life-span and increased break down of erythrocytes in *G6PD*-deficient patients. A functional missense variant in *G6PD*, rs1050828 (MAF=0.13, leading to a Val68Met amino acid substitution), was also associated with ferritin but narrowly missed genome-wide significance (p=9.1x10^{-8}). Strong LD exists between this functional variant at *G6PD* and rs141555380 (R^2=0.91, D'=1 in 1000 Genomes Project participants of African descent), and the association between rs141555380 and ferritin disappears after adjustment for rs1050828 (p = 0.55). Rs1050828 and nearby rs762516 (two SNPs in LD: R^2=0.68, D'=1.0 in HapMap YRI) have been shown to be significantly associated with multiple erythrocyte traits in AAs, including hematocrit, hemoglobin, red blood cell (RBC) count, mean corpuscular volume (MCV), and red cell distribution width (RDW) in previous GWAS or candidate gene studies (108, 109). In JHS, rs1050828 is significantly associated with MCV (p=1.7x10^{-8}), RBC count (p=9.9x10^{-16}), and RDW (p=8.9x10^{-21}) and nominally associated with hematocrit (p=8.7x10^{-7}) and hemoglobin (7.1x10^{-8}). Ferritin levels are correlated with levels of hematocrit (r=0.25), hemoglobin (r=0.27), MCV (r=0.089), RBC (r=0.15) and RDW (r=-0.13). The associations for both hematocrit (p=2.9x10^{-8}) and hemoglobin (p=4.3x10^{-10}) both became genome-wide significant after additional covariate adjustment for ferritin. Similarly, the association between rs1050828 and ferritin also became genome-wide significant after adjustment for hematocrit (7.2x10^{-9}) and hemoglobin (1.3x10^{-9}), but evidence for association with ferritin decreased after covariate adjustment for MCV (1.1x10^{-4}), RBC (5.2x10^{-6}) and RDW (7.2x10^{-3}). Since *G6DP* has been reported to play an important role in hemolysis and affect the levels of erythrocyte traits, the signal we observed at this region may help explain the relationship between hemolysis and iron metabolism. This variant is also implicated in malaria resistance, and the A- form of G6PD deficiency in Africa is under strong natural selection from the preferential protection it provides to hemizygous males and homozygous females against life-threatening malaria (110). This natural selection of G6PD deficiency in African descent may help explain the marked differences in iron measures among ethnic groups.

In summary, we report that global genetic admixture is an important predictor of iron measures in AAs, further implicating the importance of unique genetic effect alleles in the AA population. We observed SNPs in or near three genes, *TF*, *HDGFL1* and *MAF,* which were significantly associated with TIBC in JHS, and SNPs near *GAB3* that were significantly associated with ferritin. We identified a novel second independently associated SNP in the *TF* region for TIBC that was only identified after conditioning on the top SNP in the region. The two *TF* and the *GAB3* signals were replicated in a small independent AA sample from HANDLS.

Larger replication samples will be necessary to draw firm conclusions regarding the associations for the other loci. The *TF* region is known to be associated with various serum iron-related measures in subjects of European descent; we now show similar associations in AA. While the *G6DP-GAB3* region is known to be associated with multiple erythrocyte traits in AA, this is the first time it has been reported to be significantly associated with ferritin, a specific iron-related measure in AA. We have also nominally replicated four other established loci from other populations in our AA samples. Future fine-mapping studies, including rare and uncommon variants, and functional studies should be undertaken to better characterize these and other loci and ultimately to identify the functional variants directly influencing iron levels in AA.

Funding

JHS is supported by the National Heart, Lung, and Blood Institute and the National Institute on Minority Health and Health Disparities, with additional support from the National Institute on Biomedical Imaging and Bioengineering [grant numbers HHSN268201300046C, HHSN268201300047C, HHSN268201300048C, HHSN268201300049C, HHSN268201300050C]. HANDLS is supported by the Intramural Research Program of the National Institute of Health, National Institute on Aging and the National Center on Minority Health and Health Disparities (project # Z01-AG000513 and human subjects protocol # 2009-149). EML, YL and QD are partially supported by R01HG006703.

Conflicts of interest

The authors declare no conflict of interests.

Table 2.1 Descriptive statistics of JHS and HANDLS participants for the study of iron-related phenotypes

	JHS			HANDLS		
	Total	Male	Female	Total	Male	Female
Sample size	2347	1012	1335	329	189	140
Age (years)	54.5 ± 12.6	53.0 ± 12.8	55.7 ± 12.4	49.4 ± 8.3	49.0 ± 8.7	49.9 ± 7.7
BMI (kg/m2)	31.5 ± 6.9	29.9 ± 6.1	32.8 ± 7.3	28.8 ± 7.6	27.1 ± 5.4	31.2 ± 9.3
Ferritin (ng/mL)	134.0 (74.0, 232.0)	177.5 (110.8, 286.0)	105 (58.0, 185.5)	107.5 (54.0, 201.0)	137.0 (76.0, 257.0)	71.0 (35.0, 113.0)
Iron (µg/dL)	81.0 (67.0, 98.0)	86.0 (70.0, 104.0)	78.0 (64.0, 94.0)	83.0 (66.2, 103.8)	86.0 (70.0, 109.0)	81.0 (64.0, 98.0)
SAT (%)	28.0 (23.0, 35.0)	30.0 (25.0, 37.0)	27.0 (22.0, 33.0)	25.0 (20.2, 31.1)	26.4 (20.8, 33.5)	23.7 (19.3, 27.3)
TIBC (µg/dL)	284 (260.0, 314.0)	280.0 (256.0, 307.0)	288.0 (263.0, 319.0)	332.0 (306.0, 372.0)	326.0 (300.0, 368.0)	342.5 (316.8, 379.0)

Data are mean±SE, median (25th, 75th percentiles)
SAT: Transferrin saturation; TIBC: Total iron binding capacity

Table 2.2 Correlation between iron-related phenotypes

	TIBC	log_ferritin	log_iron	log_sat
TIBC	1.00	-0.32	0.20	-0.29
log_ferritin		1.00	0.08	0.23
log_iron			1.00	0.88
log_sat				1.00

Table 2.3 Association between global African ancestry estimate and iron-related phenotypes

Phenotypes	β*	SE	*P*
TIBC	-13.90	6.80	0.04
log_ferritin	0.13	0.13	0.32
log_iron	-0.20	0.05	2.4E-05
log_SAT	-0.15	0.05	0.0019

* The predicted change in the iron measure for each one-percentage point increase in estimated global African ancestry.

Table 2.4 Top SNPs that significantly ($p<5x10^{-8}$) associated with iron-related phenotypes in JHS and replication results from HANDLS.

						JHS						HANDLS					
Trait	Chr	Nearest Gene	SNP	Pos(hg19)	EA	EAF	N	R^2	β	SE	*P*	EAF	N	R^2	β	SE	*P*
TIBC	3	*TF*	rs8177253	133480192	T	0.24	2347	NA	19.86	1.34	1.8E-47	0.23	329	0.98	26.72	4.91	1.1E-07
TIBC	3	*TF*	rs9872999	133457514	T	0.62	2347	0.80	-6.55	1.34	1.1E-06	0.60	329	0.93	-14.19	4.33	0.0012
TIBC	3	*TF*	rs9872999*	133457514	T	0.62	2347	0.80	-11.72	1.28	5.4E-20	0.60	329	0.93	-19.11	4.16	6.2E-06
TIBC	6	*HDGFL1*	rs115923437	22678302	C	0.06	2347	0.91	14.84	2.6	1.1E-08	0.05	329	0.96	-5.28	9.93	0.60
TIBC	16	*MAF-DYNLRB2*	rs16951289	79790621	T	0.07	2347	0.91	13.38	2.38	2.0E-08	0.08	329	0.97	-15.95	7.92	0.04
Log_ferritin	X	*GAB3*	rs141555380	153906012	T	0.14	2347	0.94	0.17	0.03	1.1E-08	0.13	329	0.98	0.24	0.08	0.0057

*: β, SE, and P were reported for rs9872999 after adjusting for rs8177253.
Chr: chromosome; Pos(hg19): physical position of the SNP according to human genome build version 19; EA: effect allele; EAF: effect allele frequency; β: β coefficients representing the estimated change in the raw or transformed trait value associated with each additional copy of the effect allele; SE: standard error; R^2: R^2 represents the imputation quality provided by minimach. "NA" indicates that the actual genotype data from Affy 6.0 array were used in the analyses.

Table 2.5 Top SNPs associated with TIBC at $P<10^{-6}$

Chr	Nearest Gene	# of SNPs	Most Significant SNP	Pos(hg19)	EA	EAF	β	SE	*P*	Imputed (Y/N)	R^2	Function
1	*BAI2*	1	rs182815963	32219928	T	0.00	-93.66	18.51	4.2E-07	Y	0.44	intronic
2	*SERTAD2-LOC400958*	7	rs185183332	64967342	A	0.03	-20.46	4.01	3.3E-07	Y	0.69	intergenic
2	*RPRM-GALNT13*	2	rs189855038	154428028	C	0.00	-90.97	17.96	4.1E-07	Y	0.31	intergenic
3	*LOC440970*	1	3:84190898:A_AT	84190898	AT	0.02	37.67	7.59	6.9E-07	Y	0.36	intergenic
3	***TF***	221	rs8177253	133480192	T	0.24	19.86	1.34	**1.8E-47**	N	NA	intronic
4	*DCK-SLC4A4*	1	rs147549040	72028308	G	0.00	85.49	17.17	6.4E-07	Y	0.39	intergenic
6	***HDGFL1-NRSN1***	10	rs115923437	22678302	C	0.06	14.84	2.60	**1.1E-08**	Y	0.91	intergenic
6	*MIR5695*	1	rs112647290	126435741	T	0.03	23.52	4.67	4.7E-07	Y	0.58	intronic
10	*FRMD4A-MIR1265*	1	rs59376103	14442418	T	0.04	-17.36	3.37	2.6E-07	Y	0.78	intergenic
10	*KCNMA1*	2	rs71475649	79082679	C	0.01	70.63	13.65	2.3E-07	Y	0.33	intronic
11	*LUZP2*	1	rs61877888	24602715	G	0.15	9.36	1.87	5.3E-07	Y	0.80	intronic
13	*SLITRK1*	2	rs192989741	83958956	G	0.00	-76.76	15.20	4.4E-07	Y	0.38	intergenic
13	*ATP11A-MCF2L-AS1*	1	rs114067519	113547283	C	0.04	22.38	4.48	5.9E-07	Y	0.41	intergenic
16	*IGFALS-HAGH*	1	rs2256923	1844781	G	0.72	-9.98	2.04	9.9E-07	Y	0.41	intergenic
16	*SLX4-DNASE1*	1	rs60860998	3690092	C	0.03	32.70	6.47	4.3E-07	Y	0.33	intergenic
16	***MAF-DYNLRB2***	2	rs16951289	79790621	T	0.07	13.38	2.38	**2.0E-08**	Y	0.91	intergenic
18	*PTPN2-SEH1L*	1	rs73407743	12943299	A	0.10	11.06	2.25	9.0E-07	Y	0.73	intergenic
21	*ADAMTS5-MIR5009*	1	rs13052896	28481882	G	0.08	11.04	2.25	9.8E-07	Y	0.83	intergenic
21	*HUNK*	1	21:33316076:CCTT_	33316076	C	0.04	18.04	3.63	6.5E-07	Y	0.74	intronic

Table 2.6 Top SNPs associated with log_ferritin at $P<10^{-6}$

Chr	Nearest Gene	# of SNPs	Most Significant SNP	Pos(hg19)	EA	EAF	β	SE	*P*	Imputed (Y/N)	R^2	Function
1	*RCC2-ARHGEF10L*	1	rs2477740	17831605	A	0.82	0.19	0.04	3.9E-07	Y	0.65	intergenic
3	*FHIT*	1	rs13074132	60321082	C	0.18	0.16	0.03	3.0E-07	Y	0.92	intronic
10	*RSU1*	1	rs76969309	16824397	A	0.05	0.33	0.06	8.3E-08	Y	0.72	intronic
10	*ADRA2A-GPAM*	1	rs17129425	113727270	G	0.19	-0.16	0.03	3.9E-07	Y	0.87	intergenic
18	*CEP192*	1	rs1787009	13063889	A	0.13	-0.20	0.04	7.6E-07	Y	0.76	intronic
19	*C19orf18*	4	rs192471137	58469524	A	0.00	1.72	0.32	1.2E-07	Y	0.40	intergenic
X	***G6PD-GAB3***	21	rs141555380	153906012	T	0.14	0.17	0.03	**1.1E-08**	Y	0.94	UTR-3

Table 2.7 Top SNPs associated with log_iron at $P<10^{-6}$

Chr	Nearest Gene	# of SNPs	Most Significant SNP	Pos(hg19)	EA	EAF	β	SE	*P*	Imputed (Y/N)	R^2	Function
2	*ERBB4*	5	rs139713297	212423072	A	0.00	0.74	0.14	1.6E-07	Y	0.33	intronic
2	*SPHKAP-PID1*	1	rs1727563	229669613	T	0.31	-0.06	0.01	6.1E-07	Y	0.59	intergenic
4	*CC2D2A*	1	rs188125026	15490175	C	0.00	1.61	0.32	5.2E-07	Y	0.37	intronic
4	*SLIT2*	3	rs143034438	20560387	A	0.01	0.28	0.05	8.7E-08	Y	0.48	intronic
4	*HSP90AB3P-SPP1*	1	rs143957415	88837386	T	0.03	0.15	0.03	2.1E-07	Y	0.65	intergenic
8	*EPHX2-CLU*	5	rs79882106	27414335	C	0.00	0.80	0.15	8.0E-08	Y	0.32	intergenic
8	*EYA1-MSC*	1	rs191222090	72718862	G	0.00	0.49	0.10	6.2E-07	Y	0.44	intergenic
15	*SPRED1*	1	rs1522782	38544590	G	0.88	-0.08	0.02	3.8E-07	Y	0.63	intergenic
15	*LINC00052-NTRK3*	1	rs139196658	88173631	T	0.00	1.50	0.31	8.6E-07	Y	0.36	intergenic
16	*RBFOX1*	1	rs2109459	6440752	T	0.47	0.05	0.01	6.2E-07	Y	0.64	intronic
16	*ZCCHC14*	1	16:87446849	87446849	C	0.00	0.82	0.16	2.3E-07	Y	0.37	intronic
X	*HTATFS1-VGLL1*	1	rs184017266	135612099	A	0.00	31.47	6.26	4.9E-07	Y	0.35	intergenic
X	*SETP8*	2	rs190571075	116387808	T	0.00	2.86	0.58	8.0E-07	Y	0.32	intergenic
X	*HTR2C-IL13RA2*	1	rs140631409	114166154	T	0.00	0.95	0.19	8.6E-07	Y	0.53	intergenic

Table 2.8 Top SNPs associated with log_SAT at $P<10^{-6}$

Chr	Nearest Gene	# of SNPs	Most significant SNP	Pos(hg19)	EA	EAF	β	SE	*P*	Imputed (Y/N)	R^2	Function
1	*LOC100129138-PRMT6*	1	rs113982601	106475276	A	0.10	0.09	0.02	9.4E-07	Y	0.56	intergenic
2	*SPHKAP-PID1*	1	rs6756903	229646101	C	0.31	-0.06	0.01	6.5E-07	Y	0.64	intergenic
2	*NGEF*	1	rs938575	233768788	A	0.10	-0.07	0.02	3.9E-07	Y	0.88	intronic
3	*TF*	4	rs6762719	133480817	G	0.24	-0.05	0.01	5.1E-07	Y	1.00	intronic
5	*LSM11*	1	rs145829393	157176316	T	0.02	0.25	0.05	7.8E-07	Y	0.44	intronic
8	*CSMD1-LOC100287015*	1	rs140656346	5682652	G	0.00	0.55	0.11	6.1E-07	Y	0.34	intergenic
8	*MFHAS1-ERI1*	1	rs115730735	8820986	T	0.05	0.13	0.03	8.7E-07	Y	0.60	intergenic
10	*RHOBTB1*	9	rs112298642	62634356	T	0.00	0.39	0.07	2.1E-07	Y	0.60	intronic
13	*KLF12-LINC00381*	1	rs67180317	74869018	G	0.27	0.06	0.01	2.9E-07	Y	0.67	intergenic
14	*ESRRB-VASH1*	1	rs75838009	77065324	T	0.02	0.21	0.04	1.1E-07	Y	0.75	intergenic
15	*SPRED1*	1	rs1522782	38544590	G	0.88	-0.08	0.02	4.8E-07	Y	0.63	intergenic
15	*CKMT1A-CATSPER2P1*	1	rs150805357	44016403	C	0.01	-0.30	0.06	6.3E-07	Y	0.53	intergenic
16	*ZCCHC14*	1	16:87446849	87446849	C	0.00	0.86	0.16	1.2E-07	Y	0.37	intronic
17	*ASIC2-CCL2*	1	rs142195977	32540548	T	0.03	-0.18	0.04	6.1E-07	Y	0.46	intergenic
19	*SMARCA4*	1	rs116337692	11107134	A	0.07	-0.10	0.02	6.2E-07	Y	0.63	intronic

Table 2.9 SNPs on Chromosome 3 that significantly ($P<5x10^{-8}$) associated with TIBC after adjusting for the top variant rs8177253

					Before adjusting for rs8177253			After adjusting for rs8177253					
Nearest Gene	MARKER	Pos(hg19)	EA	EAF	β	SE	*P*	β	SE	*P*	Imputed(Y/N)	R^2	Function
TOPBP1	rs112327474	133348668	A	0.03	23.15	4.87	1.96E-06	25.74	4.65	3.13E-08	Y	0.48	intronic
TOPBP1	rs113969369	133348677	A	0.03	23.17	4.86	1.87E-06	25.78	4.65	2.90E-08	Y	0.48	intronic
TOPBP1-TF	rs79213250	133391145	A	0.03	22.77	4.75	1.62E-06	25.41	4.54	2.16E-08	Y	0.55	intergenic
TOPBP1-TF	rs6806769	133431627	A	0.10	9.56	2.31	3.58E-05	13.24	2.21	2.12E-09	Y	0.69	intergenic
TOPBP1-TF	rs9830001	133433470	G	0.31	7.05	1.28	4.34E-08	7.04	1.23	1.06E-08	N	NA	intergenic
TOPBP1-TF	rs4078166	133435979	A	0.66	-6.73	1.33	4.23E-07	-11.03	1.27	4.11E-18	Y	0.86	intergenic
TOPBP1-TF	rs6782434	133438834	G	0.66	-6.76	1.33	3.46E-07	-11.06	1.27	2.65E-18	Y	0.87	intergenic
TOPBP1-TF	rs4443173	133439378	G	0.74	-4.65	1.39	8.19E-04	-7.37	1.33	2.75E-08	Y	0.90	intergenic
TOPBP1-TF	rs9843635	133440977	T	0.65	-6.71	1.35	6.42E-07	-11.29	1.29	1.85E-18	Y	0.83	intergenic
TOPBP1-TF	rs11921527	133441167	A	0.52	-0.84	1.26	5.06E-01	-6.81	1.20	1.51E-08	Y	0.85	intergenic
TOPBP1-TF	rs13066859	133442939	G	0.58	-6.65	1.41	2.43E-06	-10.97	1.35	4.20E-16	Y	0.71	intergenic
TOPBP1-TF	rs145713832	133445820	C	0.51	-8.14	1.49	4.63E-08	-10.40	1.42	2.72E-13	Y	0.62	intergenic
TOPBP1-TF	rs6804904	133447231	G	0.62	-6.30	1.34	2.83E-06	-11.08	1.29	6.41E-18	Y	0.80	intergenic
TOPBP1-TF	rs6439432	133448242	A	0.56	-5.51	1.39	7.54E-05	-9.17	1.33	5.44E-12	Y	0.72	intergenic
TOPBP1-TF	rs9820225	133449189	G	0.66	-6.77	1.31	2.26E-07	-10.94	1.25	2.08E-18	Y	0.89	intergenic
TOPBP1-TF	rs6439434	133450371	G	0.66	-6.45	1.28	5.00E-07	-10.81	1.23	1.18E-18	Y	0.93	intergenic
TOPBP1-TF	rs9869311	133451613	T	0.66	-6.48	1.28	3.75E-07	-10.79	1.22	8.74E-19	Y	0.94	intergenic
TOPBP1-TF	rs6439436	133453779	T	0.36	4.67	1.23	1.57E-04	9.75	1.21	1.07E-15	N	NA	intergenic

TOPBP1-TF	rs150431546	133455441	G	0.66	-6.69	1.29	2.05E-07	-11.01	1.23	3.96E-19	Y	0.92	intergenic
TOPBP1-TF	rs9872999	133457514	T	0.62	-6.55	1.34	9.97E-07	-11.72	1.28	5.44E-20	Y	0.80	intergenic
TOPBP1-TF	rs79354095	133458401	A	0.66	-6.82	1.31	1.99E-07	-11.29	1.25	2.30E-19	Y	0.89	intergenic
TOPBP1-TF	rs4854596	133459260	C	0.66	-6.81	1.31	2.02E-07	-11.08	1.25	9.02E-19	Y	0.89	intergenic
TOPBP1-TF	rs10935085	133459735	T	0.64	-4.86	1.28	1.51E-04	-9.54	1.23	6.50E-15	Y	0.91	intergenic
TOPBP1-TF	rs8177177	133463195	C	0.49	-0.21	1.18	8.61E-01	-6.74	1.20	1.95E-08	N	NA	intergenic
TOPBP1-TF	rs8177179	133463457	A	0.66	-6.88	1.29	9.62E-08	-11.12	1.23	1.78E-19	Y	0.92	intergenic
TOPBP1-TF	rs8177182	133464314	A	0.10	4.93	2.18	2.37E-02	11.54	2.09	3.19E-08	Y	0.80	intergenic
TF	rs1130459	133465283	G	0.66	-7.00	1.30	7.43E-08	-11.27	1.24	1.27E-19	Y	0.90	UTR-5
TF	rs148600419	133467945	T	0.10	5.11	2.23	2.21E-02	11.67	2.13	4.47E-08	Y	0.78	intronic
TF	rs8177235	133476083	A	0.06	-17.69	2.48	1.37E-12	-13.35	2.40	3.10E-08	N	NA	intronic
TF	rs8177237	133476421	G	0.35	0.62	1.32	6.41E-01	-7.49	1.27	3.28E-09	Y	0.86	intronic
TF	rs8177257	133480337	T	0.03	-26.19	3.76	3.33E-12	-19.73	3.60	4.06E-08	Y	0.82	intronic
TF	rs2715632	133485830	T	0.31	1.34	1.26	2.87E-01	9.48	1.29	2.64E-13	N	NA	intronic
TF	rs2718806	133486093	A	0.40	3.49	1.19	3.43E-03	7.15	1.14	3.59E-10	Y	0.97	intronic
TF	rs8649	133486958	C	0.31	1.57	1.25	2.10E-01	8.27	1.19	4.15E-12	Y	0.99	exonic
TF	rs1358022	133487621	G	0.31	1.51	1.26	2.31E-01	9.61	1.28	1.02E-13	N	NA	intronic
TF	rs1358021	133488877	C	0.54	12.21	1.21	6.62E-24	6.32	1.16	5.05E-08	Y	0.90	intronic

Table 2.10 Effect sizes and P values for top SNPs associated with log_ferritin (p<10^{-7}) in JHS before and after adjusting for menopause status.

					Before Adjustment			After Adjustment		
Chr	SNP	EA	EAF	RSQR	β	SE	*P*	β	SE	*P*
X	rs141555380	T	0.14	0.94	0.17	0.03	1.1E-08	0.16	0.03	1.4E-08
X	rs7885619	G	0.15	0.94	0.17	0.03	1.9E-08	0.16	0.03	2.1E-08
X	rs7063597	T	0.14	0.94	0.17	0.03	2.9E-08	0.16	0.03	3.1E-08
X	rs146474788	A	0.14	0.94	0.17	0.03	2.9E-08	0.16	0.03	3.2E-08
X	rs149621038	T	0.16	0.94	0.16	0.03	4.0E-08	0.15	0.03	5.7E-08
X	rs138941436	G	0.15	0.94	0.16	0.03	5.7E-08	0.15	0.03	8.8E-08
X	rs185814586	G	0.13	0.90	0.17	0.03	6.1E-08	0.17	0.03	6.5E-08
X	rs1050828	T	0.14	0.98	0.16	0.03	9.1E-08	0.15	0.03	9.3E-08
10	rs76969309	A	0.05	0.72	0.33	0.06	8.3E-08	0.32	0.06	7.6E-08

Note: menopause status was only available on 2136/2347 JHS

Table 2.11 Replication of signals established in prior GWAS of iron measures including Subjects of European descent

Trait	Chr	Index SNP Result in JHS [a]						Most Significant SNP in JHS [b]							Nearest Gene
		Index SNP	EA	EAF	β	SE	*P*	Most Significant SNP	EA	EAF	β	SE	*P*	LD r^2	
TIBC	3	rs3811647*(20)	A	0.22	17.55	1.43	9.7E-35	rs8177253	T	0.24	19.86	1.34	1.8E-47	1	*TF*
TIBC	3	rs1830084(20)	T	0.16	16.84	1.60	2.2E-25	rs1830084	T	0.16	16.84	1.60	2.2E-25	1	*TF*
log_ferritin	6	rs17342717*(82)	T	0.02	0.24	0.10	0.02	rs78273613*	G	0.02	0.26	0.10	0.01	0.88	*SLC17A1*
log_iron	6	rs1799945*(83)	G	0.04	0.08	0.03	7.8E-04	rs129128	C	0.03	0.09	0.02	1.0E-04	0.92	*HFE*
log_iron	6	rs1800562(20)	A	0.01	-0.03	0.04	0.37	rs1800562	A	0.01	-0.03	0.04	0.37	1	*HFE*
log_ferritin	6	rs1800562(82)	A	0.01	0.20	0.10	0.05	rs1800562	A	0.01	0.20	0.10	0.05	1	*HFE*
TIBC	6	rs1800562(20)	A	0.01	-24.63	5.13	1.7E-06	rs1800562	A	0.01	-24.63	5.13	1.7E-06	1	*HFE*
log_sat	6	rs1800562(20)	A	0.01	0.05	0.04	0.14	rs1800562	A	0.01	0.05	0.04	0.14	1	*HFE*
log_sat	6	rs13194491*(20)	T	0.02	0.02	0.05	0.75	rs35657082*	T	0.01	0.25	0.11	0.02	0.91	*HIST1H2BJ*
log_iron	22	rs855791*(22)	G	0.83	0.03	0.01	0.08	rs2072860*	A	0.73	0.03	0.01	7.7E-03	0.90	*TMPRSS6*
log_sat	22	rs855791*(22)	G	0.83	0.02	0.01	0.15	rs4820268*	A	0.73	0.02	0.01	0.03	0.90	*TMPRSS6*
log_iron	22	rs4820268*(83)	A	0.73	0.03	0.01	8.4E-03	rs9610638*	C	0.80	0.05	0.02	3.5E-03	0.83	*TMPRSS6*

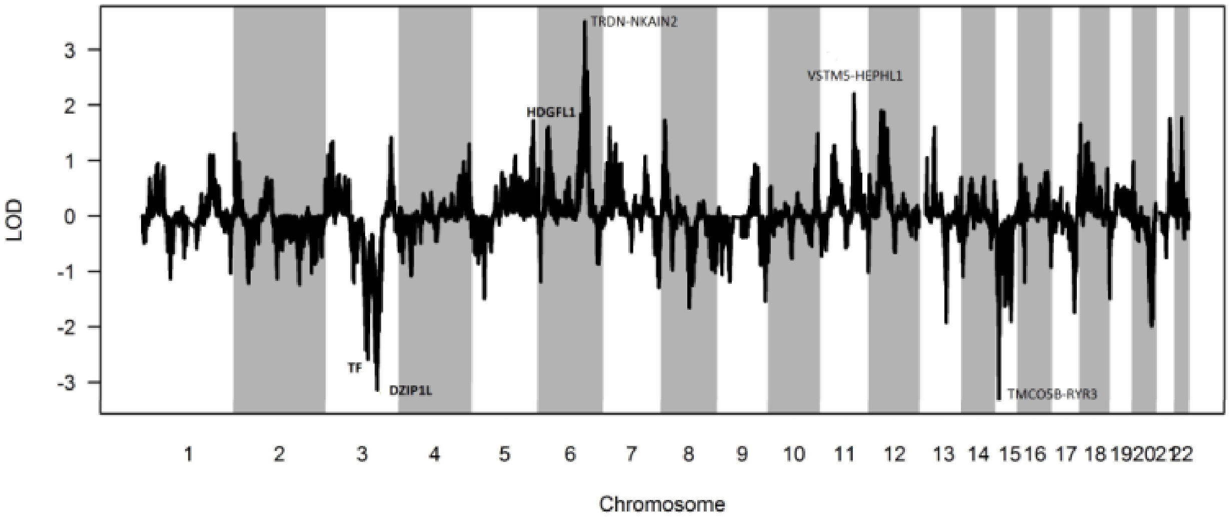
* Imputed;

( ) besides “Index SNP” contains the citation to the initial association study for each individual variant

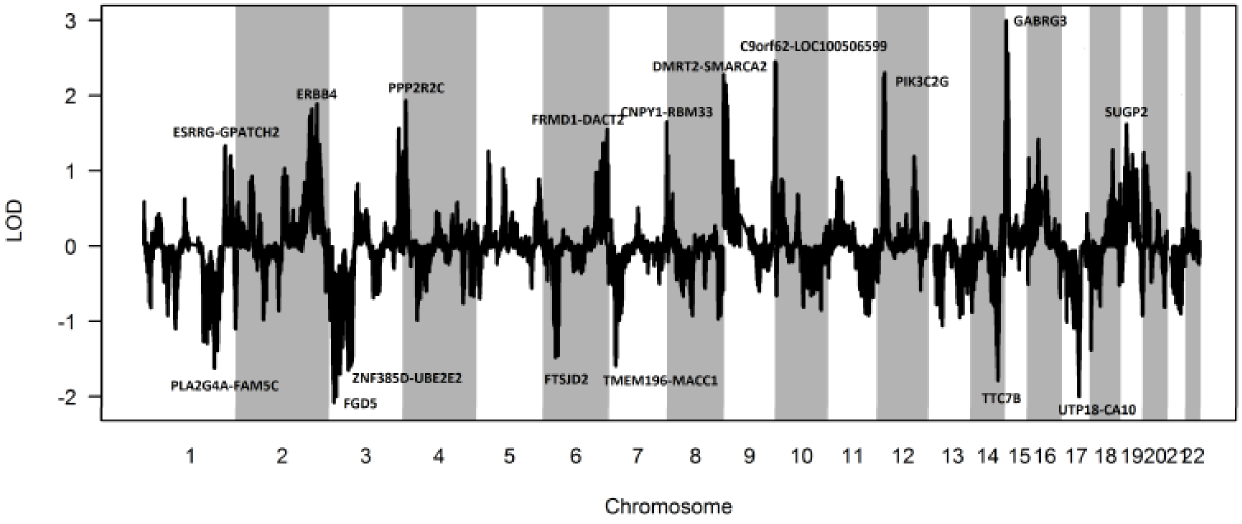
[a] Index SNP: Index SNP that was reported to be significantly ($P<5x10^{-8}$) associated with iron-traits in prior GWAS in European descents

[b] Most significant SNP: The proxy for index SNP (LD $r^2>0.8$ in CEU 1000G with the index SNP) with the smallest association P value in JHS

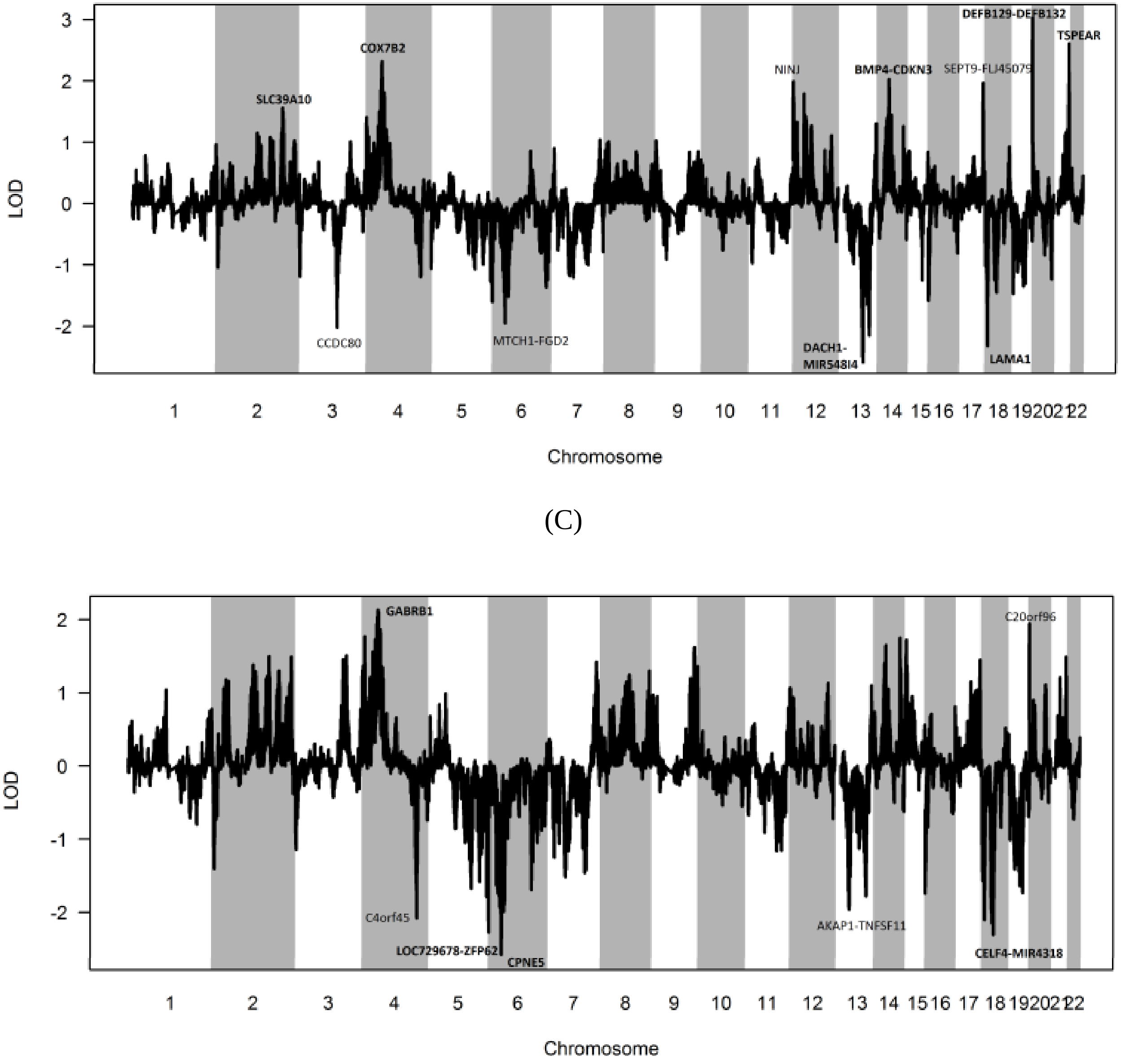
LD r^2 : r^2 with index SNP in CEU 1000G subjects


TRDN-NKAIN2
VSTM5-HEPHL1
HDGFL1
TF
DZIP1L
TMCO5B-RYR3
LOD
3
2
1
0
-1
-2
-3
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome

(A)


GABRG3
C9orf62-LOC100506599
DMRT2-SMARCA2
PIK3C2G
ERBB4
PPP2R2C
SUGP2
ESRRG-GPATCH2
FRMD1-DACT2
CNPY1-RBM33
PLA2G4A-FAM5C
ZNF385D-UBE2E2
FGD5
FTSJD2
TMEM196-MACC1
TTC7B
UTP18-CA10
LOD
3
2
1
0
-1
-2
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome

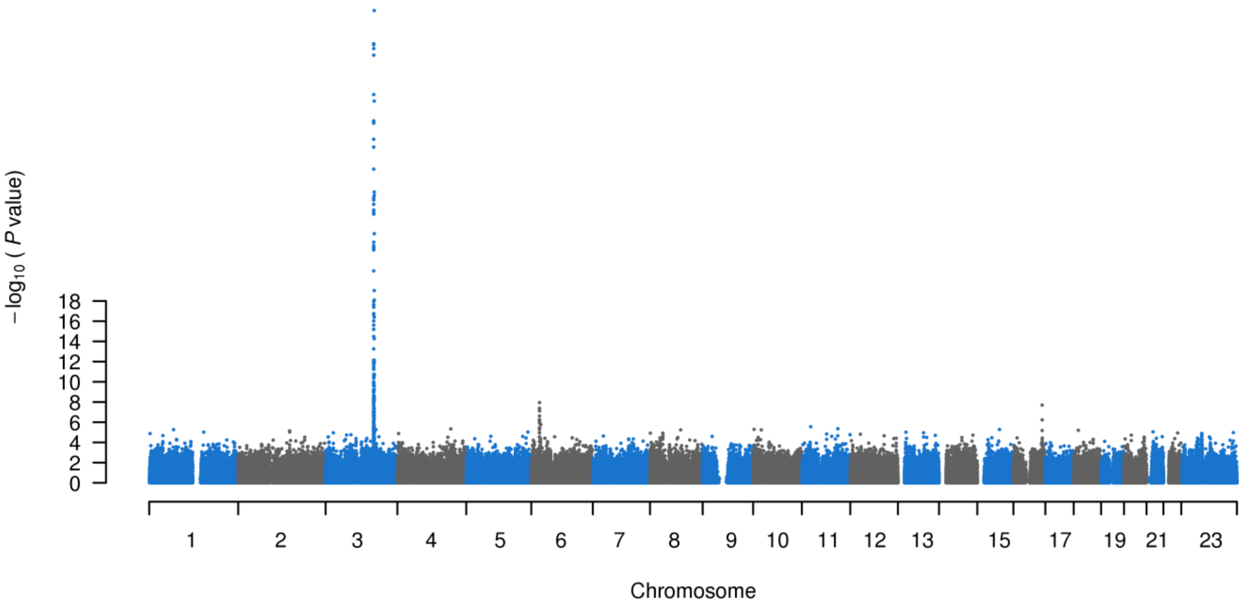
(B)



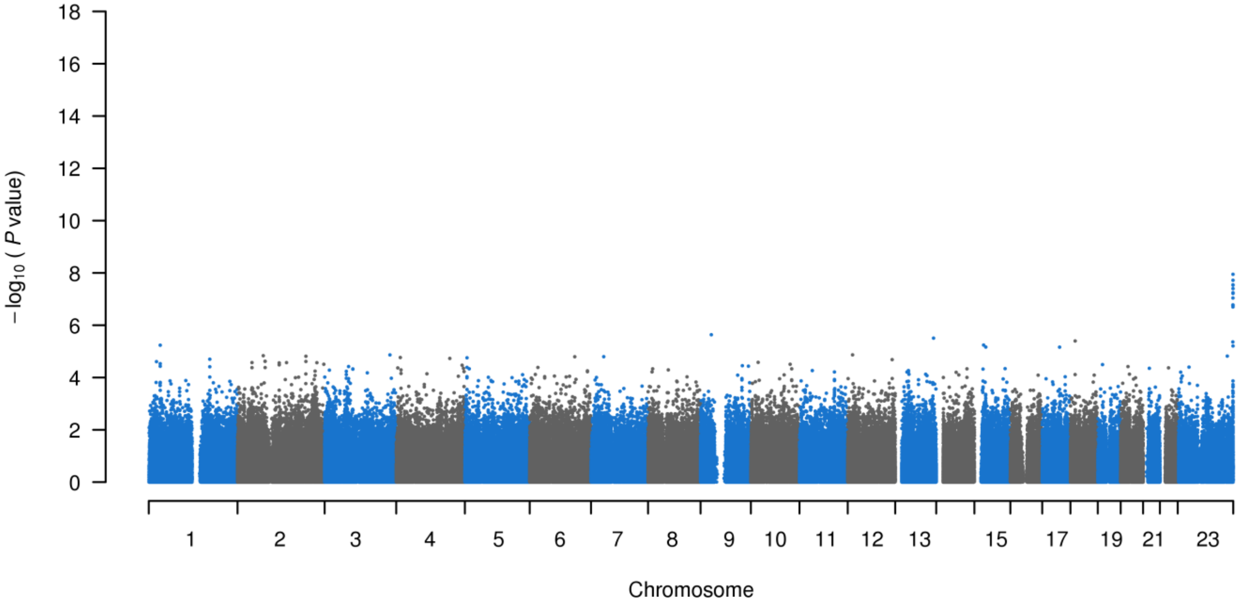
(C)

(D)

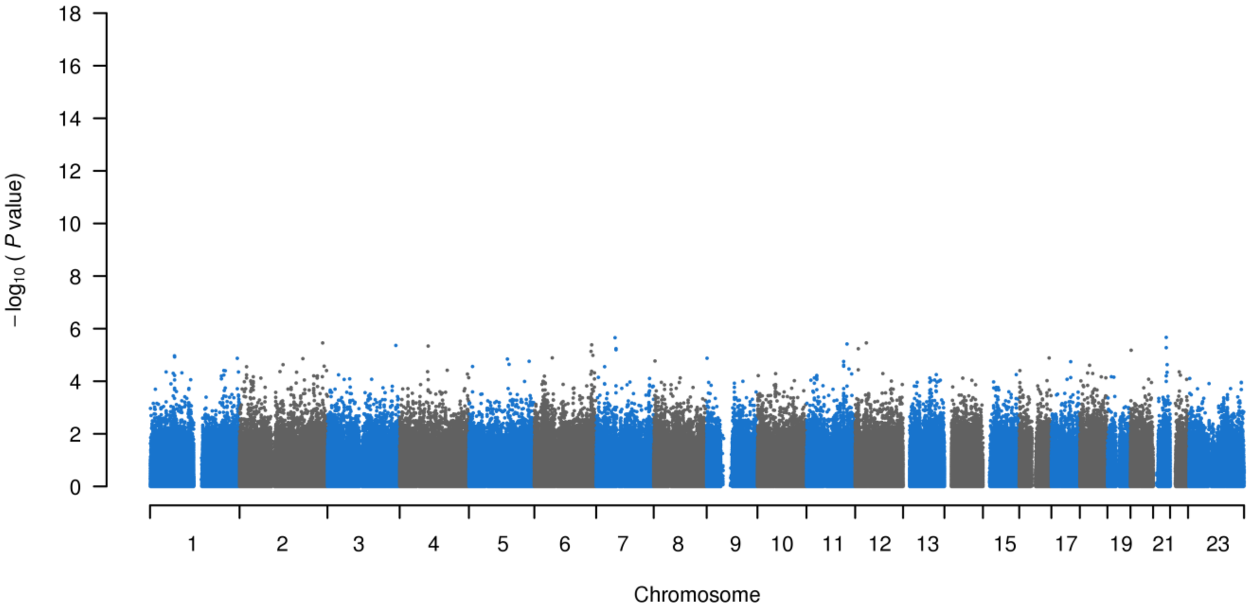
Figure 2.1 The LOD score of genome-wide admixture scan for iron-related phenotypes (A) TIBC, (B) log_ferritin, (C) log_iron, and (D) log_SAT. The LOD score is defined as the log base 10 ratio of the maximum likelihood of the data under a local-ancestry-associated disease model divided by the likelihood of the data under null model. Both alternative and null model include covariate adjustment for global ancestry. Positive LOD scores show the association of increased African ancestry with higher levels of iron measures, while negative LOD scores show the association of increased African ancestry with lower level of iron measures.


−log$_{10}$ (P value)
0
2
4
6
8
10
12
14
16
18
Chromosome
1
2
3
4
5
6
7
8
9
10
11
12
13
15
17
19
21
23

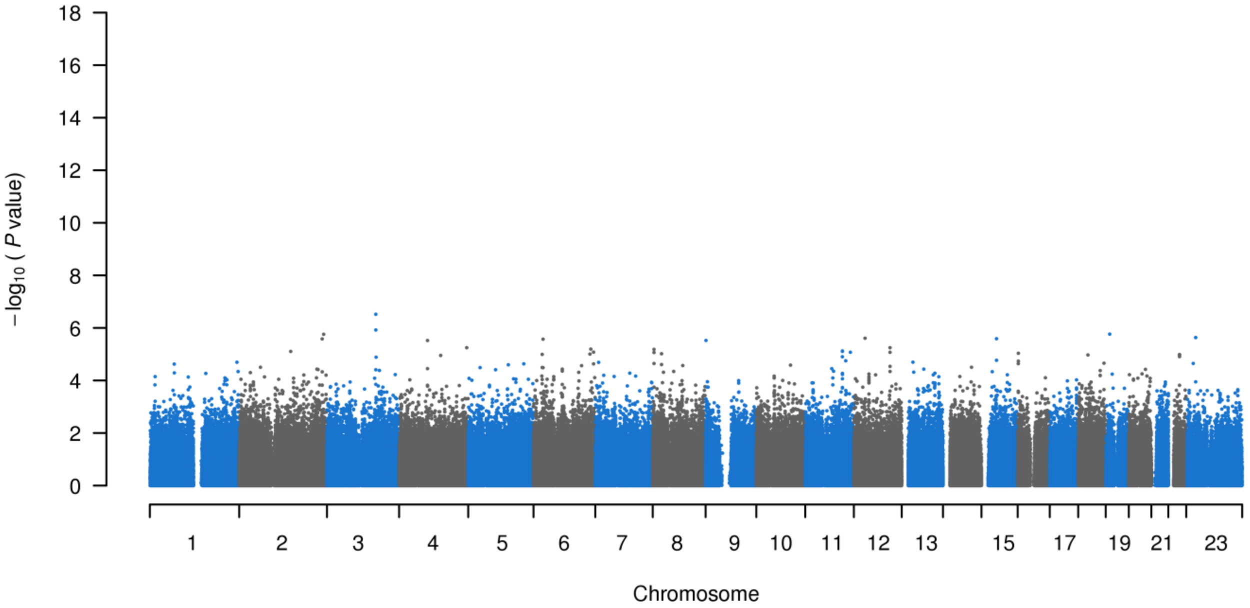
(A)


−log$_{10}$ (P value)
0
2
4
6
8
10
12
14
16
18
Chromosome
1
2
3
4
5
6
7
8
9
10
11
12
13
15
17
19
21
23

(B)

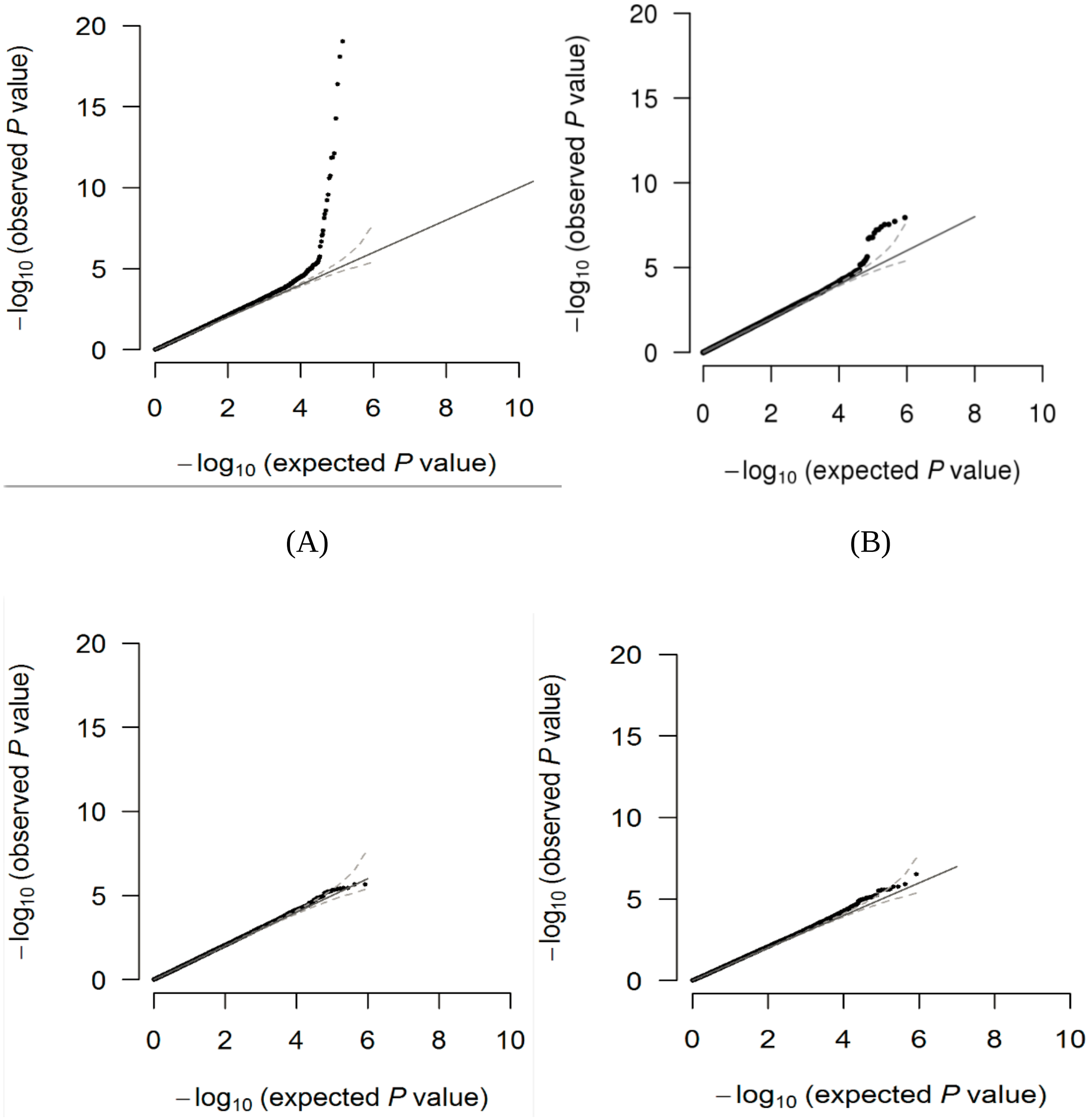


(C)



(D)

Figure 2.2 Manhattan plot of the -log$_{10}$(P) values by chromosome for iron-related phenotypes (A) TIBC, (B) log_ferritin, (C) log_iron, and (D) log_SAT.



(A) (B)

(C) (D)

Figure 2.3. Quantile-Quantile (Q-Q) plots of the P-values across all genotyped SNPs tested for association with iron-related phenotypes (A)TIBC, (B) log_ferritin, (C) log_iron, and (D) log_SAT in models adjusting for age, gender, BMI, and 10 eigenvectors calculated from principal component analysis. Horizontal and vertical lines represent expected P values under null distribution and observed P values, respectively. The straight line represents the expected distribution assuming no inflation of the statistics.

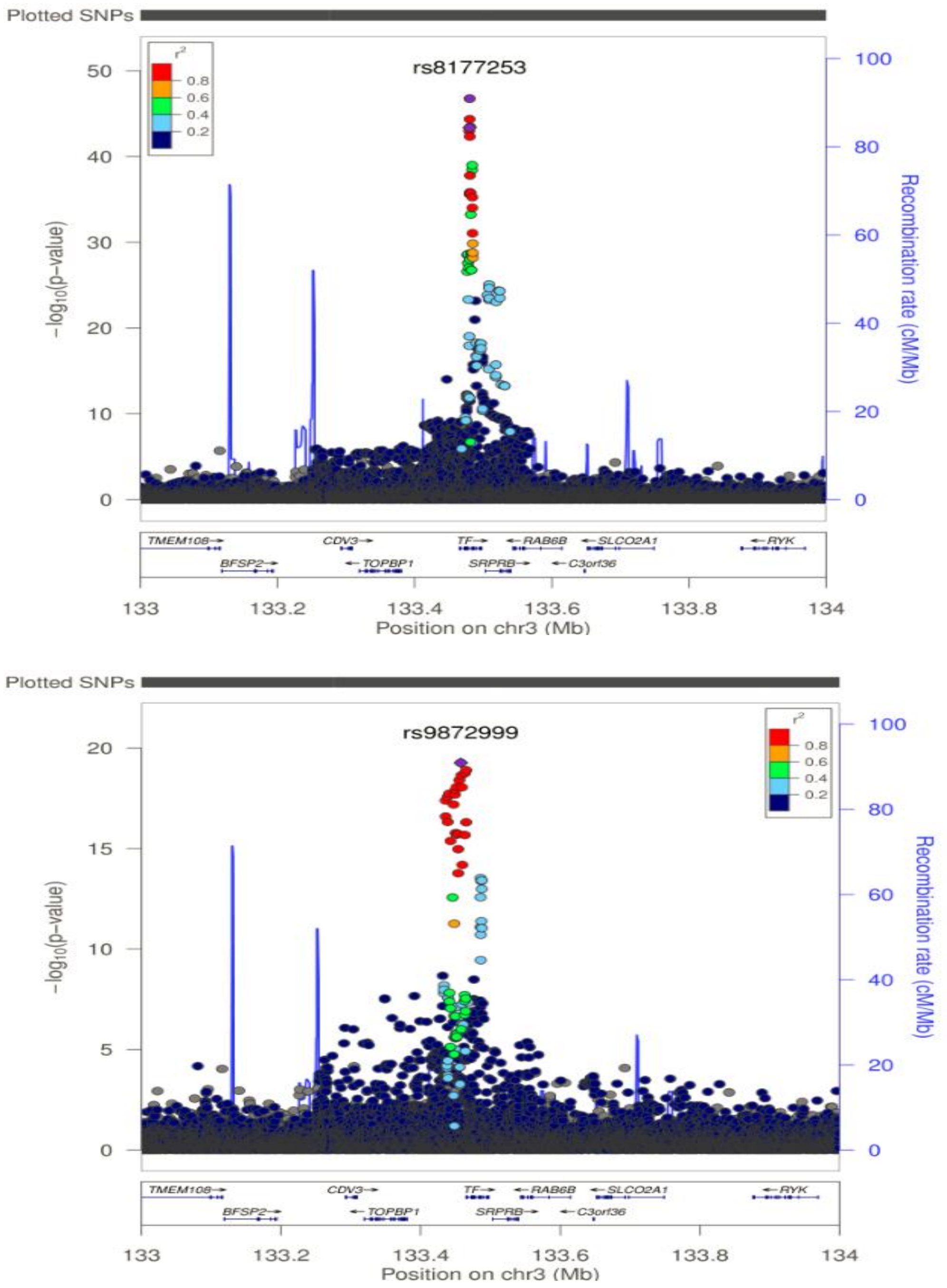


Figure 2.4 Regional plot of the -log$_{10}$(P) values for the SNPs in the *TF* region for TIBC before and after adjusting for the top SNP rs8177253 in this region (upper and lower panels, respectively). The X axis shows the human genome build 19 coordinates (Mb) and the genes in the region. The Y axis shows the -log$_{10}$ association P values of SNPs on the left, and recombination rates in cM per Mb on the right. Different colors of shading indicate the strength of linkage disequilibrium (LD) (r^2) between the top SNP and the other SNPs tested in the region.

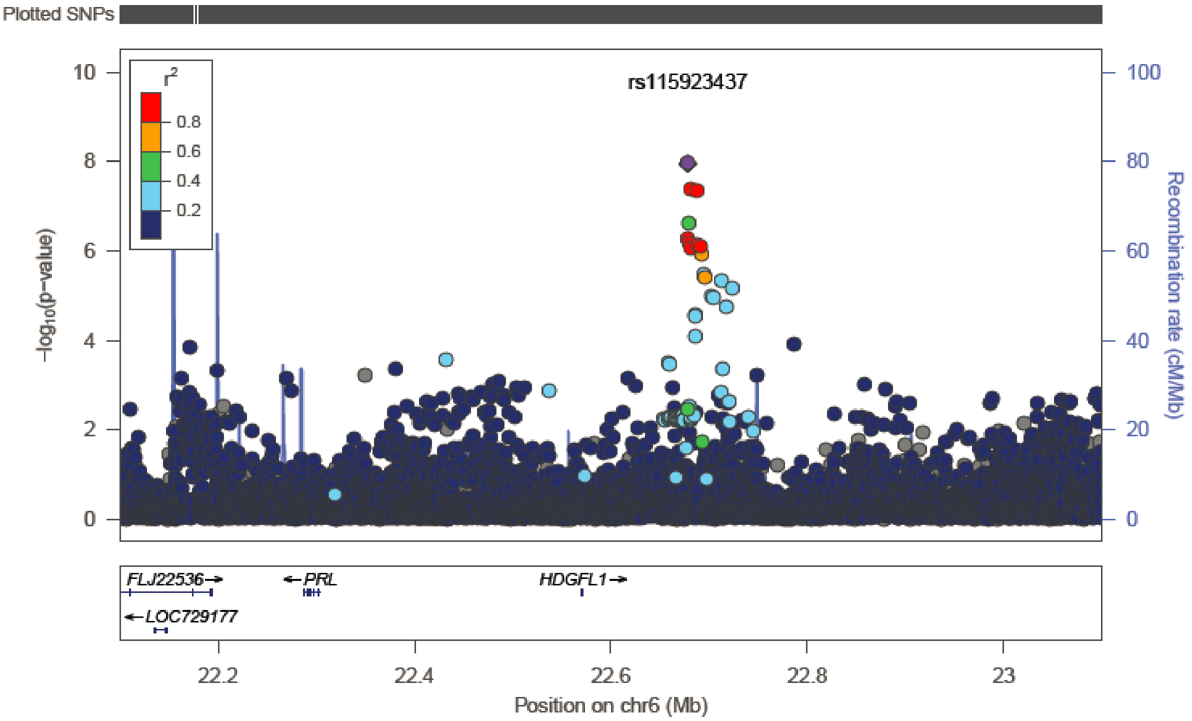


Figure 2.5 Regional plot of the -log$_{10}$(P) values for the SNPs at the *HDGFL1* risk locus for TIBC.

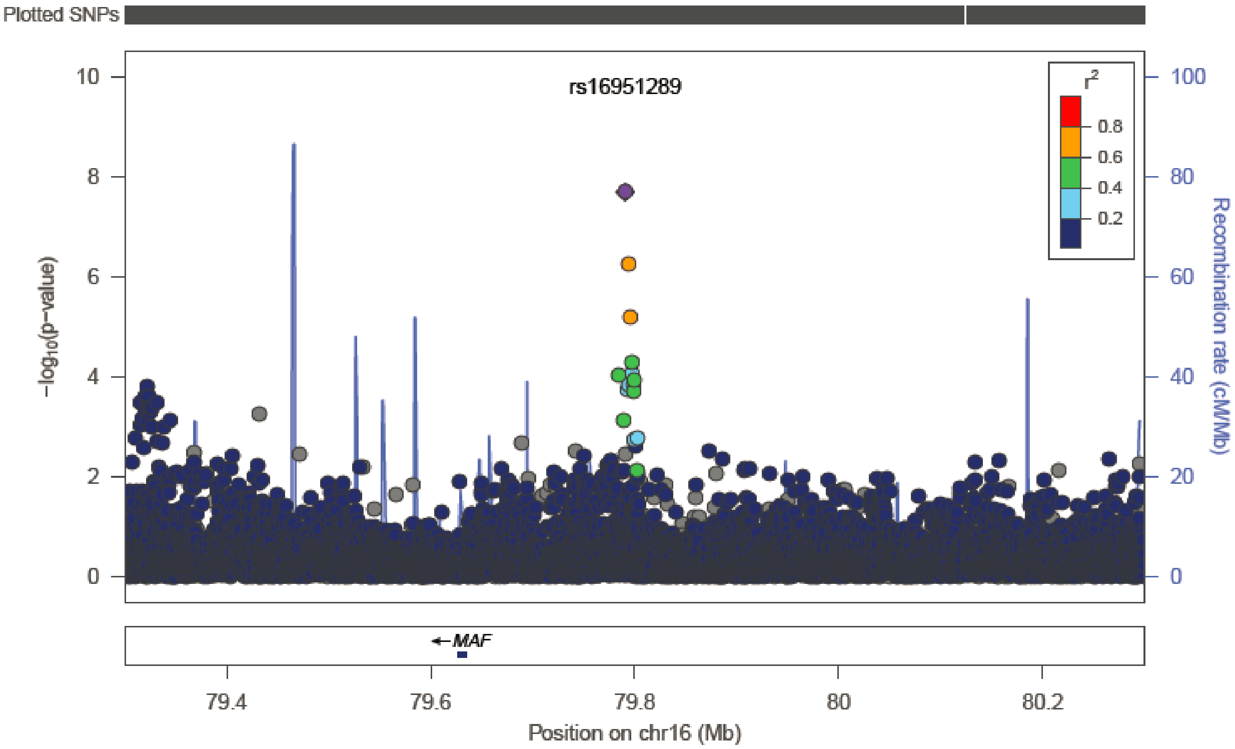


Figure 2.6 Regional plot of the -log$_{10}$(P) values for the SNPs at the *MAF* risk locus for TIBC.

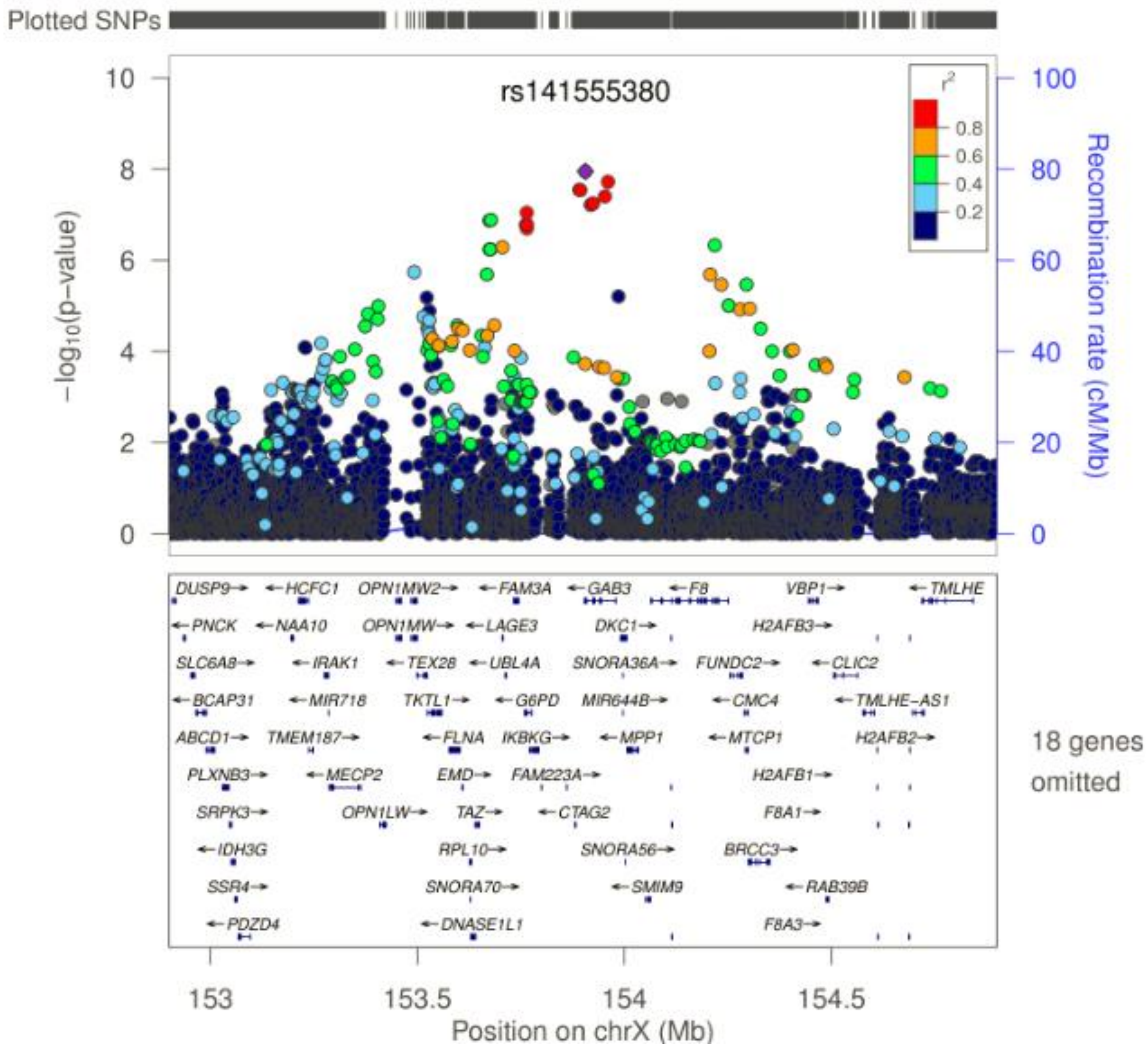


Figure 2.7 Regional plot of the -log$_{10}$(P) values for the SNPs in the *GAB3* region for log_ferritin.

CHAPTER III: GENOME-WIDE AND EXOME-WIDE ASSOCIATION STUDY OF SERUM LIPOPROTEIN (A) IN THE JACKSON HEART STUDY [2]

Introduction

Lipoprotein (a) [Lp(a)] is an independent risk factor for cardiovascular disease(27, 28). Genetic variants in *LPA* are strongly associated with both an increased level of Lp(a) and an increased risk of coronary disease(36), suggesting a causal role of Lp(a) in coronary disease. Lp(a) is a low-density lipoprotein (LDL)-like particle that consists of an apolipoprotein(a) covalently linked to apolipoprotein B100 by a disulfide bond(26). Genetic factors have a large impact on the variation of Lp(a) levels and approximately 70-90% of the total variance of Lp(a) can be attributed to variation within the *LPA* locus across worldwide populations(34, 35). In *LPA*, a copy-number variation (CNV) which encodes a Kringle(IV) type 2 domain and accounts for approximately half of variance explained by *LPA* locus(37, 38). Recent genome-wide association studies (GWAS) in subjects of European descent have identified multiple polymorphisms spanning 12.5 Mb on chromosome 6q26-27, which includes *LPA*, that are significantly associated with Lp(a) levels independent of each other and of the Kringle IV size polymorphism in *LPA* ($p<5x10^{-8}$)(37). A candidate gene study on multi-ethnic populations suggested both SNPs at 6q26-27 and the Kringle IV CNV were genomic determinants of Lp(a) level, and the proportion of total variance explained by each determinant differ across ethnic groups(39). Lp(a) levels in populations of African ancestry are much higher (2~4-fold) than in populations of European ancestry(111). A genome-wide admixture study on a population of African American

[2] This unpublished work is under review by the Journal of Human Genetics.

(AA) suggested local ancestry at 6q25.3 was significantly associated with Lp(a) after adjustment for the Kringle IV CNV(112). However, so far no genome-wide or exome-wide association studies of common and uncommon-coding variants, respectively, have been conducted in populations of African ancestry to assess the importance of other genomic regions on Lp(a) levels.

Here we present a genome-wide and an exome-wide association study of Lp(a) among AAs participating in the Jackson Heart Study (JHS). We observed numerous SNPs at the well-established *LPA* locus and a single SNP in *APOE* significantly associated with Lp(a) ($p<5x10^{-8}$). A high burden of coding variants in *LPA* and *APOE* were also associated with higher Lp(a) levels.

Materials and Methods

Study subjects and phenotypes

This study included 1106 AA male and 1789 AA female participants with measured Lp(a) levels and available genome-wide genotype data from the JHS, a longitudinal, population-based cohort from Jackson, MS(113). The design, recruitment and initial characterization of this study was described in details elsewhere(114). Serum Lp(a) levels (mg/dL) were measured using a Diasorin nephelometric assay on a Roche Cobas FARA analyzer (115). Fasting low-density lipoprotein (LDL), high-density lipoprotein (HDL), triglyceride (TG), and total cholesterol (TC) were measured as previously described(91). For each individual treated with lipid lowering therapies, the observed lipid value was multiplied by a correction factor (1.352 for LDL, 0.949 for HDL, 1.210 for TG, and 1.271 for TC)(116). The study protocol was approved by the University of Mississippi Medical Center Institutional Review Board, and written informed consent was obtained from all JHS participants. Descriptive characteristics of the JHS participants in this study were summarized in **Table 3.1**.

Genome-Wide Genotype Data and Genotype Imputation

A total of 3030 JHS participants were genotyped using the Affymetrix Genome-Wide Human SNP Array 6.0. Genotyping quality control was conducted using PLINK v1.07 (117). 874,712 SNPs with a call rate greater than 0.95 and minor allele frequency (MAF) greater than 0.01 were included in the genotype imputation target panel. Thirty-eight million SNPs were imputed, using MACH 2.0 (118), based on a reference panel consisting of the complete sample of 1000 Genome Project participants (Nov 2010, Version 3); only SNPs with imputation quality of r^2 greater than 0.3 were included in further analysis. The Kringle IV CNV was not available in JHS participants.

Exome Array

A total of 2,790 JHS participants, including 2,448 with Affymetrix 6.0 genotype data, were genotyped using the Illumina Human Exome Beadchip (version 12v1_rev5), consisting of >200,000 putative functional variants selected from >12,000 individual exome and whole-genome sequences across diverse populations and a range of common complex traits.

Statistical analyses

<u>Association of local ancestry estimates</u>

Ethnic differences in Lp(a) levels have been observed between subjects of European and African descent, with some populations of African Americans having, on average, almost 4-fold higher Lp(a) levels than European Americans(111). To assess the impact of genetic admixture within our AA population on Lp(a) levels, we first estimated the average, or global, African ancestry proportion across the genome for all subjects using the software ADMIXTURE (119) assuming K=2 contributing populations. We then tested whether this estimated proportion was associated with natural-log-transformed Lp(a) levels after adjusting for age, sex, and body mass index (BMI). Next, ANCESTRYMAP (72) was used to estimate local ancestry (probabilities of whether an individual has 0, 1, or 2 alleles of Caucasian ancestry) at 738,831 autosomal SNPs across the genome for each participant in JHS as previously described (91). In brief, local ancestry was inferred using hidden Markov models based on the genotypes from a panel of densely spaced markers with highly differential allele frequencies between African and European populations. We performed admixture mapping across the whole genome by regressing natural-log-transformed Lp(a) levels, adjusting for age, sex, BMI and global ancestry, on the local ancestry estimate at each SNP location. The LOD score for association, defined as the log-base-10 ratio of the likelihood of the data under a model including local ancestry divided by the likelihood of the data under the model excluding local ancestry, was computed at each of these local ancestry informative marker locations across the genome. For regions showing association of increased African ancestry with higher levels of Lp(a), the LOD scores were assigned positive values, and for regions showing association of increased African ancestry with lower levels of Lp(a), the LOD scores were assigned negative values. LOD scores were plotted across the whole genome, and a LOD score of 5 was assumed to be the threshold of statistical significance(72).

<u>Genome-wide association analysis</u>

Lp(a) levels were natural log-transformed to approximate normality of residuals after accounting for age, sex and BMI. The association between Lp(a) and imputed SNPs were tested using

112.w1qqmultivariable linear regression models in MACH2QTL v.1.08 (118), adjusting for age, sex, BMI and the first 10 principal components generated from EIGENSOFT (120) based on a linkage disequilibrium pruned set of SNPs with minor allele frequency (MAF) > 0.05. An additive mode-of-inheritance model was assumed for genotype; β coefficients, representing the estimated change in transformed trait value associated with each additional copy of the effect allele, and there corresponding standard errors were reported

Manhattan plots were made to illustrate the association results across the genome. Quantile-quantile (Q-Q) plots of observed versus expected -$\log_{10}$(P-values) were made to assess any systematic inflation of the regression test statistics across the genome before and after removing the SNPs in the 6q25.3 region widely reported to be significantly associated with Lp(a). The genomic inflation factor (λ), defined as observed median value of the chi-squared statistic divided by 0.456, was calculated excluding 6q25.3 SNP results. The observed chi-square statistics were divided by λ to obtain the corrected chi-square statistics and corrected p-values. In regions with significant evidence for association, multivariable "conditional" regression models that included the imputed genotype data of the most strongly associated SNP as an additional covariate were performed to assess the evidence for multiple independently associated SNPs in the region. If a second signal also reached genome-wide significance after conditioning on the top variant, multivariable regression models were repeated to include the genotypes of both SNPs as covariates. Region-specific plots were made to show the magnitude of association between all SNPs and Lp(a) levels as well as the estimated linkage disequilibrium (LD) between each SNP in the region and the most strongly associated SNP. Finally, in order to control our association results for possible confounding due to ancestry, for any observed associated SNP we identified the genetic position of the most strongly associated SNP in the associated region, selected the local ancestry estimate at the location closest to that SNP, and performed multivariable regression models as described above but now including estimated local ancestry proportion as an additional covariate.

Identifying the most important 6q associated SNPs using LASSO-based resample model averaging

In order to identify the SNPs that most likely contribute to the observed association at 6q, we applied the LASSO local automatic regularization resample model averaging (LLARRMA) method, a method that combines LASSO variable shrinkage and selection with resample model averaging and multiple imputation (121), to estimate the probability of each SNP to be included in a multi-SNP model that best explains the Lp(a) outcome across alternative realizations of the data. We first extracted the genotypes for SNPs on the Affy 6.0 array that mapped within the 1 Mb region centered around the top genotyped SNP (rs9457986) identified in the initial GWAS scan. Then we used fastPHASE (122) to impute the missing genotypes for SNPs in this region which failed genotyping on Affy 6.0. Finally, we fitted the LASSO models by regressing the residuals of Lp(a) adjusting for age, gender, BMI, 10 principle components and the local ancestry estimate of rs9457986 on the genotypes using the LLARRMA package in R, and calculated a resample model inclusion probability (RMIP) score for each SNP in this 1Mb region on chromosome 6. Linkage disequilibrium (LD) statistics (R^2 and D' based on 1000G YRI subjects) between variants with RMIP >0.75 were calculated using Gold(123) and plotted using Haploview(124).

Haplotype analyses

Haplotype analyses were conducted using the 'haplo.stats' R package(125), to examine specific combinations of allelic variants and whether the observed association signal is likely attributable to a less common unmeasured genetic variant. Haplotypes were constructed among the five genotyped variants in the *LPA* gene with RMIP score >0.75. The 'haplo.glm' function implemented in the 'haplo.stats' R package was used to calculate effect coefficients (β), standard errors (SE) and P-values for each haplotype relative to the most common reference haplotype. The 'haplo.score' function was used to calculate the global score statistic to test the overall association between haplotypes and log-transformed Lp(a). The same set of covariates used in the genotype analyses were used in the haplotype analyses.

Single variant and gene-based analysis of SNPs on Exome Array

Single variant and multivariant, gene-set or "gene-burden", association analyses were performed for variants appearing on the Exome Array. Gene-burden tests performed included the Madsen-Browning test(126) and the SKAT-Optimal (127) test, which picks the 'best' combination of the SKAT(128) and a Madsen-Browning(126) test for gene-based testing. We performed three levels of gene-based analyses: (Level 1) the combination of stop-loss, stop-gain, and splice-site regardless of MAF; (Level 2) the combination of SNPs in Level 1 and all variants which are predicted to be "damaging" using PolyPhen(129) and SIFT(130) regardless of MAF; (Level 3) the combination of stop-loss, stop-gain, splice-site and non-synonymous variants with MAF upper limits of 3%.

To investigate how much common or rare SNPs associated with Lp(a) in the single-variant analysis could account for gene-based test, conditional analysis was performed by including the allele count at these lead SNPs as covariates. In order to control for the possible confounding of Lp(a) association results due to an association between both SNP and Lp(a) with other lipid traits (HDL, LDL, TG and TC), we performed general linear mixed model regression as described above but now including the lipid traits an additional covariate.

Statistical significance

A significance threshold of $P < 5\times10^{-8}$ was used to define genome-wide significance for all individual SNP results (both GWAS and Exome Array). Test statistics for all individual SNP results, both GWAS and Exome-Chip, were adjusted by the genome inflation factor (λ) prior to calculating significance (p-values) to account for any possible systematic bias in results. A gene-based association result was defined to be significant if P<0.05/number of genes. The number of genes for Level 1, Level 2, and Level 3 inclusion criteria were 4752, 13658, and 15963, resulting in significance thresholds of 1.1 $\text{x}10^{-5}$, $3.7\text{x}10^{-6}$, and $3.1\text{x}10^{-6}$, respectively.

Results

Descriptive statistics of the JHS participants in this study were summarized in **Table 3.1**.

Admixture mapping for determinants of Lp(a)

Consistent with observed higher levels of Lp(a) in AAs versus EAs, higher levels of estimated global, or average, African ancestry was significantly associated with higher levels of log-transformed Lp(a) (β=0.76, p=$1.3\text{x}10^{-9}$). The estimated global proportion of African ancestry obtained from ADMIXTURE was highly correlated with the first principal component from EIGENSTRAT (correlation=0.998). Admixture mapping showed a highly significant association between increased African ancestry at chromosome 6q25.3 and increased Lp(a) levels (**Figure 3.1**) after adjusting for the global proportion of African ancestry. The estimated global proportion of African ancestry became non-significant after including local ancestry at 6q25.3. SNP rs505000, upstream of *SLC22A3* was the local ancestry informative SNP most strongly associated with Lp(a) levels (p= $8.2\text{x}10^{-27}$, LOD=24.95).

GWAS SNPs associated with Lp(a) at $P<5x10^{-8}$

Q-Q plots revealed evidence for systematic inflated association results likely due to left-censoring of the Lp(a) measures in a subset of JHS participants. After excluding all the SNPs on chromosome 6, the distribution of the remaining P values across the genome still demonstrated inflation. After controlling for the genomic inflation factor (λ=1.14), Q-Q plots revealed no substantial evidence for inflation (**Figure 3.2**). The genomic inflation factor was applied to all individual SNP results. Only chromosome 6q region reached genome-wide significance for Lp(a) levels **(Figure 3.3)** after adjusting for the inflation factor**.** Overall, 804 SNPs reached genome-wide significance ($p<5x10^{-8}$) (all on chromosome 6), and the top SNPs that reached genome-wide significance are listed in **Table 3.2.**

Lp(a) GWAS Results on Chromosome 6

All 804 significant SNPs on chromosome 6 mapped to the 6q region, spanning from 153,917,144 to 163,745,411bp and containing more than 10 genes (**Figure 3.4**). The strongest signal (rs115848955, $p=3.1x10^{-55}$, MAF=0.05) mapped to the *LPA* gene. *LPA* gene encodes a modified form of low density lipoprotein, in which a large glycoprotein (Apo(a)) is covalently bound to apolipoprotein B by a disulfide bridge(131), and structurally, the Apo(a) chain contains a region homologous with plasminogen, which gives Lp(a) anti-fibrinolysis activity by competing with plasminogen's binding to fibrin. After adjusting for the top SNP (rs115848955), the top associated variant was rs9355814 ($p=7.8x10^{-21}$) (**Figure 3.4**) and 469 SNPs in the 6q region spanning from 159,092,125- 163,745,411 bp remained genome-wide significant. However, after adjusting for the local ancestry informative marker closest to rs115848955 (i.e. rs6923917), only 406 SNPs in a relatively narrow region spanning from 160,633,560 to 161,342,219 bp remained significant ($p<5x10^{-8}$). The top SNP in the conditional analysis including local ancestry was rs138429428 ($p=4.2x10^{-50}$, **Figure 3.4**). The local ancestry estimate closest to rs115848955 explained an estimated 4.4% of the total variation of Lp(a) after accounting for age, gender, BMI and the first 10 PCs. SNP rs115848955 explained an estimated 8.8% of the total variation of Lp(a) after accounting for age, gender, BMI, the first 10 PCs, and the local ancestry estimate nearest this location.

Evidence for multiple associated SNPs in the *LPA* region on chromosome 6

Fifteen directly genotyped SNPs in the 1Mb region surrounding the top associated SNP (rs9457986) on chromosome 6 were identified as having stable, independent associations with Lp(a) (RMIP score >0.75) in multivariant models using LLARRMA, as shown in **Table 3.3**. The LD estimates between these SNPs were not high, as shown in **Figure 3.5**. Five of the SNPs (rs6415084 [RMIP=0.93], rs3798221 [RMIP=1], rs9457986 [RMIP=1], rs1367211 [RMIP=0.97], and rs1406888 [RMIP=1]) were in the *LPA* gene. The five Lp(a) SNPs together explained an estimated 7.3% of the total variance of Lp(a) after accounting for age, gender, BMI, the first 10 PCs, and local ancestry estimate at rs9457986. These 5 SNPs were used to construct haplotypes and to estimate the effect on Lp(a) levels for each additional copy of a particular haplotype, compared to the reference haplotype (**Table 3.4**). There was significant evidence for an overall association between haplotypes and Lp(a) (global $P=1.12x10^{-55}$) ; 6 individual haplotypes were minimally nominally associated ($p < 0.05$) with Lp(a) levels.

Single variant and gene-based analysis for Exome Array

Seven SNPs on the Exome Array were associated ($p<5x10^{-8}$) with Lp(a) (**Table 3.5**) after adjustment for the genomic inflation factor, six of them were in the chromosome 6q *LPA* region, and the other was the widely reported chromosome 19 *APOE* SNP rs7412. Five of the six associated chromosome 6q SNPs remained genome-wide significant after adjusting for top

GWAS SNP rs115848955 (**Table 3.5**). Five of the six SNPs (excepting rs41272114) were intronic or in intergenic regions. Of the seven SNPs, only rs7412 was not significant in the 1000G imputed GWAS results. Rs7412 (p=3.3x10^{-8}) was poorly imputed based on 1000G data and was not included in the original GWAS analyses. In the gene-based analyses, the *LPA* gene reached exome-wide significance for all three levels of SNP inclusion when using the Madsen-Browning test, as well as for Level 2 and Level 3 SNPs in the SKAT-O test. The *APOE* gene reached exome-wide significance for Level 2 SNPs using the Madsen-Browning test (**Table 3.6**).

Uncommon functional variants in *LPA* and *APOE*

All splice-site, stop altering and non-synonymous SNPs in *LPA* and *APOE* on the Exome Array are listed in **Table 3.7** and **Table 3.8**, respectively. The most significant individual SNP in *LPA* on the Exome Array was rs41272114 (p=6.5x10^{-12}, MAF=0.01), which is a splicing-altering variant. *LPA* SNPs were significantly associated with Lp(a) levels in the gene-burden tests (p=1.2x10^{-21} for Level 3 SNPs using Madsen-Browning test). The Level 3 gene-based Madsen-Browning test remained significant after removing the top two individual SNP results (i.e. rs41272114 and rs41272110) (**Table 3.9**). Only three *APOE* SNPs, including the aforementioned rs7412, were successfully genotyped and informative on the Exome Array. A second nonsynonymous *APOE* SNP (rs769455, MAF=0.02) that is specific to populations of African descent, demonstrated strong nominal results that were independent of rs7412 (p=2.0x10^{-5} before adjustment for rs7412 and p = 3.5x10^{-5}after adjustment for rs7412). Together, rs7412 and rs769455 explain ~2.3% of the total variation of Lp(a) after adjusting for age, gender, BMI and the first 10 PCs.

Since *APOE* is a strong risk factor for lipid traits (LDL and TC), we next investigated whether the association between *APOE* and Lp(a) is attenuated when adjusting for lipid traits.

We re-tested the association between signal at *APOE* and Lp(a) adjusting for the untreated LDL and TC levels, and found that the evidence for association decreased, but the signal did not totally disappear (**Table 3.8**, P value for rs7412 dropped from 3.2x10^{-8} to 7.6x10^{-4}).

Discussion

We conducted genome-wide and exome-wide association studies for Lp(a) in 2,896 AA participating in the Jackson Heart Study. Higher level of estimated global African ancestry was significantly associated with higher level of Lp(a), and this association of global ancestry was largely explained by the association between Lp(a) and local ancestry in the chromosome 6q25.3 region. We observed significant (P<5x10^{-8}) associations for hundreds of SNPs spanning ~10Mb region on 6q surrounding the *LPA* gene. Interestingly, after adjusting for local ancestry, the region containing significantly associated SNPs got much narrower and was centered over the *LPA* gene (<1Mb). Significant haplotypic effects were also detected in the *LPA* region that implicates numerous causal variants. A single *APOE* SNP, rs7412 on the Exome Array also reached genome-wide significance. Gene-burden tests found significant associations between Lp(a) and aggregate collections of SNPs in *LPA* and *APOE*.

Previously, Deo et al. performed a targeted study of the Lp(a) region, using haplotype tagging SNPs, in 4,464 JHS participants and 1,726 AA participants from the Dallas Heart Study.[11] This study also performed a genome-wide admixture analyses based on a panel of 1,447 ancestry informative markers including subjects in the upper and lower quintile of the Lp(a) distribution. Herein, we dramatically expand the coverage of both common and rare variants across the entire *LPA* region; estimate local genetic admixture at 738,831 autosomal SNP locations and perform admixture mapping including all subjects with Lp(a) measures. Eight-hundred-four SNPs in the

6q region (spanning ~10 Mb) were significantly associated with Lp(a) levels ($P<5x10^{-8}$). After adjusting for the most strongly associated SNP rs115848955, multiple signals at 6q region spanning ~5 Mb remained significantly associated with Lp(a). After adjusting for the local ancestry at 6q25.3, the region that harbors SNPs significantly associated with Lp(a) ($p<5x10^{-8}$) became much narrower (from 9.8Mb to 0.7Mb) and was centered around the three genes *SLC22A*, *LPL2* and *LPA*. This result suggests confounding between local ancestry and SNPs spanning the larger 6q region identified to be associated with Lp(a). Given the relatively recent admixture in the African American population, local ancestry can confound associations across a relatively large region surrounding the population-specific, or population-enriched, causal variant(s)(132, 133). The observation that the associations in and near *LPA* remains robust after adjustment for local ancestry at *LPA* while the evidence for association further away dramatically declines suggests that the ancestry-specific (or highly-enriched) causal risk variant(s) resides in or near *LPA* and that most, if not all, of the observed associations outside this narrower region are spurious associations. Interestingly, a similar extended region of association with Lp(a) surrounding *LPA* has been observed in more homogeneous European populations. An obvious candidate to explain some of the differences in association results between European and African populations is the Kringle IV polymorphism, which has not been measured in JHS participants. Deo et al. demonstrated that the Kringle IV polymorphism explains some of the ancestry effect differences, but noted that several associated common SNPs, with strong allele frequency differences between populations of African and European ancestry, in and around *LPA* explain the majority of the population differences in Lp(a) levels.[11]

A common non-synonymous variant at *APOE* on Exome Array, rs7412, was identified to be significantly associated with Lp(a) in single variant analysis (MAF=0.11, leading to an Arg to

Cyc substitution, p=3.2x10^{-8}). Another low-frequency non-synonymous variant at *APOE,* rs769455, was also nominally associated with Lp(a) (MAF=0.02, leading to an Arg to Cyc substitution, p=2.0x10^{-5}). Prior studies have investigated the relationship between *APOE* genotypes and Lp(a) levels, but the results were inconsistent. Some studies reported no impact of *APOE* genotypes on Lp(a) levels(134-138), while others reported significant associations between them(139-144). A recent study found that among African Americans, lower Lp(a) levels were observed in *APOE* ε2 carriers, and this association was only observed in subjects with large apoA size (defined as >26 Kringle IV repeats) but not in the subjects with small apoA size(145). Another study on Caucasian males also reported that the effect of *APOE* on Lp(a) levels was only observed in subjects of largest quartile of apoA size, but with lower Lp(a) levels for *APOE* *ε*4 carriers(141). In our study, *APOE* ε2 genotype was associated with lower Lp(a) levels, which is consistent with prior study on African Americans.

In summary, we observed that local ancestry at 6q25.3 was an important risk factor for Lp(a) in AA, and that SNPs at the well-established *LPA* locus were significantly associated with Lp(a) (p<5x10^{-8}) after adjusting for the local ancestry at 6q25.3. Prior to covariate adjustment for local ancestry at 6q25.3, the observed region containing associated SNPs spanned ~10 Mb. After covariate adjustment the associated region was only 700 kb. We also observed a significant association for a non-synonymous variant in *APOE.* Future large multi-ethnic studies which include high-coverage sequence data, and the Kringle IV polymorphism, would be ideally suited to better understand the complex genetic architecture of the *LPA* region that leads to strong population differences in Lp(a) levels.

Conflicts of Interest

The authors declare no conflict of interests.

Table 3.1 Descriptive statistics of JHS participants for the study of Lp(a)

	Total	Male	female
N	2896	1107	1789
Age(years)	54.4 ± 12.9	53.8 ± 13.0	54.8 ± 12.8
BMI	32.1 ± 7.5	30.0 ± 6.3	33.3 ± 7.9
Lp(a) (mg/dL)	47 (25, 80)	42 (23,74)	50 (26, 84)
LDL-c (mg/dL)[1]	128 (103, 153)	130 (106, 155)	127 (101, 151)
HDL-c (mg/dL)[2]	49 (40, 59)	43 (37, 51)	52 (44, 62)
TG (mg/dL)[3]	93 (66, 131)	98 (70, 139)	89 (62, 125)
TC (mg/dL)[4]	199 (175, 229)	198 (175, 228)	200 (175, 230)

Note:
Data are mean ± SE, median (25th, 75th percentiles); lipid levels are untreated;
1: LDL-c, low-density lipoprotein cholesterol;
2: HDL-c, high-density lipoprotein cholesterol;
3: TG, triglycerides;
4: TC, total cholesterol

Table 3.2 Top SNPs that significantly ($p<5x10^{-8}$) associated with Lp(a) after controlling for genomic inflation factor

Chr[1]	# of SNPs[2]	Most significant SNP	Pos(hg19)[3]	EA[4]	EAF[5]	RSQR[6]	β[7]	SE[8]	*P*[1]	Nearest Gene	Function
6	5	rs9322428	153917946	A	0.85	0.95	0.19	0.03	1.3E-09	*RGS17-OPRM1*	intergenic
6	6	rs17539620	154896235	T	0.05	0.85	-0.33	0.05	3.7E-10	*CNKSR3-SCAF8-TIAM2*	intergenic
6	24	6:156789295:T_TAC	156789295	I	0.16	0.74	-0.22	0.03	4.7E-11	*NOX3-ARID1B*	intergenic
6	5	rs6909229	158572379	C	0.07	0.92	-0.27	0.04	3.7E-10	*SYNJ2-SERAC1-GTF2H5-TULP4*	intronic
6	19	6:159106232:G_GTC	159106232	I	0.60	0.95	0.18	0.02	4.8E-15	*SYTL3*	intronic
6	41	rs926657	159463452	T	0.42	0.99	0.14	0.02	4.6E-11	*C6orf99-RSPH3-TAGAP-FNDC1-SOD2-PNLDC1-MAS1-IGF2R*	intergenic
6	261	rs149565105	160878078	A	0.02	0.97	0.78	0.06	1.7E-34	*SLC22A1-SLC22A2-SLC22A3*	intronic
6	55	rs185414370	160889898	C	0.03	0.92	0.82	0.06	2.9E-40	*LPL2*	ncRNA_intronic
6	177	rs115848955	161031660	T	0.05	0.90	0.79	0.05	1.3E-62	*LPA*	intronic
6	130	rs144788267	161181875	A	0.03	0.75	0.83	0.07	4.6E-30	*PLG*	intergenic
6	25	rs142799378	161305763	G	0.02	0.85	0.81	0.08	6.8E-23	*MAP3K4*	intergenic
6	24	rs3757037	161697400	G	0.66	0.68	0.20	0.03	3.3E-13	*AGPAT4*	intergenic
6	27	rs7769089	162147727	T	0.67	1.00	0.16	0.02	4.6E-12	*PARK2*	intronic
6	5	rs6927207	163740089	A	0.22	0.75	-0.19	0.03	1.3E-11	*PACRG-AS1*	ncRNA_intronic

1: Chr, chromosome ;2: # of SNPs, number of SNPs that reached genome-wide significance at each locus; 3: Pos(hg19), physical position of the SNP according to human genome build version 19;
4: EA, effect allele; 5: EAF, effect allele frequency; 6: RSQR, RSQR represents the imputation quality provided by MACH. SNPs with RSQR<0.3 were excluded from analyses;
7: β, β coefficients representing the estimated change in the log-Lp(a) level associated with each additional copy of the effect allele; 8: SE, standard error; 9: *P*, p-value after genomic control
Models were adjusted for age, gender, BMI, and 10 PCs.

Table 3.3 LD statistics of 15 SNPs in the 1Mb hit region on chromosome 6 with RMIP>0.75 for Lp(a)

Gene	RMIP score	SNP	rs7757997	rs3850659	rs377551	rs2457576	rs7754188	rs6415084	rs3798221	rs9457986	rs1367211	rs1406888	rs9458005	rs783147	rs1406891	rs1247568	rs1247340
SLC22A2	0.78	rs7757997	1.00	0.36	0.07	0.00	0.02	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00
SLC22A2	1	rs3850659	**0.63**	1.00	0.07	0.01	0.06	0.00	0.00	0.02	0.04	0.01	0.02	0.01	0.00	0.00	0.00
SLC22A3	1	rs377551	**0.58**	**0.55**	1.00	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SLC22A3	0.95	rs2457576	**0.04**	**0.25**	**0.78**	1.00	0.03	0.00	0.00	0.03	0.04	0.01	0.00	0.02	0.01	0.00	0.00
LPAL2	0.98	rs7754188	**0.21**	**0.32**	**0.06**	**0.47**	1.00	0.00	0.05	0.02	0.03	0.00	0.01	0.01	0.00	0.00	0.00
LPA	0.93	rs6415084	**0.07**	**0.05**	**0.07**	**0.11**	**0.02**	1.00	0.11	0.25	0.00	0.11	0.02	0.00	0.00	0.00	0.01
LPA	1	rs3798221	**0.01**	**0.23**	**0.11**	**0.05**	**0.36**	**0.99**	1.00	0.03	0.00	0.00	0.02	0.04	0.00	0.00	0.00
LPA	1	rs9457986	**0.05**	**0.15**	**0.08**	**0.74**	**0.18**	**0.91**	**0.94**	1.00	0.23	0.01	0.06	0.01	0.00	0.00	0.00
LPA	0.97	rs1367211	**0.23**	**0.39**	**0.02**	**0.39**	**0.27**	**0.04**	**0.10**	**1.00**	1.00	0.20	0.01	0.07	0.01	0.00	0.00
LPA	1	rs1406888	**0.06**	**0.25**	**0.03**	**0.14**	**0.12**	**0.39**	**0.08**	**0.24**	**0.66**	1.00	0.00	0.11	0.01	0.00	0.01
PLG	0.89	rs9458005	**0.11**	**0.13**	**0.12**	**0.17**	**0.10**	**0.22**	**0.62**	**0.26**	**0.24**	**0.02**	1.00	0.00	0.00	0.01	0.00
PLG	0.86	rs783147	**0.27**	**0.43**	**0.04**	**0.13**	**0.21**	**0.00**	**0.25**	**0.47**	**0.51**	**0.44**	**0.23**	1.00	0.01	0.06	0.00
PLG	0.86	rs1406891	**0.03**	**0.01**	**0.06**	**0.15**	**0.03**	**0.00**	**0.11**	**0.00**	**0.13**	**0.09**	**0.03**	**0.17**	1.00	0.06	0.00
PLG	0.98	rs1247568	**0.04**	**0.03**	**0.04**	**0.17**	**0.03**	**0.01**	**0.21**	**0.06**	**0.01**	**0.10**	**0.08**	**0.82**	**0.32**	1.00	0.00
PLG	0.93	rs1247340	**0.08**	**0.07**	**0.04**	**0.06**	**0.05**	**0.11**	**0.07**	**0.08**	**0.06**	**0.13**	**0.07**	**0.12**	**0.03**	**0.05**	1.00

Note:
The lower left part showed the D’ and the upper right part showed the R^2;
RMIP score: resample model inclusion probability score.

Table 3.4 Lp(a) association with haplotypes consisting of SNPs on *LPA* gene reprioritized by LLARRMMA

Haplotype	rs6415084	rs3798221	rs9457986	rs1367211	rs1406888	Freq	β	SE	*P*
A	**T**	**G**	**T**	**T**	**T**	**0.04**	**0.86**	**0.06**	**2.7E-45**
B	**T**	**G**	**T**	**T**	**C**	**0.12**	**0.27**	**0.04**	**1.2E-11**
C	T	G	C	T	C	0.02	-0.12	0.07	0.10
D	**T**	**G**	**C**	**C**	**T**	**0.17**	**0.17**	**0.04**	**1.2E-06**
E	T	G	C	C	C	0.05	0.07	0.06	0.20
F	C	T	C	T	C	0.07	0.03	0.05	0.54
G	**C**	**T**	**C**	**C**	**T**	**0.05**	**-0.17**	**0.05**	**1.7E-03**
H	**C**	**G**	**C**	**T**	**C**	**0.20**	**0.20**	**0.03**	**1.3E-09**
I	**C**	**G**	**C**	**C**	**T**	**0.06**	**0.23**	**0.06**	**7.0E-05**
Rare	*	*	*	*	*	0.04	-0.05	0.07	0.45
Base	C	G	C	C	C	0.17			

Global score= 303.3, df=15, global p-value=1.1E-55
Freq: Haplotype frequency
Effect size(β), standard errors (SE), and P values were calculated for each haplotype compared with the reference haplotype.
Significant associations (P<0.05) are in boldface.
Models were adjusted for age, gender, BMI, 10 PCs, and local ancestry estimate at the top SNP rs9457986.

Table 3.5 SNPs on Exome-chip that significantly ($p<5x10^{-8}$) associated with Lp(a) after controlling for genomic inflation factor

									after adjusting for rs115848955				
Chr[1]	Gene	Function	SNP	Pos(hg19)[2]	MAF[3]	β[4]	SE[5]	*P*[6]	β.adj[7]	SE.adj[7]	*P*.adj[7]	RSQR.in.GWAS[8]	*P*.in.GWAS[9]
6	*MIR1202-ARID1B*	intergenic	rs9478712	156858484	0.14	-0.23	0.03	7.7E-11	-0.20	0.03	6.4E-09	0.99	1.3E-08
6	*SYTL3*	intronic	rs894124	159096121	0.39	-0.15	0.02	8.8E-10	-0.14	0.02	3.1E-09	0.99	7.1E-12
6	*LPA*	intronic	rs6919346	160960359	0.03	-0.43	0.07	1.5E-09	-0.39	0.06	6.0E-09	0.99	2.1E-09
6	*LPA*	splicing	rs41272114	161006077	0.01	-0.80	0.11	6.5E-12	-0.72	0.11	8.8E-11	0.55	3.2E-11
6	*LPA*	intronic	rs1652507	161082461	0.08	-0.36	0.04	4.4E-16	-0.30	0.04	2.6E-12	0.99	6.3E-19
6	*PARK2*	intronic	rs6455767	162148335	0.28	-0.15	0.03	1.4E-08	-0.12	0.03	1.7E-06	0.94	5.2E-10
19	*APOE*	exonic;nonsynonymous	rs7412	45412079	0.11	-0.21	0.04	3.3E-08	-0.21	0.04	3.3E-08	0.28	3.3E-03

Notes:
1: Chr, chromosome; 2: Pos(hg19), physical position of the SNP according to human genome build version 19; 3: MAF, minor allele frequency;
4: β, β coefficients representing the estimated change in the log-Lp(a) level associated with each additional copy of the minor allele;
5: SE, standard error; 6: P, p-value after genomic control;
7: β.adj, SE.adj, and P.adj were reported after adjusting for the lead SNP from GWAS signal, rs115848955;
8: RSQR.in.GWAS, the imputation quality of the Exome-chip SNP in GWAS 1000G; 9: *P*.in.GWAS, the p-value after genomic control in GWAS;
Models were adjusted for age, gender, BMI, and 10 PCs.

Table 3.6 Genes on Exome-chip associated with Lp(a) in gene-based analysis

	Madsen-Browning test						SKAT-O test			
	Gene	*P*	β	SE	cmafTotal[1]	cmafUsed[2]	Gene	*P*	cmafUsed[2]	# of SNPs
level 1	*LPA*	7.4E-14	-0.76	0.10	0.01	0.01				
level 2	*LPA*	1.5E-13	-0.44	0.06	0.04	0.04	*LPA*	1.0E-06	0.04	14
	APOE	1.1E-12	-0.24	0.03	0.13	0.13				
level 3	*LPA*	1.2E-21	-0.40	0.04	0.55	0.08	*LPA*	6.2E-17	0.08	24

Note:
Lp(a) levels were log transformed and residuals were adjusted for age, gender, BMI, 10 PCs, and family relatedness;
Gene-based tests were carried out using Madsen-Browning test and SKAT-O test respectively;
level 1 is the combination of stop-loss, stop-gain, and splice-site regardless of MAF;
level 2 is the combination of SNPs in level 1 and all variants which are predicted to be “damaging” using PolyPhen regardless of MAF;
level 3 is the combination of stop-loss, stop-gain, splice-site and non-synonymous variants with MAF upper limits of 3%;
A gene-based association was defined to be significant if P<0.05/number of genes;
Number of genes for level 1, level 2, and level 3 were 4752, 13658, and 15963, corresponding to a p value of 1.05 x10^{-5}, 3.66x10^{-6}, and 3.13x10^{-6};
1: cmafTotal, cumulative MAF of total Exome-chip variants in the *LPA* gene;
2: cmafUsed, cumulative MAF of Exome-chip variants in the *LPA* gene that were included in the gene-based test.

Table 3.7 All stop-altering, splice-site and non-synonymous variants at *LPA* region

SNP	Pos(hg19)	Gene	MAF	A1	A2	β	SE	Function	LOF (level 1)	Damaging (level 2)	*P*
rs186413938	160952780	*LPA*	8.53E-04	T	C	-0.28	0.40	exonic;nonsynonymous	FALSE	FALSE	0.50
rs41267807	160952816	*LPA*	1.49E-03	C	T	-0.34	0.30	exonic;nonsynonymous	FALSE	TRUE	0.29
rs41267809	160953642	*LPA*	3.62E-03	G	A	-0.60	0.19	exonic;nonsynonymous	FALSE	FALSE	3.01E-03
rs3798220	160961137	*LPA*	8.53E-03	C	T	0.23	0.12	exonic;nonsynonymous	FALSE	TRUE	0.08
rs139145675	160966559	*LPA*	9.81E-03	A	G	-0.47	0.12	exonic;nonsynonymous	FALSE	TRUE	1.35E-04
rs144281871	160968889	*LPA*	5.12E-03	A	C	-0.02	0.16	exonic;nonsynonymous	FALSE	FALSE	0.89
rs143431368	160969693	*LPA*	8.53E-04	C	T	-1.24	0.39	splicing	TRUE	TRUE	2.21E-03
rs200099994	160998167	*LPA*	2.13E-04	T	C	-1.48	0.80	splicing	TRUE	TRUE	0.07
rs41267813	160998199	*LPA*	4.26E-04	A	G	-1.41	0.57	exonic;nonsynonymous	FALSE	TRUE	1.64E-02
rs140720828	160998277	*LPA*	1.71E-03	A	G	-0.67	0.28	exonic;nonsynonymous	FALSE	TRUE	2.11E-02
rs41272114	**161006077**	***LPA***	**1.07E-02**	**T**	**C**	**-0.80**	**0.11**	**splicing**	**TRUE**	**TRUE**	**6.52E-12**
rs76144756	161006084	*LPA*	8.53E-04	A	G	-0.90	0.43	exonic;nonsynonymous	FALSE	FALSE	4.25E-02
rs41272112	161006105	*LPA*	0.05	T	C	-0.12	0.05	exonic;nonsynonymous	FALSE	FALSE	2.82E-02
rs41272110	161006172	*LPA*	2.62E-02	G	T	-0.35	0.07	exonic;nonsynonymous	FALSE	FALSE	2.36E-06
rs7765781	161007496	*LPA*	0.42	G	C	-0.08	0.02	exonic;nonsynonymous	FALSE	FALSE	8.88E-04
rs41267817	161010615	*LPA*	2.13E-04	C	T	-0.62	0.80	exonic;nonsynonymous	FALSE	FALSE	0.45
rs41267819	161011993	*LPA*	2.13E-04	A	G	-1.90	0.80	exonic;nonsynonymous	FALSE	FALSE	2.17E-02
rs200561706	161015041	*LPA*	4.90E-03	A	G	-0.59	0.16	exonic;nonsynonymous	FALSE	TRUE	6.01E-04
rs142720914	161016427	*LPA*	2.13E-04	A	G	0.94	0.74	exonic;nonsynonymous	FALSE	TRUE	0.22
rs199952286	161020632	*LPA*	4.26E-04	A	G	0.10	0.57	exonic;stopgain	TRUE	TRUE	0.87
rs113020022	161020641	*LPA*	6.40E-04	T	G	-0.37	0.46	exonic;nonsynonymous	FALSE	FALSE	0.44
rs41259144	161022107	*LPA*	1.07E-03	T	C	0.46	0.35	exonic;nonsynonymous	FALSE	TRUE	0.21
rs201480327	161026077	*LPA*	4.26E-04	T	C	-0.06	0.49	splicing	TRUE	TRUE	0.91
rs139937718	161026078	*LPA*	4.26E-04	A	G	0.03	0.57	exonic;splicing;nonsynonymous	TRUE	TRUE	0.96
rs200802664	161027512	*LPA*	6.40E-04	C	G	-0.68	0.46	exonic;nonsynonymous	FALSE	FALSE	0.16
rs200163192	161027551	*LPA*	6.41E-04	A	T	-0.26	0.45	exonic;nonsynonymous	FALSE	FALSE	0.57

LOF: loss of function, including stop-loss, stop-gain, and splice-site regardless of MAF; level 1 of gene-based analysis include SNPs with LOF=TRUE.

Damaging: LOF variants plus non-synonymous variants which are predicted to be “damaging” using PolyPhen; Level 2 of gene-based analysis include SNPs with damaging=TRUE
P: p-value after genomic control

Table 3.8 All stop-altering, splice-site and non-synonymous variants at *APOE* region

											after adjusting for LDL			after adjusting for log_TC		
SNP	Gene	Pos (hg19)	Function	amino acid	MAF	A1	A2	β	SE	*P*	β	SE	*P*	β	SE	*P*
rs7412	*APOE*	45412079	exonic;nonsynonymous	R>C	0.11	T	C	-0.21	0.04	3.2E-08	-0.13	0.04	7.6E-04	-0.15	0.04	2.0 E-04
rs769455	*APOE*	45412040	exonic;nonsynonymous	R>C	0.021	T	C	-0.36	0.08	2.0E-05	-0.27	0.08	1.0E-03	-0.31	0.08	2.3E-04
chr19:45412056	*APOE*	45412056	exonic;nonsynonymous	R>H	6.4E-04	A	G	0.09	0.42	0.84	0.33	0.44	0.48	0.30	0.45	0.52

P: p-value after genomic control

Table 3.9 Gene-based analysis in *LPA* region conditioning on the top two variants identified in single-variant analysis

Gene	*P*	β	SE	nsnpsTotal	nsnpsUsed	
LPA	1.2E-21	-0.40	0.04	45	24	
	2.1E-14	-0.34	0.04	44	23	after adjusting for rs41272114
	5.7E-09	-0.33	0.06	43	22	after adjusting for rs41272114 and rs41272110

Gene-based analysis is level 3 Madsen-Browning test (including stop-altering, splice-site, and non-synonymous variants with MAF <0.03);
rs41272114: the strongest associated variant in *LPA* region identified in single-variant analysis, p=6.5E-12;
rs41272110: the second strongest associated variant in *LPA* region identified in single-variant analysis, p=2.4E-06;
nsnpsTotal: total number of Exome-chip variants in the *LPA* gene;
nsnpsUsed: number of Exome-chip variants in the *LPA* gene that were included in the level 3 Madsen-Browning test.

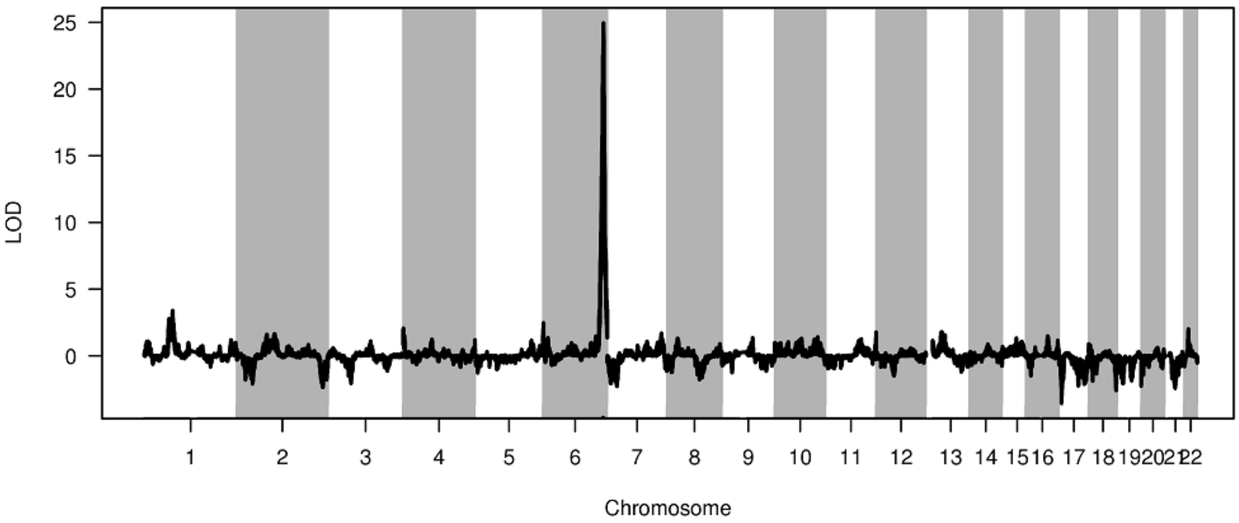


Figure 3.1 The LOD score of genome-wide admixture scan for Lp(a). The LOD score is defined as the log base 10 ratio of the maximum likelihood of the data under a local-ancestry-associated disease model divided by the likelihood of the data under null model. Both alternative and null model include covariate adjustment for global ancestry. Positive LOD scores show the association of increased African ancestry with higher levels of Lp(a), while negative LOD scores show the association of increased African ancestry with lower level of Lp(a).

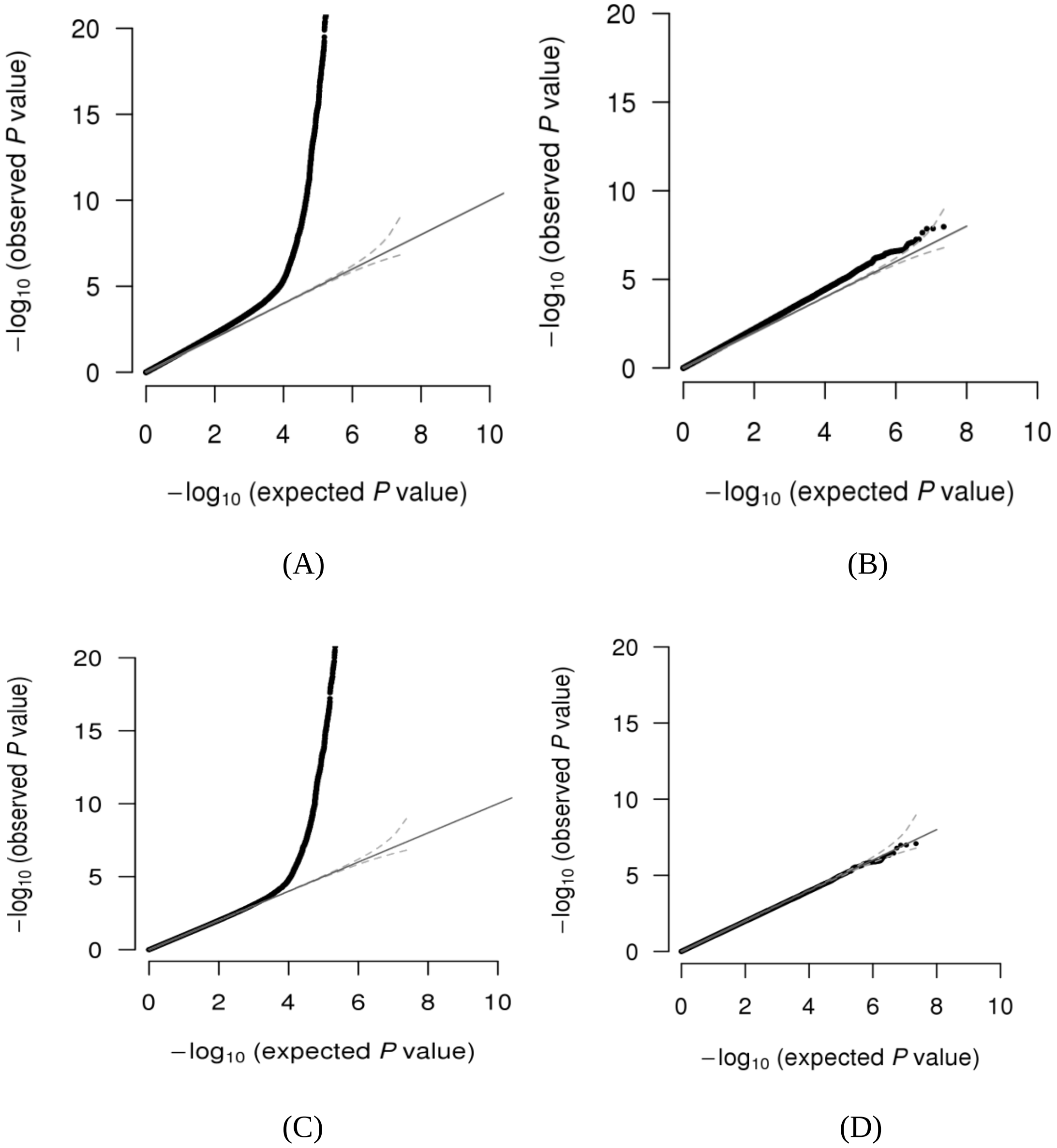


Figure 3.2 Quantile-Quantile (Q-Q) plots of the P-values tested for association with Lp(a) across (A) all SNPs, (B) all remaining SNPs after excluding SNPs on chromosome 6, (C) all SNPs after controlling for inflation factor λ, and (D) all remaining SNPs after excluding SNPs on chromosome 6 and after controlling for inflation factor λ.

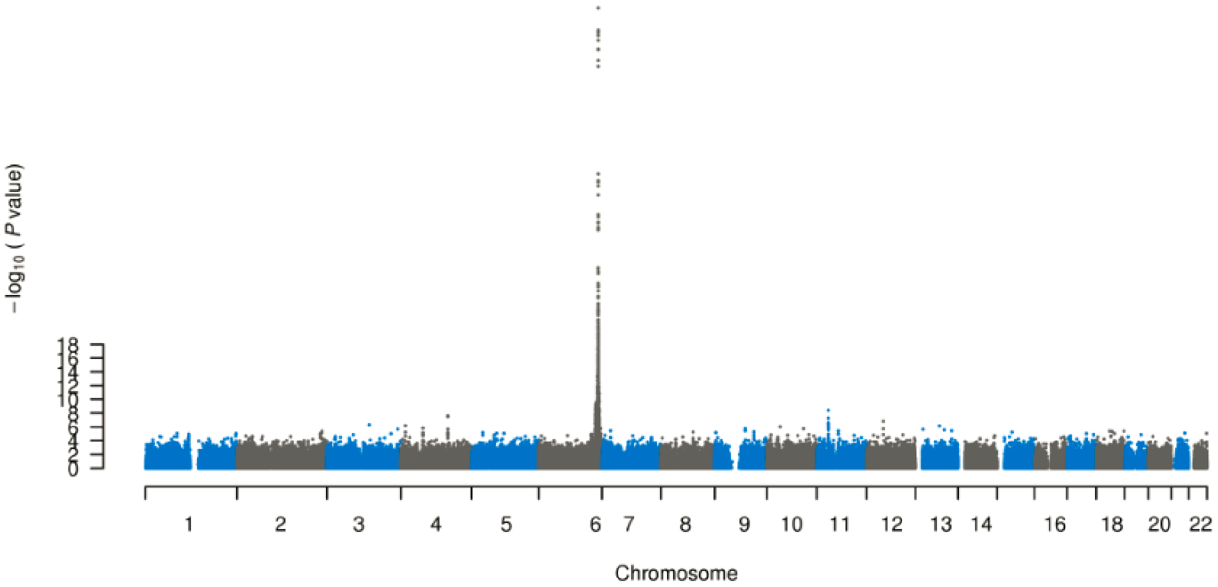
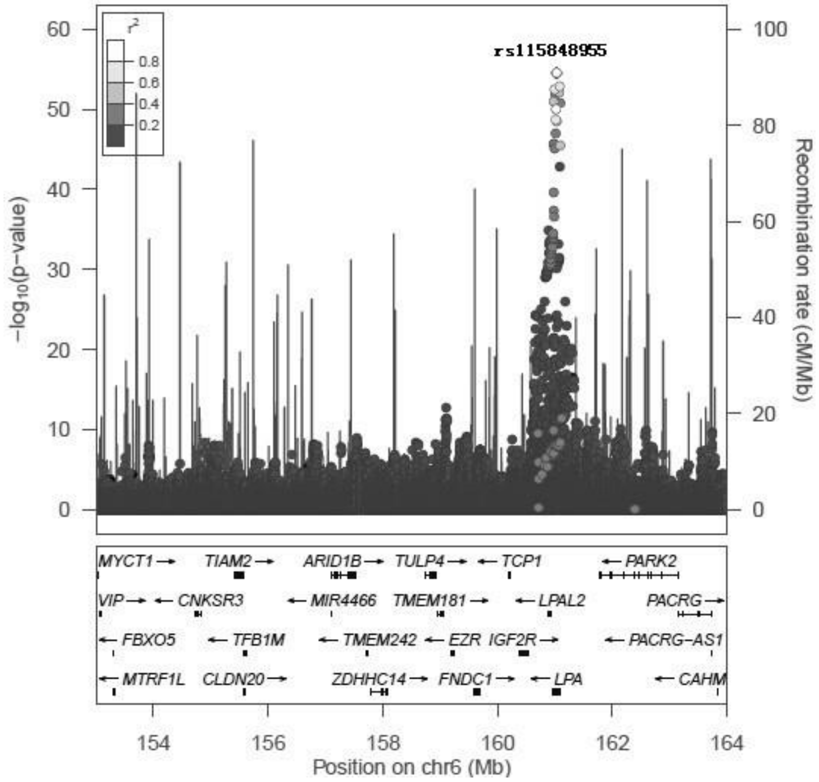
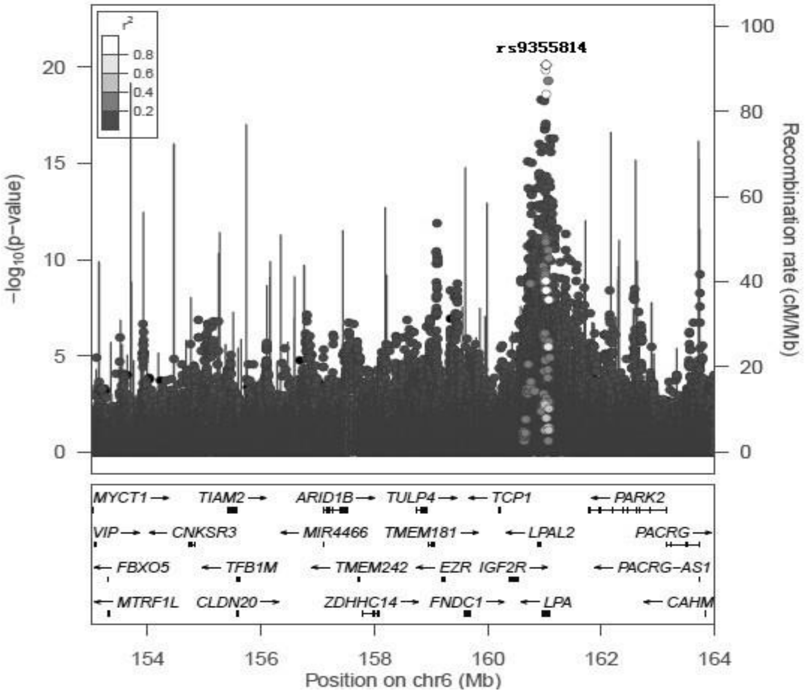


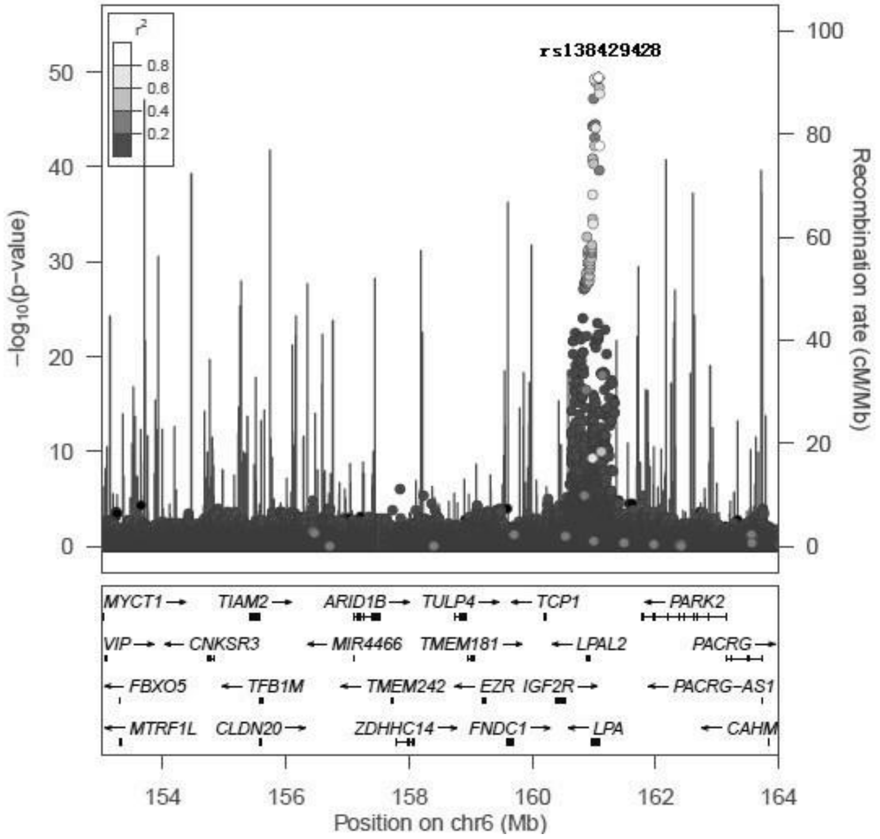
Figure 3.3 Manhattan plot of the -log$_{10}$(P) values by chromosome for Lp(a)


r²
0.8
0.6
0.4
0.2
rs115848955
−log10(p−value)
0
10
20
30
40
50
60
Recombination rate (cM/Mb)
0
20
40
60
80
100
MYCT1→
TIAM2→
ARID1B→
TULP4→
←TCP1
←PARK2
VIP→
←CNKSR3
←MIR4466
TMEM181→
←LPAL2
PACRG→
←FBXO5
←TFB1M
←TMEM242
←EZR
IGF2R→
←PACRG−AS1
←MTRF1L
CLDN20→
ZDHHC14→
FNDC1→
←LPA
←CAHM
154
156
158
160
162
164
Position on chr6 (Mb)

(A)


r²
0.8
0.6
0.4
0.2
rs9355814
−log10(p−value)
0
5
10
15
20
Recombination rate (cM/Mb)
0
20
40
60
80
100
MYCT1→
TIAM2→
ARID1B→
TULP4→
←TCP1
←PARK2
VIP→
←CNKSR3
←MIR4466
TMEM181→
←LPAL2
PACRG→
←FBXO5
←TFB1M
←TMEM242
←EZR
IGF2R→
←PACRG−AS1
←MTRF1L
CLDN20→
ZDHHC14→
FNDC1→
←LPA
←CAHM
154
156
158
160
162
164
Position on chr6 (Mb)

(B)



(C)

Figure 3.4 Regional plot of the -log$_{10}$(P) values for the SNPs in the chromosome 6q region for Lp(a) (A) before adjusting for any SNPs, (B) after adjusting for the top SNP rs115848955 at this locus, and (C) after adjusting for the local ancestry estimate rs6923917. The X axis shows the human genome build 19 coordinates (Mb) and the genes in the region. The Y axis shows the -log$_{10}$ association P values of SNPs on the left, and recombination rates in cM per Mb on the right. Different colors of shading indicate the strength of linkage disequilibrium (LD) (r^2) between the top SNP and the other SNPs tested in the region.

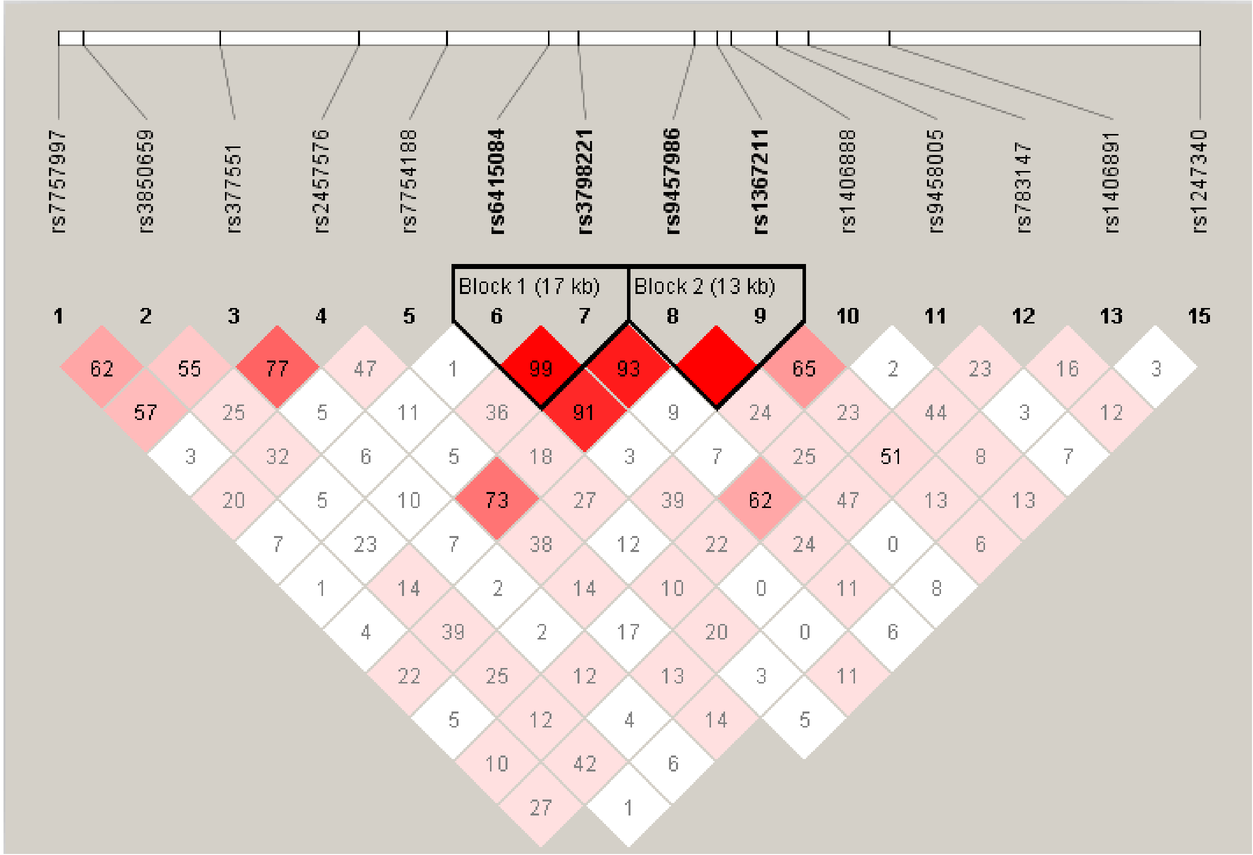


Figure 3.5 Haploview screen showing an LD plot of 15 reprioritized SNPs (RMIP >0.75) in the 1 Mb hit region on chromosome 6. Each square displays the LD D' value between a pair of SNPs. The strength of LD between two SNPs is displayed by the intensity of the color of a square. Thick black triangles depict haplotype blocks by default definition used in Haploview. SNP rs6415084, rs3798221, rs9457986, rs1367211, and rs1406888 were on the *LPA* gene.

CHAPTER IV: A VARIANT NEAR *FGF5* HAS STRONGER EFFECTS ON BLOOD PRESSURE IN CHINESE WITH A HIGHER BODY MASS INDEX [3]

Introduction

Hypertension is an important risk factor for cardiovascular disease (CVD), the leading cause of mortality all over the globe (40). Hypertension is a worldwide problem(146) and its prevalence is increasing in China(147). According to the data from 2007-2008 China National Diabetes and Metabolic Disorders Study, 26.6% of Chinese adults have hypertension(147), causing considerable health and economic burden(148, 149).

The etiology of high blood pressure involves the interplay among many factors. Risk factors include body mass index (BMI), tobacco use, salt and alcohol intake, and physical activity(46). An estimated 30-60% of blood pressure variation is explained by genetic factors(49). Many genetic variants have been identified by genome-wide association studies (GWAS) conducted in multiple populations. Blood pressure is a complex phenotype, and the effects of some variants may be stronger in individuals exposed to specific environmental factors. However, relatively few studies have investigated the possible interaction between genetic variants and environmental factors in affecting blood pressure.

Several loci first identified in subjects of European descent, including *ATP2B1* on chromosome 12q21.33, *FGF5* on 4q21.21, *CYP17A1* on 10q24.32 and *CSK* on 15q24.1 have been validated in East Asians(59, 150). In this study, we genotyped index variants from these

[3] A version of this work was previously published as Li J et al. Am J Hypertens. 2015 Jan 23. pii: hpu263 [Epub ahead of print]

four candidate loci and examined whether the associations with systolic (SBP) and diastolic (DBP) blood pressure could be replicated in 7,319 adults (18 years and older) from 9 provinces in the China Health and Nutrition Survey (CHNS). We further tested whether the associations were influenced by environmental factors including age, sex and BMI. Finally, we attempted to replicate specific findings in 1,996 men from the Fangchenggang Area Male Health and Examination Survey (FAMHES).

Materials and methods

Study design

The China Health and Nutrition Survey (CHNS) is a nationwide longitudinal survey designed to examine a series of economic, sociological, demographic, and health questions in the Chinese population. The design of CHNS has been described in detail elsewhere(151). This study was conducted in nine provinces in China that vary significantly in terms of socioeconomic, health-related, and nutritional status. The sample from each province was drawn using a multistage, random cluster procedure, designed to select a stratified probability sample in each province, with selection of larger cities and smaller suburban and rural villages, from which households were randomly selected. Approximately 19,000 individuals from ~4,400 households participated in the overall survey. The first round of data was collected in 1989, followed by eight additional waves of data collected in 1991, 1993, 1997, 2000, 2004, 2006, 2009, and 2011. The study was approved by the ethics committee at the National Institute of Nutrition and Food Safety at the China Center for Disease Control and Prevention and the Institutional Review Board at the University of North Carolina at Chapel Hill. Written consent was obtained from subjects surveyed in 2009.

Data collection and phenotypes

For this study, we used dietary, clinical, and anthropometric data collected from participants who were age 18 years or older and not pregnant at the time of the 2009 survey. Blood pressure was measured three consecutive times on the same day, with 10 minutes of seated rest before the first measurement and 3-5 minutes intervals between each measurement. SBP and DBP were determined by the first and fifth phase Korotkoff sounds(152), respectively. The average of the three measurements was used for analyses. For participants who take antihypertensive medications, 10 and 5 mm Hg was added to the average SBP and DBP, respectively(153) (154). The smoking responses were dichotomized to define current smoking status. Total salt intake (grams) was estimated based on a combination of three consecutive 24-hour food recalls at the individual level and a food inventory at the household level(155).

Genotyping

Fasting venous blood was collected and stored at -80 °C, and DNA was extracted from buffy coat using the FlexiGene DNA kit (Qiagen, Valencia, CA, USA), according to the manufacturer's instructions. Four candidate variants for blood pressure(150), including rs11105378 at *ATP2B1*, rs1458038 at *FGF5*, rs1004467 at *CYP17A1*, and rs1378942 at *CSK-CYP1A1*, were genotyped using TaqMan chemistry (Applied Biosystems). TaqMan genotyping assays with probes labeled with the fluorophores FAM and VIC were purchased from Applied Biosystems. The Universal PCR Master Mix from Applied Biosystems was used in a 5 µl total reaction volume with 10 ng DNA per reaction. Allelic discrimination was measured automatically on ABI Prism 7900HT (Applied Biosystems) with Sequence Detection Systems. Among 8,221 individuals genotyped, the success rate for each variant was >98%. Genotype distributions were consistent with Hardy-Weinberg equilibrium expectations (*P* > 0.05). The concordance rate between 301 duplicate pairs across all variants was >99.9%.

Statistical analyses

A total of 7,319 non-pregnant adults, age 18 years or older at the time of blood pressure measurement, with complete phenotype, covariate and genotype data were included in tests of association between variants and SBP or DBP. The distributions of SBP and DBP, after accounting for age, sex and BMI, showed no significant deviation from normality; untransformed traits were analyzed. Linear mixed-effects models(156), with a random effect for household to account for the unaccounted for correlation between blood pressure measurements from members in the same family, were used to test the variant additive genetic main effects. The base models included adjustment for covariates significantly (*P* < 0.05) associated with SBP or DBP, age, sex, province and BMI. Additional models also included covariate adjustment for current smoking status and total salt intake. Each variant that showed evidence for a main effect association (*P* < 0.05) with either SBP or DBP was further analyzed for variant-by-environment interactions with age, sex, and BMI by including an interaction term to the above base linear mixed-effects model. Stratified analyses were performed for variants demonstrating evidence of a significant interaction ($P_{\text{interaction}} < 0.05$) by categorizing participants according to quartiles of the environmental factor and testing main effects of the variants within each quartile.

Follow-up sample

We followed up our genotype-by-BMI interaction finding at rs1458038 near *FGF5* in 1,996 men from the FAMHES study using linear regression models with covariate adjustment for age and the main effects for genotype and BMI. As with CHNS, we also performed stratified analyses for the main effects of rs1458038 on SBP and DBP in strata defined by BMI quartiles of the FAMHES participants. As described in Yang et al(157), the FAMHES project was conducted in Fangchenggang city, Guangxi, southern China in 2009. A total of 4,303 Chinese men ranged from age 17 to 88 years old were recruited(158). The subjects used in our current study were limited to men aged 20 to 69 years old (average age 37.5 ± 11.1 years) of self-reported southern Chinese Han ethnicity. 50.8% of men were reported smokers and the average BMI was 23.1 ± 3.4 kg/m^2). Written informed consent to participate in the study was provided by all participating men. The Illumina HumanOmni1-Quad BeadChip was used to perform genome-wide assay of all samples. Trained nurses obtained a single measure of blood pressure on each FAMHES participant by applying a mercury sphygmoma nometer to the right arm of the participants when they were seated in a comfortable sitting position after an at least 5-minute rest period. Participants were asked to avoid vigorous exercise, drinking, and smoking for at least 30 minutes prior to the measurement. Medication history related to hypertension was not available.

Results

Overall, 7,319 CHNS subjects (3,987 females and 3,332 males) with complete phenotype (average SBP and DBP), essential covariates (age, sex, BMI, province), and genotype data were included in the analyses (**Table 4.1**). Higher age, male sex, larger BMI, and current smoking were each associated with higher SBP and DBP (data not shown, *P* < .05). Higher salt intake was associated with higher DBP. Province was also associated with blood pressure levels. Blood pressure levels were higher in residents from Northern provinces than in residents from Southern provinces (**Table 4.2**), which is consistent with the north-south gradient in the prevalence of hypertension in China observed in another cross-sectional survey(159). When all of these covariates were included in a single model, current smoking status and total salt intake no longer showed evidence for association with SBP or DBP. Current smoking status was missing on 4 subjects and total salt intake was missing on 256 subjects. Thus our base model, Model 1, included covariates for age, sex, BMI and province while our secondary model, Model 2, had additional covariates for current smoking status and total salt intake.

Four candidate variants that were previously reported to be associated with SBP and/or DBP in prior studies (49) (59) (150) (160) (161) were tested for association with these traits in the CHNS study. Two variants (rs11105378 near *ATP2B1* and rs1458038 near *FGF5*) were significantly associated ($P < 0.00625 = 0.05/8$) with both SBP and DBP (**Table 4.3).** Variant rs1378942 near *CSK* was nominally associated with SBP ($P = 7.0\text{x}10^{-3}$). The direction and effect size for all eight association tests were consistent with the previous reports. Additional covariate adjustment for current smoking status and total salt intake resulted in very similar findings compared to the base model without these covariates (data not shown).

The three variants associated ($P < 0.05$) with SBP and two variants associated with DBP were tested for interaction with each of three environmental factors (age, sex, and BMI) that significantly affected blood pressure when controlling for all other covariates. Among the 15 tests performed, four showed nominally significant results ($P_{\text{interaction}} < 0.05$). The observed interaction between rs1458038 and BMI on SBP was significant ($P_{\text{interaction}} = 0.0018$; $\beta=0.25$) after Bonferroni correction for multiple tests ($P_{\text{interaction}} < 0.0033$, 0.05/15 tests). The interaction between rs1458038 and BMI was also nominally associated ($P_{\text{interaction}} = 0.049$, $\beta = 0.10$) with DBP. The two remaining nominally significant interactions were between rs1378942 and sex on SBP ($P_{\text{interaction}} = 0.045$, $\beta = 1.54$) and between rs11105378 and age on DBP ($P = 0.026$, $\beta = 0.03$).

Further analyses stratified by BMI levels were conducted in CHNS to interrogate the interactions between rs1458038 and BMI on SBP and DBP. We stratified the samples according to BMI quartiles and performed main-effects regression analyses of rs1458038 on SBP and DBP, adjusting for the secondary model covariates. The T allele of rs1458038 was significantly associated with higher SBP in the highest quartile of BMI (Q4: P = 1.9x10^{-6}), but the association was not significant in lower quartiles (Q1: P = 0.69; Q2: P = 0.11; Q3: P = 5.7x10^{-2}) (**Table 4.4**). The magnitude of the effect size estimates increased with BMI (Q1: β = 0.21; Q2: β = 0.88; Q3: β = 1.05; Q4: β = 2.93). A similar pattern of interaction of this variant with BMI was observed in its effects on DBP. Here too, rs1458038 genotype had stronger main effects on DBP in individuals with the largest BMI (Q1: β = 0.33, P = 0.34; Q2: β = 0.12, P = 0.72; Q3: β = 0.80, P = 2.1x10^{-2}; Q4: β = 1.23, P = 1.0x10^{-3}; **Table 4.4**).

In FAMHES, there was a trend ($P \leq 0.10$) with consistent direction of effects to CHNS, for main effects of rs1458038 on both DBP (P = 0.058) and SBP (P = 0.10) (**Table 4.5**). The interaction effects between rs1458038 and BMI affecting SBP and DBP were in the same direction as CHNS but did not reach statistical significance in FAMHES ($P_{interaction}$ = 0.36 for DBP; $P_{interaction}$ = 0.62 for SBP). Consistent with CHNS, the estimates of the main effects of rs1458038 on SBP (Q1: β = 1.10; Q2: β = 0.37; Q3: β = 1.02; Q4: β = 1.63) and DBP (Q1: β = 0.44; Q2: β = 0.68; Q3: β = 0.49; Q4: β = 1.59) were strongest in men grouped in the highest quartile of BMI (**Table 4.6**). The association between rs1458038 and DBP was nominally significant (P = 0.034) only in the highest BMI quartile. Because FAMHES only contained males, we could not assess the interaction between rs1378942 and sex on SBP. Finally, due the absence of evidence for main effects at rs11105378 (P = 0.96), we did not assess the evidence for an interaction between this SNP and age on DBP in FAMHES.

Discussion

In this study, we tested the association of four candidate variants with SBP and DBP in 7,319 Chinese individuals, from nine provinces, participating in the CHNS and followed up our top findings in 1996 Chinese men from the province of Guangxi participating in FAMHES. Rs1458038, upstream of *FGF5* on chromosome 4 and rs11105378, upstream of *ATP2B1* on chromosome 12 were significantly associated with both SBP and DBP ($P < 0.00625$) in CHNS while rs1378942, an intronic variant in *CSK*, was nominally associated with SBP ($P = 7.0 \text{x} 10^{-3}$). Most interestingly, in CHNS, we detected a significant interaction between rs1458038 and BMI affecting SBP and a nominally significant interaction between this same variant and BMI affecting DBP. Specifically, we showed that the effect of the rs1458038 risk genotype is considerably stronger in subjects in the heaviest quartile of the BMI distribution.

Rs1458038, 5' of *FGF5*, was identified in Europeans(160), and also confirmed in Japanese(150). Located at chr4:81164723, rs1458038 is ~23 kb upstream of the transcription staring site of *FGF5* (chr4:81187742). The association of nearby variant rs16998073 (chr4:81184341, LD $r^2 = 0.917$ in CEU) was identified in East Asians(59) (162) (163) and Europeans(161). Rs11105378, an intronic variant in *ATP2B1*, was identified by the CHARGE and Global BPgen Consortia (49), and confirmed in Japanese (150). Several other variants that are in high LD with rs11105378 ($r^2 > 0.8$ in CEU) were also identified in different ethnicities (49, 59, 160, 164). Differences in *ATP2B1* mRNA expression levels in umbilical artery smooth muscle cells were observed among different rs11105378 genotypes (150). To our knowledge, no study has provided any evidence that either rs1458038 or rs16998073 are in any promoter or enhancer regions. We replicated the associations of these two variants with both SBP and DBP, reaffirming the importance of these variants on blood pressure in the Chinese population. Rs1378942, an intronic variant 2,306 bp from a splice site in *CSK*, was significantly associated with SBP and DBP in Europeans(160); however, only nominal association was observed in East Asians(59) and Japanese(150). In our study, we only replicate (*P* < 0.05) its association with SBP, although the effect size estimates for both SBP and DBP were similar in both direction and size to the previous study in Japanese. We failed to replicate blood pressure associations with rs1004467, an intronic variant 35 bp from a splice-site in *CYP17A1*. Here too, our directions of effect were consistent with the previous Japanese study. However, the size of the effects for DBP and SBP were about half of those reported previously for this variant.

An interesting finding of this study is the evidence for interaction between rs1458038 5' of *FGF5* at 4q21 and BMI in affecting both SBP and DBP in CHNS. BMI is a strong risk factor that affects blood pressure; the prevalence of hypertension and mean levels of SBP and DBP increase as BMI increases(165, 166). While variants 5' of *FGF5* at 4q21 are associated with blood pressure in the Chinese population(162, 163), the interaction between these two risk factors on affecting blood pressure has not been previously reported. In our study, the increasing differences in blood pressure between rs1458038 TT, CT and CC carriers with increasing BMI support the hypothesis that larger BMI enhances the effects of the rs1458038 risk allele on increasing blood pressure. The strongest evidence for association between rs1458038 and the blood pressure measures in CHNS were observed in the highest quartile of BMI. In FAMHES, the strongest effects of rs14358038 for both blood pressure traits were also observed in the highest quartile of BMI. The observed main effects of rs14358038 in the highest BMI quartile of FAMHES men were ~50% larger than in any other quartile for SBP and more than twice as large as the estimated effects in any other quartile for DBP. While the tests of rs14358038-BMI interaction on the blood pressure measures were not significant in FAMHES, they were directionally consistent with CHNS.

When assessing the overall impact of the FAMHES results, some differences between CHNS and FAMHES are worth noting. FAMHES includes only males from the province of Guangxi, while CHNS includes both males and females from nine provinces, including Guangxi. The sample size of FAMHES (n=1,996) is considerably smaller than that of CHNS (n=7,319). FAMHES men were also younger, on average, than CHNS individuals (average age = 37.5 years of age in FAMHES and 50.8 years of age in CHNS). Both the average SBP and DBP measures were lower in FAMHES men compared to CHNS men; however, these results are consistent with the younger ages in FAMHES and the lower blood pressure values observed in CHNS men from Guangxi. SBP and DBP were measured only once in FAMHES compared to three measurements (we applied the average) in CHNS. Finally, no hypertensive medication history was available in FAMHES. In CHNS, 399 subjects were known to be on hypertensive medication and an offset was applied to the SBP and DPB measures in CHNS to account for medication use.

An interaction between variants at 4q21 and BMI in affecting blood pressure is biologically plausible. Four genes including *ANTXR2*, *PRDM8*, *FGF5*, and *C4orf22* are located near rs1458038. *ANTXR2* encodes anthrax toxin receptor 2, a protein involved in angiogenesis(167). A recent functional study(168), using in vivo small interfering RNA (siRNA) silencing in mice, suggested *ANTXR2* is the most likely causative gene in the 4q21 region that regulates individual differences in blood pressure in humans. The study proposed that the lower *Antxr2* expression can lead to a decrease in the proliferation of endothelial cells, prevent the formation of capillary network, and result in microvascular rarefaction and increase of BP.

Many studies observed a developing microvascular rarefaction within skeletal muscle during the metabolic syndrome including obesity, insulin resistance/type II diabetes mellitus, dyslipidemia, and hypertension. (169, 170). Microvascular dysfunction is a potential mechanism in the pathogenesis of obesity-associated insulin resistance and hypertension(171, 172). In addition to *ANTXR2*, *FGF5* is also possibly involved in the metabolic syndrome. *FGF5* encodes fibroblast growth factor 5, a member of the fibroblast growth factor (FGF) family. Several members of the FGF family have been shown to affect obesity by regulating fatty acid oxidation and lipid metabolism. Treatment with exogenous recombinant human FGF21 protein via infusion or injection can lead to weight loss and improvement of lipid profiles in diet-induced obese (DIO) mice(173) and diabetic rhesus monkeys(174). Transgenic mice expressing human *FGF19* display increased metabolic rate and decreased adiposity(175). Although FGF5 has not yet been shown to have a direct effect on BMI, this member of the FGF family may also play a role in regulating metabolism and affecting obesity. The role of PRDM8 and C4orf22 in the pathogenesis of obesity and hypertension is not known.

In conclusion, the association of rs11105378 near *ATP2B1* and rs1458038 near *FGF5* with both SBP and DBP were replicated in CHNS. In addition, the magnitude of the associations between rs1458038, 5' of *FGF5*, and blood pressure were modified by BMI in CHNS individuals. While we were unable to formally replicate the interaction in a second Chinese cohort, evidence from both studies implicate that the risk genotype at rs1458038 is particularly important in Chinese individuals with higher BMI. To our knowledge, this is the first reported interaction between a variant in or near *FGF5* and BMI on blood pressure. Further studies in Chinese and other populations are needed to confirm this finding.

Conflicts of interest

The authors stated no conflict of interest.

Table 4.1 Characteristics of the China Health and Nutrition Survey participants analyzed

Characteristics	Female (n=3,987)		Male (n=3,332)		Total(n=7,319)	
	mean ± SD	median (range)	mean ± SD	median (range)	mean ± SD	median (range)
SBP (mm Hg)	124.3 ± 20.8	120.0 (76.7, 266.7)	126.8 ± 18.2	122.0 (80.0, 229.3)	125.4 ± 19.7	121.0 (76.7, 266.7)
DBP (mm Hg)	79.4 ± 11.6	80.0 (44.0, 152.0)	82.5 ± 11.3	80.7 (50.0, 136.0)	80.8 ± 11.6	80.0 (44.0, 152.0)
Age (years)	50.9 ± 15.0	51.1 (18.0, 98.9)	50.7 ± 15.1	51.0 (18.0, 92.3)	50.8 ± 15.0	51.0 (18.0, 98.9)
BMI (kg/m^2)	23.4 ± 3.5	23.0 (13.4, 38.8)	23.4 ± 3.4	23.2 (13.4, 37.2)	23.4 ± 3.5	23.1 (13.4, 38.8)
Current smoker (%)	3.70%	-	55.00%	-	27.04%	-
Total salt intake (grams)	4.5 ± 2.6	3.9 (0.1, 22.2)	4.9 ± 2.7	4.3 (0.2, 21.3)	4.7 ± 2.6	4.1 (0.1, 22.2)

Values are means ± SD, medians (range), or %. The average of the three measurements of SBP or DBP was used for analysis. Total salt intake was based on a combination of three consecutive 24-hour food recalls at the individual level and a food inventory at the household level. The sample size is 7,315 with data on current smoking status and 7,063 with data on total salt intake.

SBP: systolic blood pressure; DBP: diastolic blood pressure; BMI: body mass index.

Table 4.2 Mean levels of SBP and DBP in residents from each province

Province Number	Female	Male	Total	Sample Size	Province Name	Range of Latitude
SBP (mm Hg)						
province=23	124.7 ± 21.2	127.3 ± 16.0	125.9 ± 19.0	768	Heilongjiang	(43.2°, 53.3°)
province=21	130.3 ± 21.7	131.0 ± 19.8	130.6 ± 20.9	719	Liaoning	(38.4°, 43.3°)
province=37	126.8 ± 16.2	130.8 ± 16.8	128.6 ± 16.6	835	Shandong	(34.2°, 38.2°)
province=41	122.8 ± 19.3	125.6 ± 17.3	124.0 ± 18.5	776	Henan	(31.2°, 36.2°)
province=32	125.2 ± 20.8	128.8 ± 18.5	126.9 ± 19.8	1004	Jiangsu	(30.4°, 35.2°)
province=42	122.9 ± 22.5	126.7 ± 18.6	124.6 ± 20.9	729	Hubei	(29.0°, 33.2°)
province=43	123.8 ± 23.9	124.5 ± 18.5	124.1 ± 21.6	969	Hunan	(24.4°, 30.1°)
province=52	119.8 ± 19.7	125.4 ± 19.7	122.3 ± 19.9	564	Guizhou	(24.4°, 29.1°)
province=45	121.7 ± 19.4	121.6 ± 16.7	121.7 ± 18.2	955	Guangxi	(20.5°, 26.2°)
DBP (mm Hg)						
province=23	82.9 ± 13.2	85.5 ± 11.4	84.1 ± 12.4	768	Heilongjiang	(43.2°, 53.3°)
province=21	83.5 ± 11.6	85.7 ± 11.3	84.5 ± 11.5	719	Liaoning	(38.4°, 43.3°)
province=37	81.3 ± 10.0	85.3 ± 10.5	83.1 ± 10.4	835	Shandong	(34.2°, 38.2°)
province=41	80.1 ± 11.1	83.8 ± 10.5	81.8 ± 11.0	776	Henan	(31.2°, 36.2°)
province=32	79.7 ± 11.0	83.5 ± 10.9	81.5 ± 11.1	1004	Jiangsu	(30.4°, 35.2°)
province=42	76.9 ± 11.8	80.0 ± 10.6	78.3 ± 11.4	729	Hubei	(29.0°, 33.2°)
province=43	76.8 ± 11.8	79.8 ± 11.8	78.2 ± 11.9	969	Hunan	(24.4°, 30.1°)
province=52	77.4 ± 11.4	81.1 ± 13.0	79.1 ± 12.3	564	Guizhou	(24.4°, 29.1°)
province=45	77.0 ± 10.6	78.7 ± 9.9	77.8 ± 10.3	955	Guangxi	(20.5°, 26.2°)

Values for SBP and DBP are mean ± SE.

Table 4.3 Main effect association of variants with SBP and DBP in CHNS

						Model 1				Model 2			
Variant	Chr	Nearby Gene	EA	EAF	Trait	Sample	β	SE	*P*	Sample	β	SE	*P*
rs1458038	4	*FGF5*	T	0.42	SBP	7,272	1.29	0.28	**2.8E-06**	7,013	1.28	0.28	**5.9E-06**
					DBP		0.70	0.17	**5.2E-05**		0.68	0.18	**1.1E-04**
rs1004467	10	*CYP17A1*	A	0.65	SBP	7,308	0.37	0.29	0.20	7,048	0.44	0.29	0.13
					DBP		0.11	0.18	0.54		0.13	0.18	0.49
rs11105378	12	*ATP2B1*	C	0.66	SBP	7,187	1.22	0.29	**3.3E-05**	6,929	1.29	0.30	**1.8E-05**
					DBP		0.73	0.18	**6.9E-05**		0.80	0.19	**1.8E-05**
rs1378942	15	*CSK*	C	0.84	SBP	7,276	1.02	0.38	**7.0E-03**	7,018	0.97	0.39	**0.012**
					DBP		0.35	0.24	0.14		0.36	0.24	0.14

β coefficients represent the estimated change in the level of blood pressure associated with each additional copy of the effect allele, designed as the blood pressure raising allele. Covariates for model 1 were age, gender, province, and BMI. Covariates for model 2 were age, gender, province, BMI, current smoking status, and total salt intake.
EA: effect allele; EAF: effect allele frequency; SE: standard error

Table 4.4 Mean levels of SBP and DBP by rs1458038 genotype and BMI quartile in CHNS.

	All participants	CC	CT	TT	N	β (SE)	*P*
SBP							
Q1	118.9 ± 18.6	118.7 ± 18.3	118.7 ± 18.2	120.1 ± 20.0	1818	0.21 (0.54)	0.69
Q2	122.8 ± 18.2	121.7 ± 17.4	123.0 ± 18.1	124.7 ± 20.2	1818	0.88 (0.55)	0.11
Q3	127.1 ± 18.7	126.2 ± 18.6	127.2 ± 18.1	128.6 ± 20.4	1818	1.05 (0.55)	0.057
Q4	132.8 ± 20.4	130.8 ± 19.2	133.4 ± 20.7	135.1 ± 21.6	1818	2.93 (0.61)	**1.9E-06**
DBP							
Q1	76.4 ± 10.9	76.0 ± 10.6	76.4 ± 10.8	77.1 ± 11.6	1818	0.33 (0.34)	0.34
Q2	78.9 ± 10.8	78.7 ± 10.8	79.1 ± 10.6	78.9 ± 11.4	1818	0.12 (0.34)	0.72
Q3	82.2 ± 10.8	81.4 ± 11.0	82.5 ± 10.4	83.0 ± 11.3	1818	0.80 (0.34)	**0.021**
Q4	85.7 ± 11.7	85.1 ± 11.7	85.6 ± 11.9	87.4 ± 11.2	1818	1.23 (0.38)	**1.0E-03**

Q1-Q4: the lowest BMI quartile to the highest BMI quartile
Q1: BMI <20.93; Q2: 20.93≤BMI<23.11; Q3: 23.11≤BMI<25.65; Q4: BMI≥25.65. Values for SBP and DBP are mean ± SE.

Table 4.5 Main effect association of rs1458038 with SBP and DBP in FAMHES

Variant	Chr.	Nearest Gene	EA	EAF	Trait	Sample	β	SE	P
rs1458038	4	*FGF5*	T	0.42	SBP	1996	0.80	0.49	0.10
					DBP	1996	0.62	0.33	0.058

Table 4.6 Mean levels of SBP and DBP by rs1458038 genotype and BMI quartile in FAMHES.

	All participants	CC	CT	TT	β (SE)	*P*
SBP						
Q1	113.3 ± 13.2	112.7 ± 13.6	112.8 ± 13.1	115.1 ± 12.9	1.10 (0.82)	0.18
Q2	116.6 ± 14.4	116.4 ± 15.7	116.6 ± 14.2	117.1 ± 12.2	0.37 (0.92)	0.69
Q3	118.6 ± 14.8	117.7 ± 14.0	118.8 ± 14.8	119.7 ± 16.6	1.02 (0.94)	0.28
Q4	124.3 ± 17.1	122.9 ± 15.8	124.8 ± 16.8	126.0 ± 20.6	1.63 (1.12)	0.15
DBP						
Q1	73.5 ± 8.8	73.3 ± 9.3	73.2 ± 8.7	74.3 ± 8.6	0.44 (0.55)	0.43
Q2	75.0 ± 9.3	74.2 ± 9.9	75.6 ± 9.4	75.3 ± 8.2	0.68 (0.60)	0.26
Q3	77.5 ± 9.2	77.0 ± 9.1	77.8 ± 8.8	77.9 ± 10.4	0.49 (0.58)	0.40
Q4	81.6 ± 11.5	80.1 ± 10.3	82.2 ± 11.3	83.0 ± 14.0	1.59 (0.75)	**0.034**

Q1-Q4: the lowest BMI quartile to the highest BMI quartile
Q1: BMI <20.80; Q2: 20.80≤BMI<23.04; Q3: 23.04≤BMI<25.50; Q4: BMI≥25.50. Values for SBP and DBP are mean ± SE.

CHAPTER V: USE OF SELECTIVE PHENOTYPING TO INCREASE POWER OF TWO-STAGE GENETIC ASSOCIATION STUDIES FOR BOTH GENOTYPED AND IMPUTED DATA [4]

Introduction

Genome-wide association studies (GWASs) are a powerful method for identifying common genetic variants contributing to human disease(176). GWASs require genotyping a large number of genetic markers on a large number of subjects. Cost-efficient sample selection for GWASs is often necessary due to budget limitations. In the past, the expense of genotyping has been the main cost constraint. Today, chip-based high-throughput technologies have dramatically decreased genotyping costs and, as a consequence, genotype data is routinely available on large numbers of potential study subjects. Much of the burden of expense has now shifted from the cost of genotyping to the cost of phenotyping new traits, especially when phenotyping involves mRNA abundance data obtained from microarray experiment, novel blood-based biomarkers or complex physiological and behavioral traits. Selective phenotyping can be a powerful approach for increasing the power of genetic association studies that are under trait-measurement-related sample size constraints (77). The approach utilizes available genetic information on the complete sample to select a subset of individuals for phenotyping that will improve statistical power for the markers of greatest interest.

Two-stage GWASs have been routinely applied during the past decade to increase statistical power and reduce genotyping costs. Herein, we describe a novel two-stage selective-

[4] A version of this work will be submitted for publication.

phenotyping design that will increase the power for GWAS studies, with existing genetic data and to-be-measured newly quantitative traits. To date, selective-phenotyping designs have been proposed for single-stage genetic association studies, where the markers of greatest interest have already been determined. A GWAS is typically a discovery-based study, where the identification of the "interesting" markers is the overarching goal of the study. Thus, it stands to reason that before we can apply selective phenotyping in a meaningful way, we first have to identify which markers we should focus the phenotype selection on. This dynamic is a natural fit for a two-stage selective-phenotyping design, where in the first stage the investigators phenotype a random subset of subjects to identify the most promising markers and then in the second stage the investigators use their remaining resources to carefully select a fraction of the remaining subjects to phenotype based on their observed genotypes for these most promising markers. We show that combining the data from the two stages results in greater statistical power for discovery compared to the default method of simply selecting the same number of random subjects to phenotype for a single-stage association study.

Methods to maximize power using selective phenotyping have been described previously, albeit for considerably smaller studies than a GWAS (78-81). The common rationale behind these methods is to identify the subset of subjects who are as genetically dissimilar as possible with respect to distributions of marker genotype data across all markers of interest. If only a single marker is of interest, then selecting an equal number of homozygotes for the minor and major allele for that marker to phenotype would be the optimal strategy for most plausible inheritance models (i.e. if we ignore the possibility of overdominance). The maximization of genotypic diversity simultaneously across multiple markers is a complex optimization problem, as some subjects who have the less common homozygous genotype for some markers of interest could also have the more common homozygous genotypes for other markers of interest. The computation complexity of the problem grows quickly as the number of target markers (e.g. those markers determined to be of interest for the phenotype-selection optimization) increases.

Simulated annealing (SA) is a computationally efficient algorithm for complex optimization problems, such as finding the global minimum or maximum of functions containing many variables (177, 178)’(179). In this study, we describe a selective-phenotyping method that uses SA to identify the optimal subset of subjects to be phenotyped in Stage 2 of a two-stage GWAS, based on their available directly genotyped or imputed dosage data across multiple markers of interest, as determined by the results from a subset of randomly phenotyped subjects used in Stage 1. We then combine the results from the two stages using a joint-analysis approach to assess the overall significance of all markers, including those not included in the Stage 2 selection process, as all samples will have available phenotypic and genetic data on all available genotyped and imputed markers. Critical parameterization of the two-stage approach includes the proportion of available subjects to phenotype in Stages 1 versus Stage 2 and number of markers to include in the phenotype-selection process in Stage 2. To underscore this point, power is ideally optimized when the truly, and only truly, associated markers are selected for the Stage 2 optimization process that is allowed to consider the largest possible sample of subjects available for phenotyping. The probability of the true markers being included in the Stage 2 selective-phenotyping optimization can be increased by either increasing the proportion of subjects to be randomly phenotyped in Stage 1 or by increasing the number of markers included in the Stage 2 selection algorithm, however, the trade-off is that either fewer samples are then available for the Stage 2 selection strategy or the proportion of true markers included in the optimization algorithm is likely diluted thereby reducing the benefits of the phenotype-selection algorithm.

Materials and methods

Simulated-annealing algorithm for selective phenotyping

In this study, we use a SA-based algorithm to maximize the leverage of specific targeted markers by selecting subjects to phenotype from the larger cohort pool that maximizes the average genotype dispersion across that set of targeted genotyped markers included in the selection process. For a targeted marker, increasing the dispersion, or variance, of the marker genotypes will result in greater expected precision of the regression coefficient that describes the relationship between the outcome measure and the marker, leading to greater statistical power to identify a significant association between the outcome and the marker when such a relationship exists. For a given marker j, the dispersion = $\sum_{i=1}^{n}(x_{ij} - \overline{x}_{\cdot j})^2$ which we label as D_j, is the sum of the squared difference between each of the *n* subjects genotypes and the overall genotypic mean, where genotype is scored 0, 1 or 2 for each subject corresponding to the carried number of copies of the minor allele. We further define the average of D_j across *m* target markers as $\overline{D} = \frac{\sum_{j=1}^{m} D_j}{m} = \frac{\sum_{j=1}^{m}\sum_{i=1}^{n}(x_{ij}-\overline{x}_{\cdot j})^2}{m}$.

Parameters in our simulated annealing algorithm include the function to be optimized $\overline{D} = \frac{\sum_{j=1}^{m} D_j}{m} = \frac{\sum_{j=1}^{m}\sum_{i=1}^{n}(x_{ij}-\overline{x}_{\cdot j})^2}{m}$, the initial temperature (t^0), the rate of cooling (α) and the convergence criteria (ε). We assume that we have access to genotype data on $\boldsymbol{n.geno}$ subjects across $\boldsymbol{M}$ (=1,000,000) markers prior to phenotype sample selection, but we can only afford to measure the phenotypes of $\boldsymbol{n.pheno}$ subjects ($\boldsymbol{n.pheno} < \boldsymbol{n.geno}$) in total. We refer to $\boldsymbol{pi.sample}$ as the proportion of these $\boldsymbol{n.pheno}$ subjects to be phenotyped in Stage 1, so that $\boldsymbol{n.pheno} \times \boldsymbol{pi.sample} = \boldsymbol{n1}$ subjects are phenotyped in Stage 1 and $\boldsymbol{n.pheno} \times (\boldsymbol{1} -$

$\boldsymbol{pi.sample}) = \boldsymbol{n2}$ subjects are phenotyped in Stage 2. We refer to $\boldsymbol{pi.marker}$ as the proportion of markers to be used in the SA algorithm for selecting the $\boldsymbol{n2}$ subjects, from the $\boldsymbol{n.geno - n1}$ possible subjects, to be phenotyped in Stage 2; thus, ($\boldsymbol{M \times pi.marker = m2}$**)** candidate markers are used in the SA algorithm to select subjects for phenotyping in Stage 2. We denote $\boldsymbol{S_i}$ as subject *i*, $\boldsymbol{Q_1}$as the set of $\boldsymbol{n1}$ subjects in Stage 1 so that $\boldsymbol{Q_1}$={$\boldsymbol{S_1, S_2, \ldots S_{n1}}$}, $\boldsymbol{Q_2^k}$ as a set of $\boldsymbol{n2}$ subjects selected in Stage 2 from the complete ($\boldsymbol{n.geno - n1}$) sample at iteration $\boldsymbol{k}$ so that $\boldsymbol{Q_2^k}$**={**$\boldsymbol{S_{n1+1}, S_{n1+2}, \ldots S_{n1+n2}}$**}.** We denote $\boldsymbol{Q_{1+2}^k}$as a set of $\boldsymbol{n.pheno}$ subjects after combining the $\boldsymbol{n1}$ subjects in Stage 1 and $\boldsymbol{n2}$ subjects selected in Stage 2 at iteration $\boldsymbol{k}$ together, so that $\boldsymbol{Q_{1+2}^k = Q_1 + Q_2^k}$. We denote $\boldsymbol{Q_3^k}$ as the set of ($\boldsymbol{n.geno - n1 - n2}$) subjects left at iteration $\boldsymbol{k}$ so that $\boldsymbol{Q_3^k}$**={**$\boldsymbol{S_{n1+n2+1}, S_{n1+n2+2}, \ldots S_{n.geno}}$}. Let $\boldsymbol{Q_2^{k+1}}$ = $\boldsymbol{Q_2^k - S_a + S_b}$, where $\boldsymbol{S_a}$ is a subject *a* randomly selected from $\boldsymbol{Q_2^k}$ and removed from $\boldsymbol{Q_2^k}$, and $\boldsymbol{S_b}$ is a subject *b* randomly selected from $\boldsymbol{Q_3^k}$ and added to $\boldsymbol{Q_2^{k+1}}$. Thus, $\boldsymbol{Q_{1+2}^{k+1}}$=$\boldsymbol{Q_1 + Q_2^{k+1}}$. Let $\overline{D}^k = \frac{\sum_{j=1}^{m2}\sum_{i=1}^{n2}(x_{ij}-\overline{x}_{.j})^2}{m2}$ for set of subjects $i \in \boldsymbol{Q_2^k}$, respectively. Our goal is to iteratively maximize $\overline{D}$ using simulated annealing. The iterative approach is as follows:

1) Select a random $\boldsymbol{Q_1}$ (Stage 1 samples) and to be removed from possible Stage 2 samples.
2) Set t^0, the starting temperature, α, the constant cooling rate, and ε, the user-defined threshold for convergence.
3) Randomly select initial random Stage 2 sample $\boldsymbol{Q_2^0}$, and calculate $\overline{D}^0$based on $\boldsymbol{Q_2^0}$.
4) Create $\boldsymbol{Q_2^1}$ by randomly removing one subject from $\boldsymbol{Q_2^0}$ and randomly adding one subject from $\boldsymbol{Q_3^0}$.
5) Calculate $\overline{D}^1$based on $\boldsymbol{Q_2^1}$.

6) If $\overline{D}^{1} > \overline{D}^{0}$, sample $\boldsymbol{Q_2^0}$ is replaced by sample $\boldsymbol{Q_2^1}$ as the new “best” sample. Else, sample $\boldsymbol{Q_2^0}$ is replaced by sample $\boldsymbol{Q_2^1}$ with probability = $\exp\{(\overline{D}^{1} - \overline{D}^{0})/t^0\}$.That is, always accept a favorable move and sometimes choose the less favorable move with decreasing probability as the system cools.

7) Define $t^1 = \alpha \times\ t^0$.

8) Continue steps 8-10 iteratively until step 11 is satisfied, create sample $\boldsymbol{Q_2^{k+1}}$ from sample $\boldsymbol{Q_2^k}$ in a similar fashion, and calculate $\overline{D}^{k+1}$ and $\overline{D}^{k}$ based on $\boldsymbol{Q_2^{k+1}}$ and $\boldsymbol{Q_2^k}$, respectively.

9) If $\overline{D}^{k+1} > \overline{D}^{k}$, sample $\boldsymbol{Q_2^k}$ is replaced by sample $\boldsymbol{Q_2^{k+1}}$ as the new “best” sample. Else, sample $\boldsymbol{Q_2^k}$ is replaced by sample $\boldsymbol{Q_2^{k+1}}$ with probability = $\exp\{(\overline{D}^{k+1} - \overline{D}^{k})/t^k\}$.

10) $t^k = \alpha \times\ t^{k-1}$

11) Stop when $t^k < \varepsilon$.

The final sample of $\boldsymbol{Q_2}$ at the last iteration is then selected as the optimal sample to be phenotyped in Stage 2.

The algorithm is programmed in R (version 2.11.0; www.r-project.org). The parameters used for the runs of the SA algorithm described herein were as follows: t^0=400, α=0.9999, ε=0.000001. The impacts of values assigned to these parameters were explored during the course of the study.

Data simulations

Two-stage approach using genotyped data

We assumed a normally distributed trait with mean 0 and variance 1 that is associated, in an additive fashion, with an "index" marker (a SNP scored 0, 1 or 2 for number of copies of the minor allele). We considered a range of allele frequencies [see below] for the index marker. For each simulation condition, we assigned a minor allele frequency (MAF) for the index marker and a corresponding effect size β for the index marker given its' MAF based on our assumption that the proportion of total variation explained by the associated index marker with respect to MAF is constant. We assumed that the MAFs for the remaining ($\boldsymbol{m2} - 1$) non-index markers (important for Stage 2) follow a uniform distribution within the range from 0.05 to 0.5 and that these markers are not associated with the trait. Genotype data for each genotyped marker was simulated using the binomial distribution given their corresponding MAF assuming marker genotypes follow Hardy-Weinberg equilibrium. The phenotype data were simulated, using random draws from a normal distribution, conditional on the genotype at the index marker and its corresponding β. Statistical tests of significance (testing H_0: β = 0 vs H_A: β ≠ 0) were performed using *t*-tests implemented in R.

We performed 500 simulations for each model we tested. For each simulation, we first simulated the index marker data and trait, conditional on the index marker, on the $\boldsymbol{n1}$ subjects phenotyped in Stage 1 and then assessed the significance of the index marker using a standard *t*-test. If this p-value was less than $\boldsymbol{pi.marker},$ we then we applied the SA-based approach to select $\boldsymbol{n2}$ subjects to be phenotyped in Stage 2 from the remaining $\boldsymbol{n.geno - n1}$ subjects, so that $\frac{\sum_{j=1}^{m2}\sum_{i=1}^{n2}(x_{ij}-\bar{x}_{\cdot j})^2}{m2}$ is maximized across the $\boldsymbol{m2}$ Stage 2 selected markers; otherwise, we randomly selected $\boldsymbol{n2}$ subjects to be phenotyped in Stage 2 from the remaining $\boldsymbol{n.geno - n1}$ subjects (Stage 2 sample selection is assumed to be random with respect to the index marker if that marker is not included in the SA optimization). Phenotypes were then simulated for the $\boldsymbol{n2}$ Stage 2 participants conditional on their genotype at the index marker.

We considered two previously described alternative strategies for evaluating the significance of our two-stage findings (180), namely the "joint-analysis" and the "replication-based analysis" approaches. For the joint-analysis approach, after selecting $\boldsymbol{n2}$ Stage 2 subjects, the index marker genotype-phenotype data were combined across all subjects from Stages 1 and 2 and we calculated the significance of the association between the trait and the index marker using the final sample of $\boldsymbol{n1 + n2 \ = \ n.pheno}$ subjects. The overall power for the joint-analysis design was defined by the proportion of simulations (out of 500) that achieve a $p<5\text{x}10^{-8}$ (a standard GWAS significance threshold that is likely conservative in this setting of evaluating only directly genotyped markers). For the replication-based analysis design, power was defined as the proportion of simulations where 1) the index marker was selected for inclusion in Stage 2 AND 2) the index marker achieved a $p<0.05/\boldsymbol{m2}$ using only the $\boldsymbol{n2}$ Stage 2 samples. Power for both approaches was compared to the power obtained from a single-stage study based on randomly selecting all $\boldsymbol{n1 + n2}$ samples for phenotyping. To reduce random noise in the comparisons, for each simulation, the $\boldsymbol{n1}$ samples in the joint-analysis, replication-based analysis and random sample approaches were the same $\boldsymbol{n1}$ samples.

In order to investigate how the power of the two-stage designs depends on $\boldsymbol{pi.sample}$ and $\boldsymbol{pi.marker}$, and to determine good choices for $\boldsymbol{pi.sample}$ and $\boldsymbol{pi.marker}$ for index markers with different MAFs, we calculated the power for the different study designs using various combinations of parameters: MAFs [0.05, 0.1, 0.2, 0.4] × $\boldsymbol{pi.marker}$ [0.00001, 0.000025, 0.00005, 0.0001, 0.0002, 0.0005] × $\boldsymbol{pi.sample}$ [0.375, 0.5, 0.625, 0.75, 0.875]. We assumed that $\boldsymbol{n.pheno/n.geno}$ (combined # subjects phenotyped in Stage 1 and 2 / # number of total subjects with genotype data) = 2,000/4,000 in all situations. We further assumed that the proportion of total variation explained by the associated index marker was independent of MAF and equal to a constant value of 2%.

Two-stage approach using imputed genotype data

Genotype imputation is a widely used zero-genotyping-cost method that can increase both genomic coverage and power to detect genetic associations (181). In addition to MAF and effect size, statistical power for an imputed index marker is also a function of the genotype imputation quality for the variant. Unlike standard power analyses for single variants, for our selective-phenotyping approach, the power for the index SNP is also a function of the MAF and imputation quality of the other $\boldsymbol{m}$**2-1** non-index Stage 2 selected markers, as these features will impact which samples are ultimately selected for phenotyping in Stage 2.

For the index marker, we calculated statistical power for the joint-analysis design (using significance criterion $p<5x10^{-8}$), with different combinations of parameters: MAFs [0.05, 0.1, 0.2, 0.4] $\times$ $\boldsymbol{pi.sample}$ [0.625, 0.75, 0.875] $\times$ R^2 [0.95, 0.80, 0.65, 0.50], and compared these results to what would be obtained using a random phenotype sample selection approach. We set the parameter $\boldsymbol{pi.marker}$ =0.00005 and, similar to the previous simulations for genotyped markers, assumed the true marker genotype explained 2% of the total trait variation. For the index marker, we generated probabilities $p_{i,0,}$ $p_{i,1,}$ and $p_{i,2}$ of genotype AA, Aa, and aa, respectively, for each individual *i* using the Dirichlet distribution given the MAF and corresponding R^2. We calculated allelic dosage for individual *i* at marker *j* as $D_{ij} = p_{i,1,} + 2 \times p_{i,2}$. For each subject, we randomly drew the true value of the index marker genotype, from the posterior probabilities for the three possible genotypes (AA/Aa/aa), and simulated the phenotype data y_i using a normal distribution with the mean conditional on the assigned index marker genotype.

For all other markers included in the Stage 2 selection process, we did not assume values of MAF or R^2 (imputation quality). Instead, we sampled these values from their approximate empirical distributions based on Hapmap Phase II imputed data from the Jackson Heart Study (JHS). We first made a histogram of the distribution of MAFs of HapMap Phase II imputed data from the JHS and fitted a high-ordered spline function (*f*) that reasonably fitted the observed distribution (**Figure 5.2**). We then sampled MAFs under *f*, using a Monte Carlo method, to obtain the MAFs for the non-index Stage 2 markers. To be specific, each time we drew a random number $\boldsymbol{x}$ from uniform distribution within the range from 0.005 to 0.5 (the range of MAF in JHS HapMap Phase II imputed data), and a random number *y* from the uniform distribution with bounds ranging from 0 to the maximum value that can be achieved by *f*. If the point (*x*,*y*) was above the spline of *f*, we rejected the marker; if the point (*x*,*y*) was under the spline, we accepted the marker in our sample.

Imputation quality varies as a function of MAF, with less common markers having poorer imputation quality on average than more common markers. We noted that the distribution of R^2 appeared to follow a $Beta$ distribution for markers within a given MAF range. To estimate this conditional $Beta$ distribution, we estimated the mean (μ) and variance (ϑ) of the imputation quality R^2 for a given MAF. To accomplish this, we created a fine grid of consecutive MAF bins and for each bin we obtained the mean and variance of R^2 values within the bin. We then fitted higher-ordered splines to create functions of the mean and variance for a given MAF. We assumed, within each MAF bin, that R^2 follows a $Beta$ distribution with parameters a and b, and thus, $\mu(R^2) = \frac{a}{a+b}$ and $\vartheta(R^2) = \frac{ab}{(a+b+1)(a+b)^2}$. By inverting the dependent and independent variables in these two functions we solve for the parameters $a = \mu^2(\frac{1-\mu}{\vartheta} - \frac{1}{\mu})$ and $b\ = (\frac{1}{\mu} - 1)\mu^2(\frac{1-\mu}{\vartheta} - \frac{1}{\mu})$ for the corresponding Beta function. For a given selected marker we then randomly generated a R^2 from the $Beta$ distribution Beta (*a,b*) conditional on the marker's MAF. To assess the overall fit of our $Beta$ model, we overlayed the curve generated by the density function of the $Beta$ distribution Beta (*a,b*) on the corresponding empirical density of R^2 values (histograms) for a range of MAF bins (**Figure 5.1**). Based on these results, it appeared our model was reasonable

………We then performed the Stage 2 SA sample selection model on allelic dosage D_{ij} for the $\boldsymbol{m2}$ Stage 2 markers to select the subset of subjects which maximizes the value of $\frac{\sum_{j=1}^{m2}\sum_{i=1}^{n2}(D_{ij}-\overline{D}_{\cdot j})^2}{m2}$. The association p-value for the index marker *j** was calculated by regressing the simulated phenotype y_i on the allelic dosage D_{ij*}.

Example: C-Reactive Protein (CRP) in the Jackson Heart Study (JHS)

We evaluated the performance of our two-stage selective-phenotyping approach in a study with real data. JHS is a longitudinal, population-based cohort study that aims to investigate cardiovascular disease risk factors in African Americans from Jackson, Mississippi(113). The design, recruitment and initial characterization of this study was described in detail elsewhere(114). We first performed a GWAS on a total of 2987 JHS subjects with available CRP phenotype data and genome-wide genotype data imputed based on 1000 Genome Project panel. The levels of CRP were naturally log-transformed to approximate normality of residuals after accounting for age, sex and BMI. Extreme values of CRP (>100) were removed as the extreme values suggested acute infection in those subjects. Thirty-eight million markers were imputed, using MACH 2.0 (118), based on a reference panel consisting of the complete sample of the 1000 Genome Project participants (Nov 2010, Version 3); only markers with MAF > 0.05 and estimated imputation quality of $R^2 > 0.3$ were included in further analyses. The association between CRP and imputed markers were tested using multivariable linear regression models in MACH2QTL v.1.08 (118), adjusting for age, sex, BMI, smoking status (yes/no), and the first 10 principal components generated from EIGENSOFT (120). An additive mode-of-inheritance model was assumed for genotype; β coefficients, representing the estimated change in transformed trait value associated with each additional copy of the effect allele, and the corresponding standard errors were reported.

We next considered the scenario where we could only afford to measure CRP in 1500 JHS subjects in total. Based on the results of simulated data, we used the parameters setting as $\boldsymbol{pi.sample}$ =0.75 and $\boldsymbol{pi.marker}$ =0.000025. Specifically, we first randomly selected 1125 (=1500×0.75) subjects in stage 1 and performed a GWAS on this subset of subjects. Due to the high coverage of 1000 Genome Project panel and linkage disequilibrium (LD) structure among markers, we selected only the top marker from each of the top 25 loci instead of the top 25 individual markers (i.e. we selected only one SNP at each locus, based on physical location, so that the final selected 25 markers were not likely to be in high LD with each other),identified in Stage 1, to be followed up and included in SA phenotype-selection process in Stage 2. In Stage 2, we identified the 375 subjects, from the remaining 1862 (=2987-1125) JHS subjects available for phenotyping, that maximized the average genotypic dispersion across these 25 selected markers. Next, we combined the imputed genotype and phenotype data from the 1125 subjects in Stage 1 and the 375 subjects selected in Stage 2 to assess whether there was evidence of association for the same loci found at $p<5x10^{-8}$ in the complete JHS sample. In comparison, we used the same

1125 subjects in Stage 1 combined with 375 randomly selected JHS subjects from Stage 2 and made the same assessments. We performed 100 different trials, each using a different random Stage 1 sample, and calculated the number of times each established GWAS signal reached a p-value of 5x10^{-8}, 5x10^{-7}, and 5x10^{-6} using either of the two approaches.

Results

Two-stage design using directly genotyped markers

Our results showed we can get modestly increased power using a two-stage joint-analysis selective-phenotyping study over studies based on a random selection of subjects. In general, estimated power was greatest when including large samples in Stage 1 (higher $\boldsymbol{pi.sample}$) combined with higher selectivity of markers (lower $\boldsymbol{pi.marker}$) for inclusion in the Stage 2 SA phenotype-selection algorithm, where the relative increase in power approached 10% compared to using unselected samples. Power comparisons between our two-stage selective-phenotyping approach, using joint analyses, and the conventional single-stage random sample approach are shown in **Figure 5.3** and **Figure 5.4** across alternative values of $\boldsymbol{pi.sample}$ and $\boldsymbol{pi.marker}$, respectively. The relationship between the relative increase in overall power [(overall power achieved by SA - overall power achieved by random selection)/overall power achieved by random selection] and $\boldsymbol{pi.sample}$ across different choices of index marker MAF and $\boldsymbol{pi.marker}$ are shown in **Figure 5.5**. In most scenarios, the maximum benefit in power gained by our two-stage joint-analysis approach over random selection was achieved using higher values of $\boldsymbol{pi.sample}$ (0.75, 0.875). In most scenarios, the minimum benefit was achieved for markers with higher MAF (0.4). We also investigated the relationship between the relative increase in power and $\boldsymbol{pi.marker}$ for different scenarios of the index marker MAF and

$\boldsymbol{pi.sample}$ (**Figure 5.6**). No obvious pattern between relative increase in power and $\boldsymbol{pi.marker}$ was observed for lower $\boldsymbol{pi.sample}$ values, but a strong trend was observed when $\boldsymbol{pi.sample}$ was as high as 0.875 (a lesser trend was observed for $\boldsymbol{pi.sample}$ = 0.75) for lower power being associated with increasing values of $\boldsymbol{pi.marker.}$ The power achieved by the two-stage replication-based study design was considerably worse than both random sample selection and the two-stage joint-analysis design (data not shown), despite the phenotypic sample enrichment in Stage 2. Thus, we dropped the two-stage replication study design from further consideration.

Two-stage replication study using imputed genotype data

Our results using imputed genotype data were consistent with the results from the directly genotyped results, showing we can achieve increased power using the two-stage joint-analysis study over studies based on random selection of subjects. Not surprisingly, the relative power of an imputed index marker is markedly lower than when using a directly genotyped index marker when the index marker has low imputation quality. However, the relative gains or our proposed approach compared to using randomly phenotyped samples still exists even for poorly imputed index markers. The overall power depending on $\boldsymbol{pi.sample}$ is shown in **Figure 5.7.** No obvious pattern between relative increase in power and $\boldsymbol{pi.sample}$ or MAF was observed, suggesting that our approach is not overly sensitive to parameter choices and that its performance using imputed data is robust across a wide range of alternative scenarios.

C-Reactive Protein (CRP) in the Jackson Heart Study (JHS)

The top markers that reached genome-wide significance in a total of 2987 JHS subjects are listed in **Table 5.1.** Among the five signals, three of them (*CRP*, *APOE*, and *HNF1A*) were established GWAS signals, and thus we tested how many times (out of 100 random trials) each of the three established GWAS signal reached a *P* value of 5x10^{-8}, 5x10^{-7}, 5x10^{-6}, 5x10^{-5}, and 5x10^{-4} using our two-stage selective-phenotyping approach and a random phenotyping selection approach (**Table 5.2**) that included a total of n=1500 JHS participants. The *CRP* locus signal was detected at a genome-wide significance level (P<5x10^{-8}) in all 100 trials using both approaches. The *APOE* signal was detected at genome-wide significance level in 29 out of 100 trials using the selective-phenotyping approach and only 10 times using the random selection approach. The *HNF1A* signal was detected using a genome-wide significance threshold only once using both approaches, and the performance of both approaches were similar when using less stringent significance levels.

Discussion

Two-stage genetic association studies have played an important role in reducing the cost and increasing the power of genetic association studies. Historically, two-stage designs have involved genotyping a subset of phenotyped subjects (in Stage 1) on 100,000’s of markers contained on a large-scale genotyping array and then performing targeted genotyping on the remaining phenotyped subjects (in Stage 2) on a subset of the markers that demonstrated the greatest evidence for association in the Stage 1 participants. Here, we proposed an alternative two-stage genetic association study design, where it is assumed that we have an unmeasured quantitative trait of interest but that all subjects have available directly measured or imputed genome-wide genotype data. We described a selective-phenotyping method that uses SA to identify the optimal subset of subjects to be phenotyped in Stage 2 of a two-stage GWAS based on their available directly genotyped and/or imputed dosage data across markers that had the strongest evidence for association in a subset of randomly selected, newly phenotyped, Stage 1 participants. The method is particularly designed for the scenario where cost, or other, constraints prohibit the phenotyping of the entire cohort of participants that have available genetic data. Through both simulations and a real example, we show that our two-stage selective-phenotyping approach can increase the power to identify significant associations between newly measured quantitative-trait phenotypes and either directly or imputed genotyped markers compared to studies that randomly select subjects for phenotyping.

We proposed to use a computationally efficient method based on SA to find the maximal genotypic dissimilarity across multiple selected markers of interest, rather than simpler methods that only consider a single marker at a time. In comparison with other optimization algorithms, SA does not require calculation of derivatives and subsequent root finding, which can be intractable in many complicated settings. The SA optimization algorithm, as described, does not make any assumptions regarding Hardy-Weinberg equilibrium or marker independence and is remarkably flexible. While not described herein, weights can be readily included for certain markers, if desired, to increase the influence of these markers on sample selection. Along these lines, our two-stage approach could include prior information (e.g. force including a previously reported associated marker in the optimization scheme regardless of Stage 1 results) to influence phenotype selection in the remaining samples. We recommend considering LD pruning top markers prior to marker selection for Stage 2, as failure to do so will result in some regions having disproportionate influence on the phenotype selection procedure. An investigator might also choose to optimize alternative functions depending on their underlying hypothesis. For example, if an investigator wants to assume a rare-recessive mode-of-inheritance model then genotype could be rescored 0 (non-carrier) or 1 (carrier of at least on risk allele), rather than our described values of 0, 1 and 2 based on assuming an additive inheritance model, prior to conducting the SA optimization.

We attempted to identify reasonable parameter choices for selecting samples, chosen from a larger cohort, for phenotyping (for both Stage 1 and Stage 2 samples) to provide greater power than a simple random sample selection for phenotyping. The balance is between selecting a high enough proportion of subjects in Stage 1, so that the truly associated markers rise to the top of the results (so they can be included in sample selection in Stage 2), while leaving enough remaining subjects for phenotyping in Stage 2, where we can see some benefit of performing selective phenotyping for these markers. Interestingly, our approach was not overly sensitive to parameter choices and remained robust across a wide range of MAFs, $\boldsymbol{pi.marker}$, and $\boldsymbol{pi.sample}$. In general, though some benefit of increased power was observed across most combinations of $\boldsymbol{pi.marker}$, and $\boldsymbol{pi.sample}$, it appears that using a larger sample selection in Stage1 and a stricter marker inclusion in Stage 2 provides the best results.

Unlike more traditional two-stage genotyping designs, all markers analyzed initially in Stage1 will also be evaluated in the completed Stage 1 + Stage 2 samples when using our two-stage selective-phenotyping approach. All Stage 1 + Stage 2 samples will have measured genotype and phenotype data available for analyses. For the markers not included in the Stage 2 optimization procedure, there would be little expected impact on their respective parameter estimates or estimated power estimates compared to what would typically be obtained from using random sample phenotype selection for the remaining Stage 2 samples, assuming that these markers are reasonably independent of those markers included in the Stage 2 optimization.

The clear advantage of our approach is the increased power to detect novel associations compared to a random sampling design under the same fixed sample size constraint. It is important to note that our sample procedure does not directly cause the effect (β) estimates to be biased. Unlike the extremes-of-phenotype selective-genotyping study design, our procedure does not inherently change the trajectory of the effect estimate under the alternative hypothesis, rather, our increase in power is obtained by increasing the precision (shrinking the variance) of the index marker effect estimator. However, the effect estimates of the best marker results (including many of the markers which would likely be used in the Stage 2 phenotype selection) using our approach would typically be inflated, per the winner's curse phenomenon, just like they would be for any other large-scale discovery study, including studies that use random sample phenotype selection. Finally, while we describe our approach for measuring a new quantitative trait, we note that the general two-stage selective phenotyping approach should also be considered when measuring new dichotomous traits, where the probabilities of the different outcomes are a function, under the alternative hypothesis, of the marker genotypes (e.g.. we would expect to increase the proportion of cases by selectively genotyping uncommon homozygotes for an index marker).

There are also some possible limitations in the proposed two-stage study design that warrant consideration. First, our two-stage phenotyping design requires additional time to phenotype samples (phenotyping is performed in two batches, with the second batch conditional on the association results from the first batch). Investigators will have to decide if time constraints are an important factor and if additional phenotyping costs occur due to the piecemeal approach. Second, phenotyping subjects in each stage under different situations may induce batch effects in the measures of phenotypes. Such batch effects could be remedied by the simple inclusion of an additional covariate for batch in the regression models or use of meta-analysis.

Table 5.1 Top markers that reached genome-wide significance ($p<5x10^{-8}$) in 2987 subjects

Chr	# of Markers	Most significant SNP	Pos(hg19)	EA	EAF	RSQR	N	β	SE	P	Function	*Gene*
1	111	**rs73024710**	159689965	T	0.17	0.99	2978	0.44	0.04	1.19E-30	intergenic	***APCS-CRP-DUSP23***
19	10	**rs10119**	45406673	A	0.25	0.56	2978	-0.33	0.05	2.98E-13	UTR3	***TOMM40-APOE***
7	1	rs12698712	68106789	G	0.10	1.00	2978	-0.27	0.05	2.30E-08	intergenic	*STAG3L4-AUTS2*
12	1	rs11047572	24836927	A	0.26	0.88	2978	0.19	0.04	3.38E-08	intergenic	*LINC00477-BCAT1*
12	1	**rs1169284**	121419926	C	0.25	0.90	2978	-0.19	0.04	3.61E-08	intronic	***HNF1A***

Table 5.2 Times each of the 3 established GWAS signal (*CRP*, *APOE*, and *HNF1A*) reached a *P* value of $5x10^{-8}$, $5x10^{-7}$, $5x10^{-6}$, $5x10^{-5}$, and $5x10^{-4}$ using either approach (SA/random)

Chr	Gene	SNP	# P<$5x10^{-8}$	# P<$5x10^{-7}$	# P<$5x10^{-6}$	# P<$5x10^{-5}$	# P<$5x10^{-4}$
1	*APCS-CRP-DUSP23*	rs726640	100/100	100/100	100/100	100/100	100/100
19	*TOMM40-APOE*	rs1160985	29/10	44/34	65/60	90/90	100/100
12	*HNF1A*	rs1169284	1/1	8/4	16/16	41/42	74/69

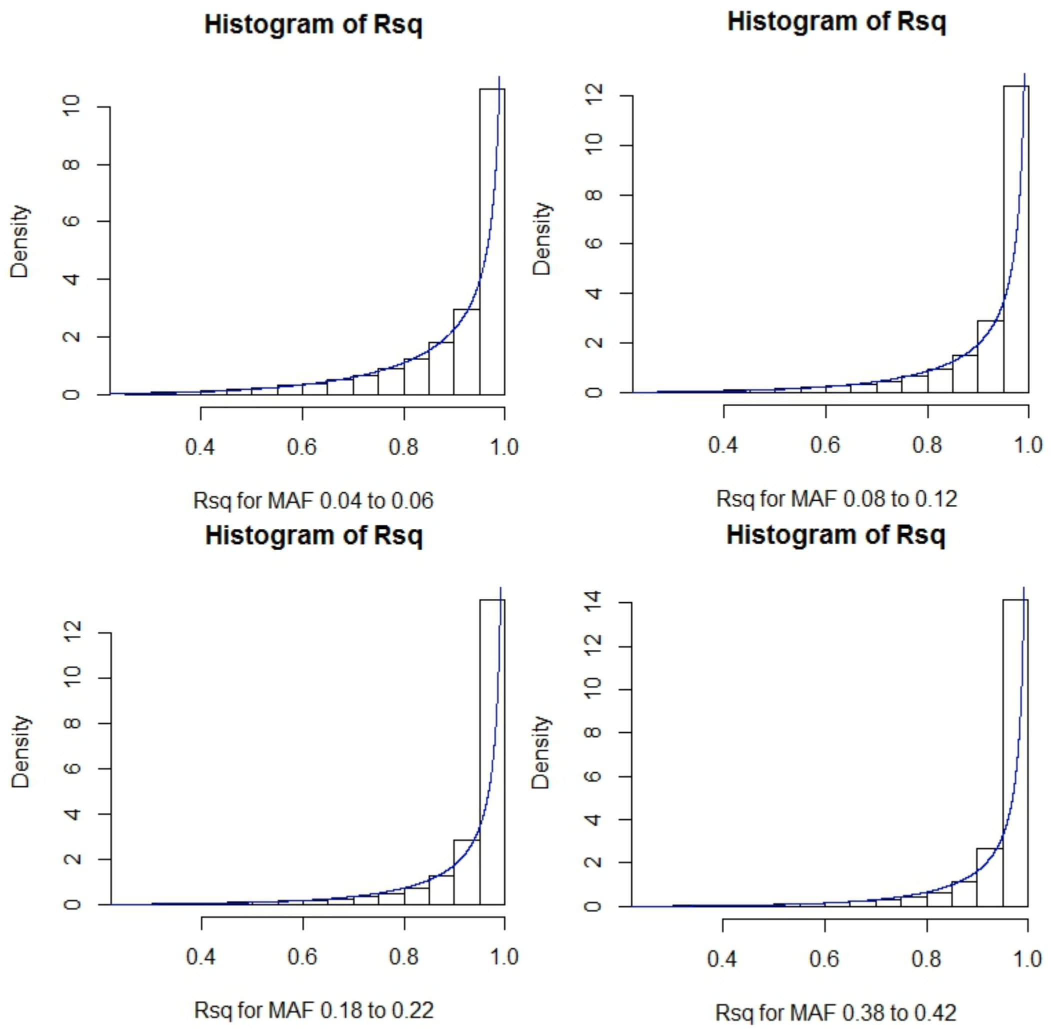


Figure 5.1: Histograms of R^2 across the spectrum of MAF bins and the curve generated by density function of Beta distribution

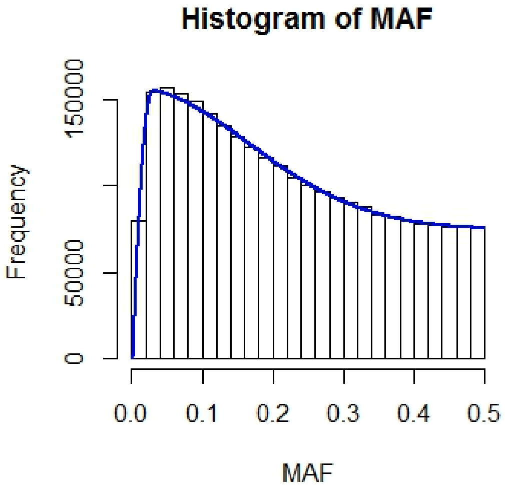
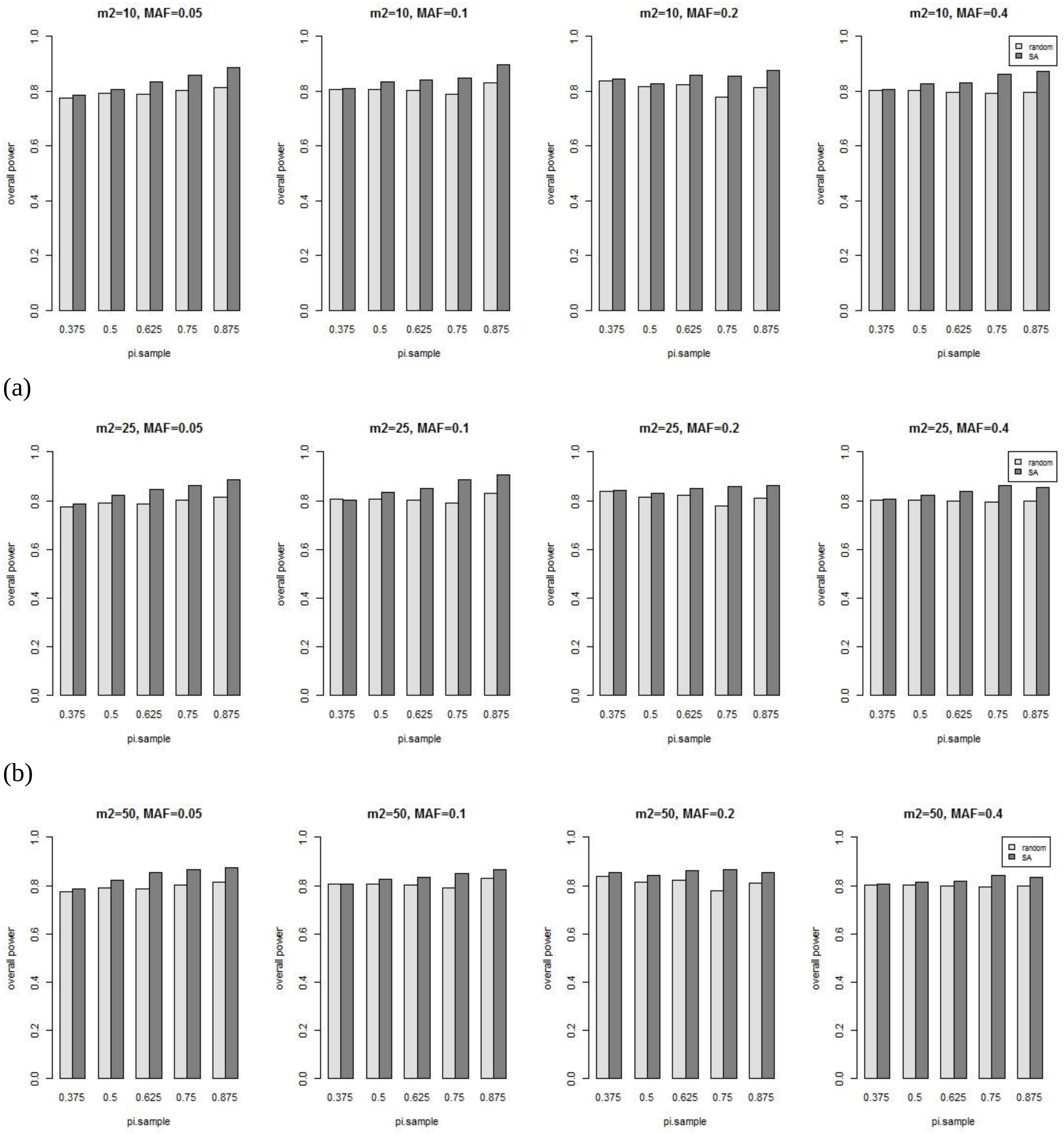


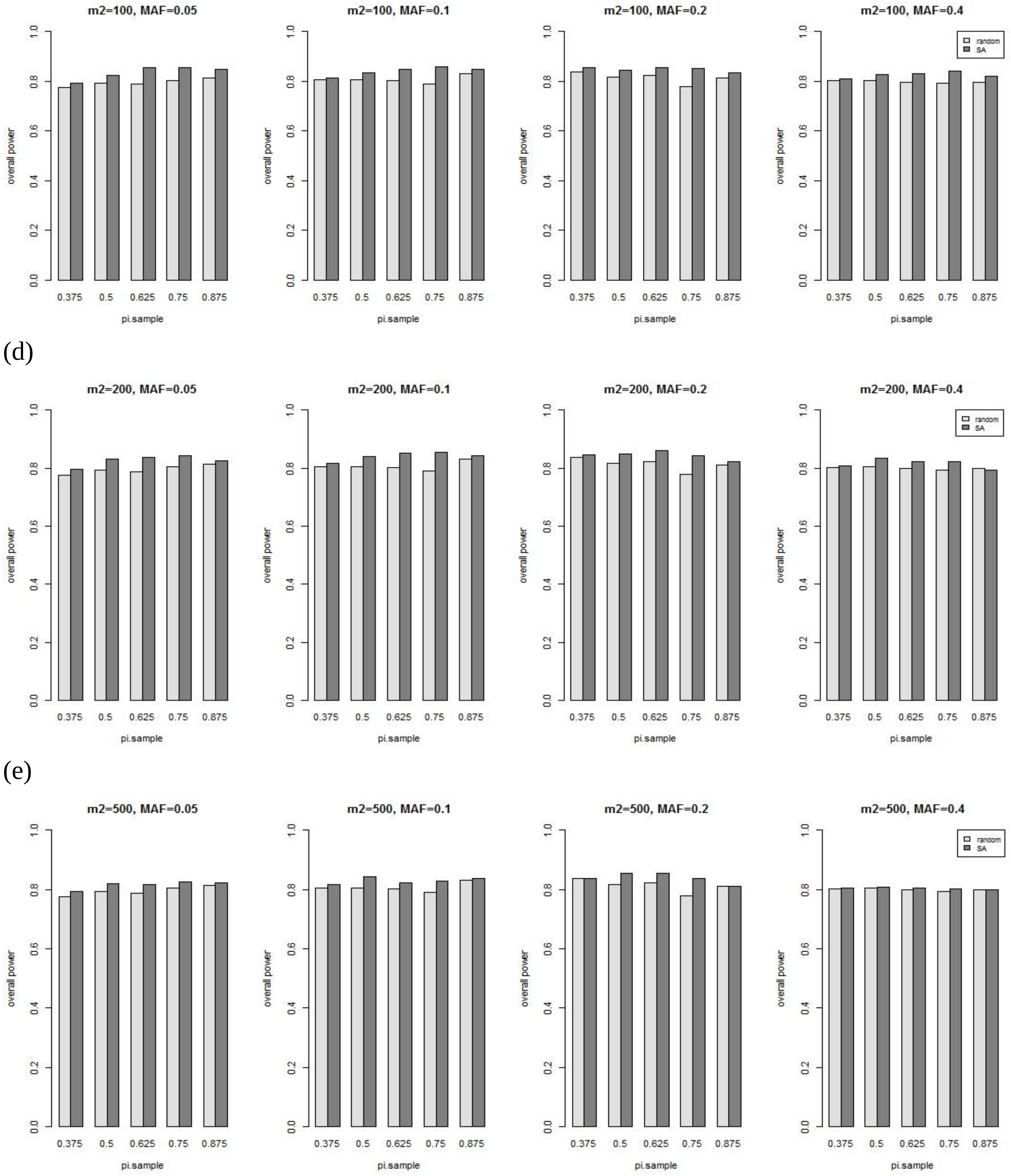
Figure 5.2: Histogram of MAF of HapMap Phase II imputed data from the Jackson Heart Study


m2=10, MAF=0.05
m2=10, MAF=0.1
m2=10, MAF=0.2
m2=10, MAF=0.4
m2=25, MAF=0.05
m2=25, MAF=0.1
m2=25, MAF=0.2
m2=25, MAF=0.4
m2=50, MAF=0.05
m2=50, MAF=0.1
m2=50, MAF=0.2
m2=50, MAF=0.4
overall power
0.0
0.2
0.4
0.6
0.8
1.0
0.375
0.5
0.625
0.75
0.875
pi.sample
□ random
▨ SA

(a)

(b)

(c)

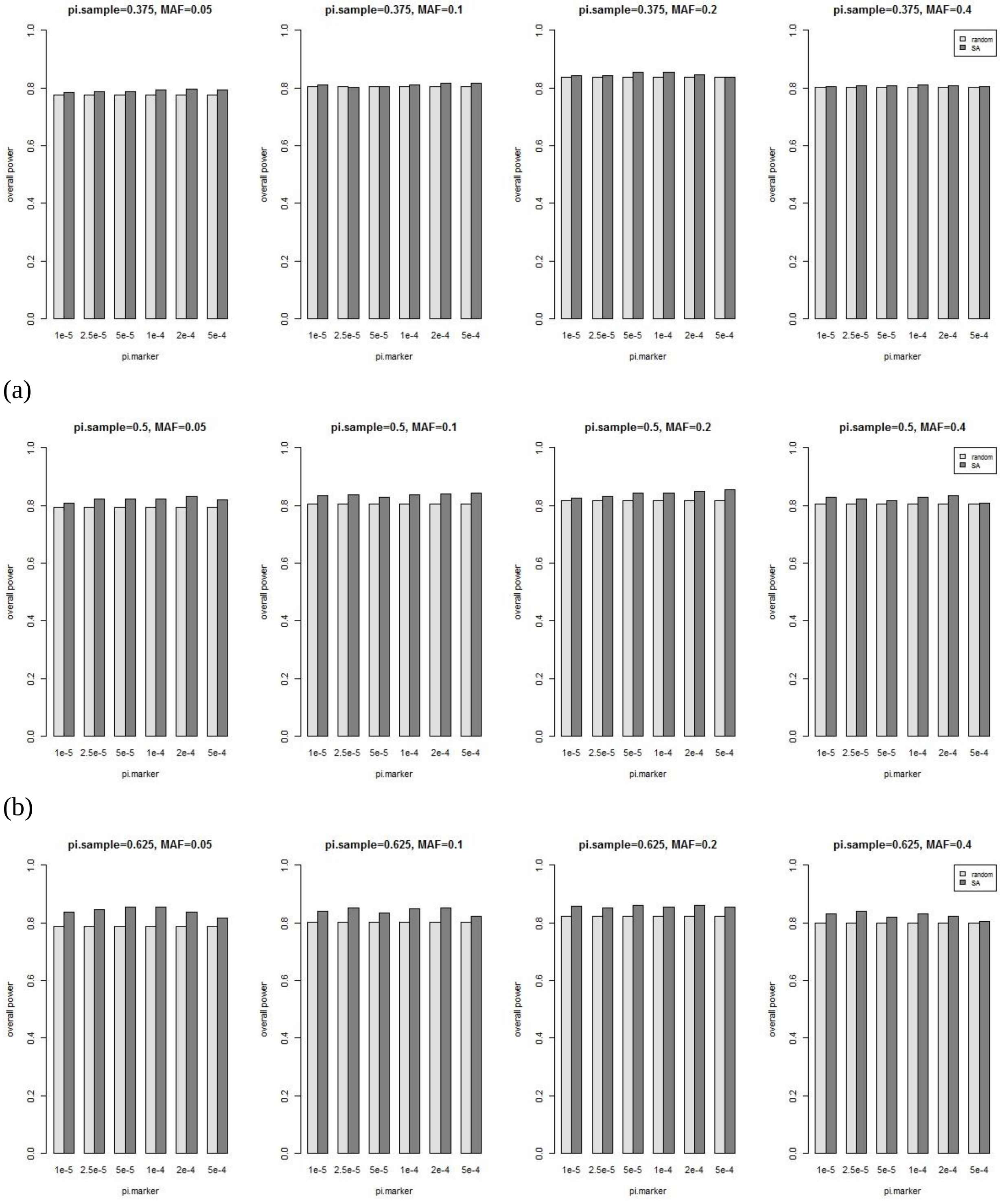


(d)

(e)

(f)

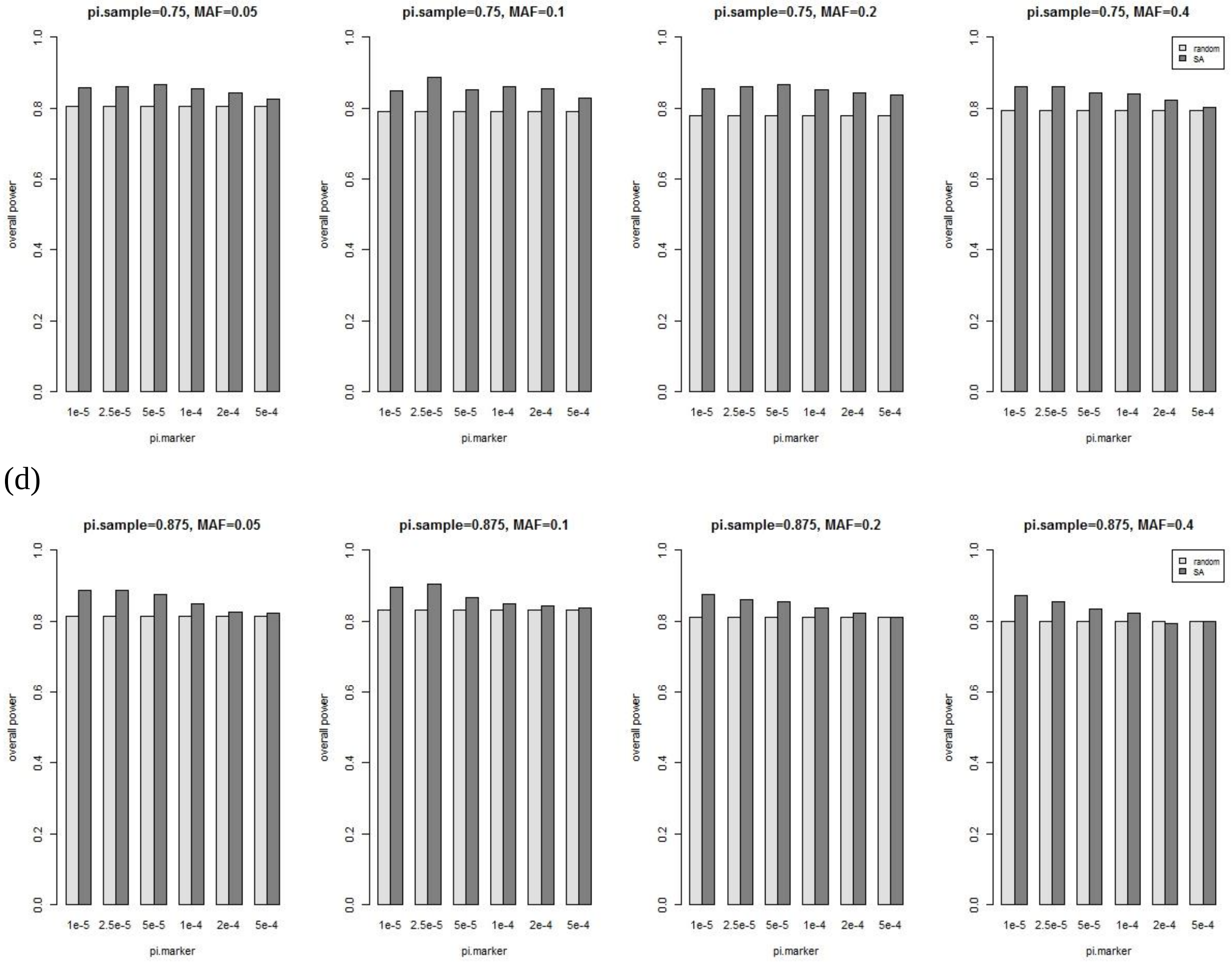
Figure 5.3 Power of a two-stage GWAS with a subset of subjects randomly selected or selected based on SA to be phenotyped in Stage 2 depending on $\boldsymbol{pi.sample}$. n.pheno=2000, n .geno=4000 a) pi.marker=0.000001. b) pi.marker=0.000025. c) pi.marker=0.00005. d) pi.marker=0.0001. e) pi.marker=0.0002. f) pi.marker=0.0005


pi.sample=0.375, MAF=0.05
pi.sample=0.375, MAF=0.1
pi.sample=0.375, MAF=0.2
pi.sample=0.375, MAF=0.4
pi.sample=0.5, MAF=0.05
pi.sample=0.5, MAF=0.1
pi.sample=0.5, MAF=0.2
pi.sample=0.5, MAF=0.4
pi.sample=0.625, MAF=0.05
pi.sample=0.625, MAF=0.1
pi.sample=0.625, MAF=0.2
pi.sample=0.625, MAF=0.4
overall power
0.0
0.2
0.4
0.6
0.8
1.0
1e-5
2.5e-5
5e-5
1e-4
2e-4
5e-4
pi.marker
□ random
■ SA

(a)

(b)

(c)



(d)

(e)

Figure 5.4 Power of a two-stage GWAS with a subset of subjects randomly selected or selected based on SA to be phenotyped in Stage 2 depending on $\boldsymbol{pi.marker}$. n.pheno=2000, n .geno=4000 a) pi.sample=0.375; b) pi.sample=0.5; c) pi.sample = 0.625; d) pi.sample = 0.75; e) pi.sample = 0.875

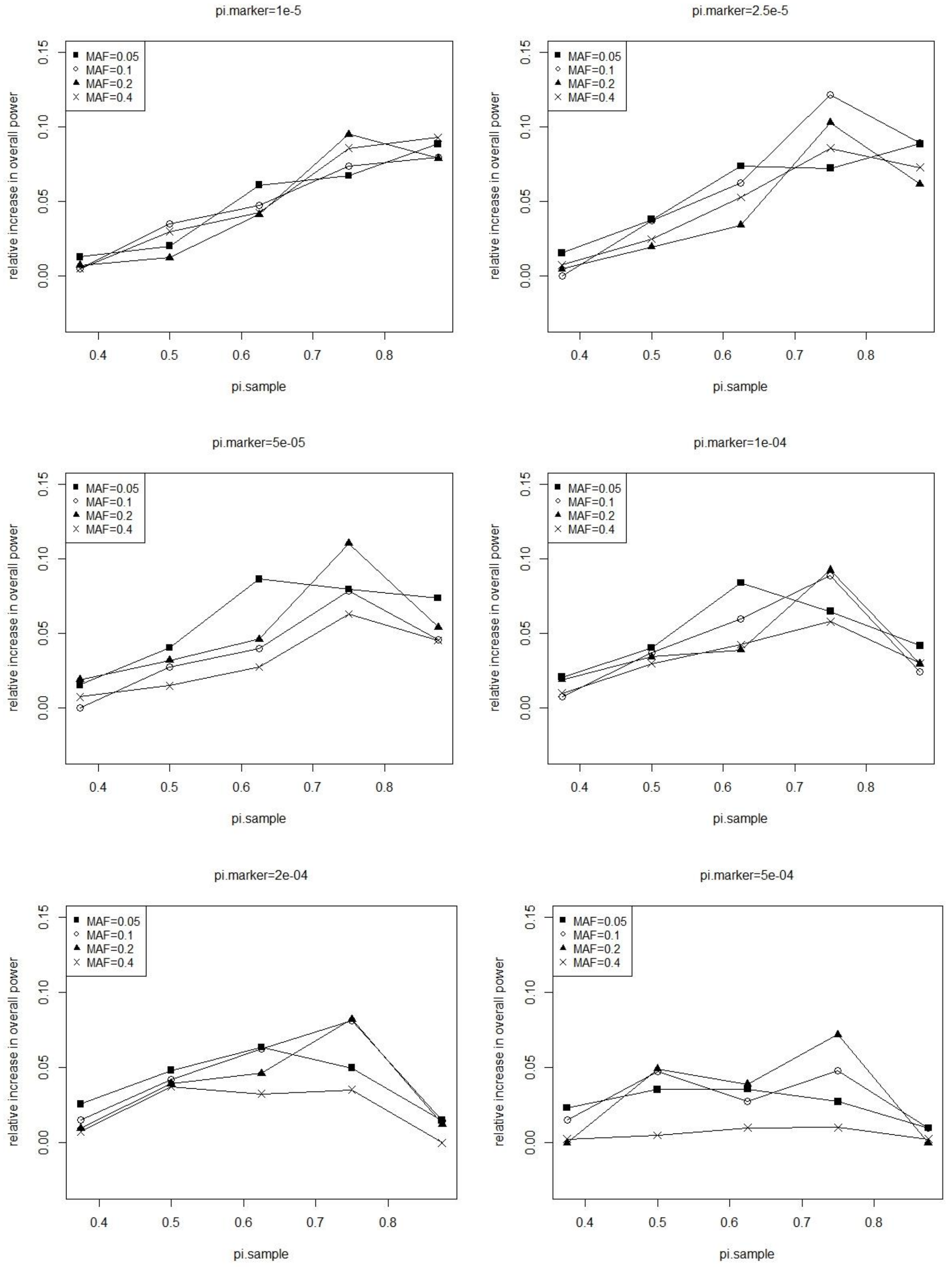


Figure 5.5 The relationship between relative power and $pi.sample$ for different MAF.

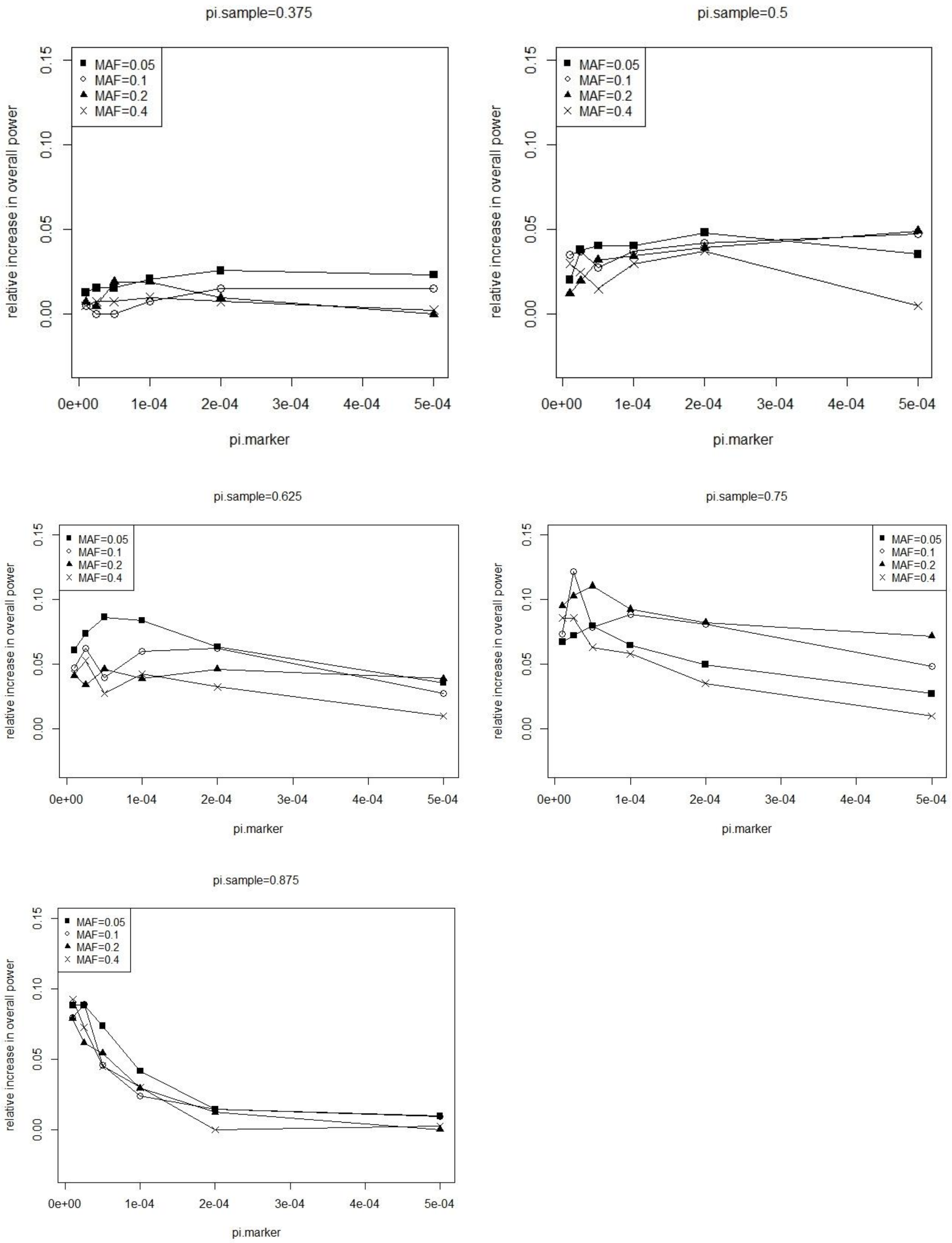
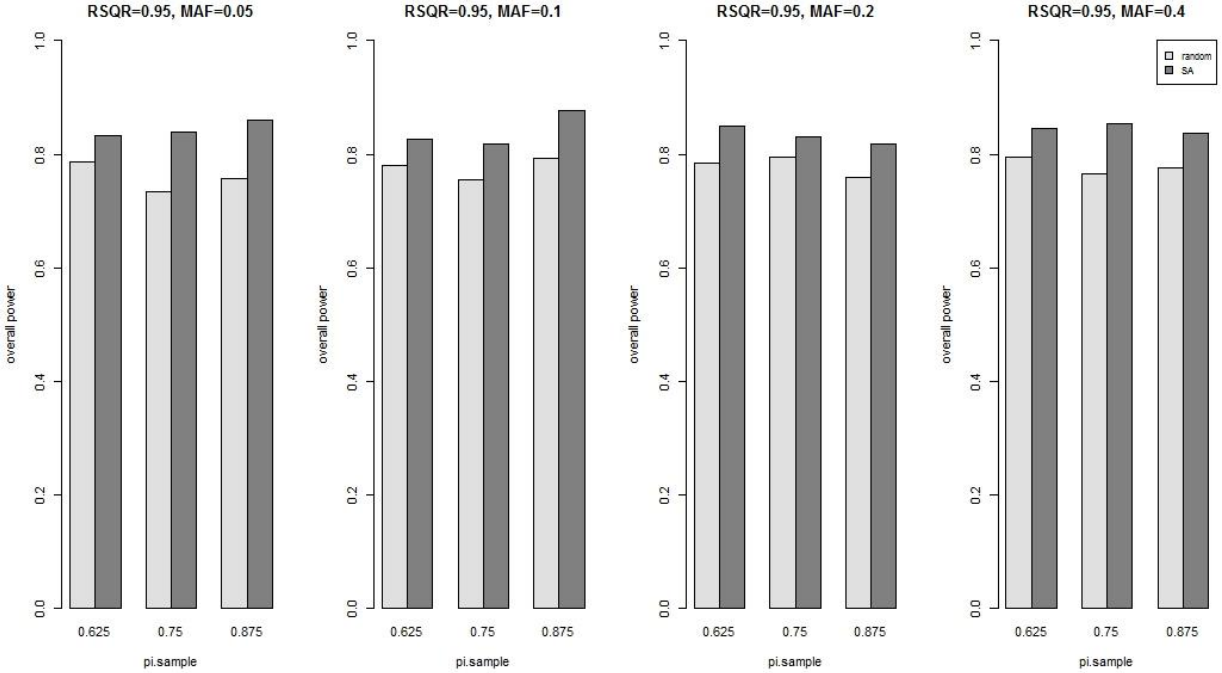
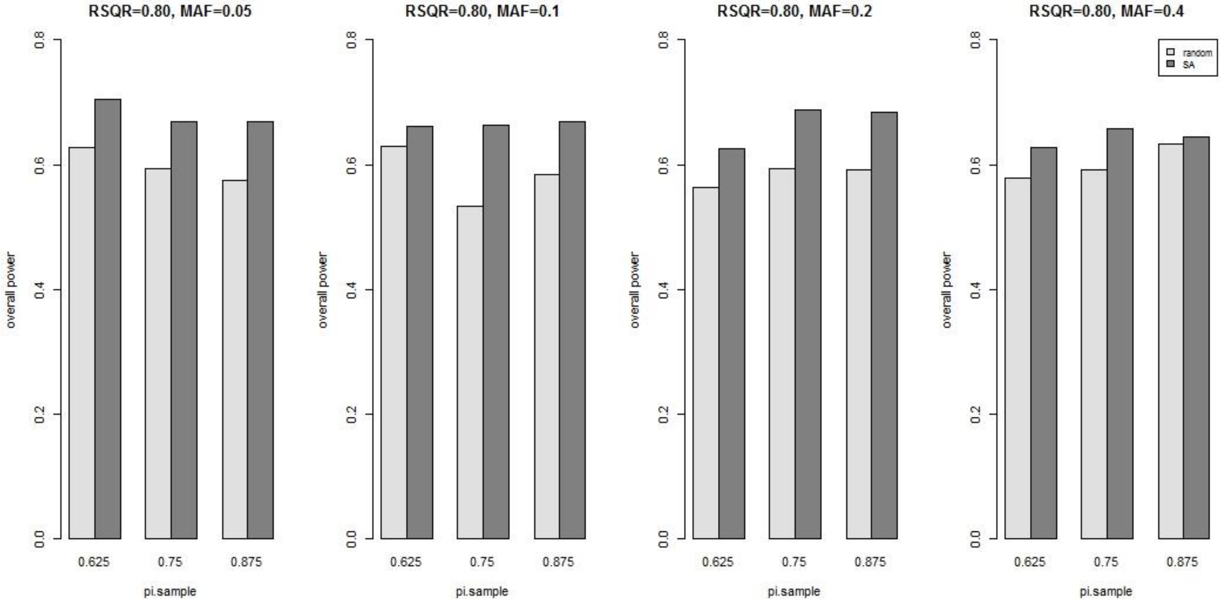


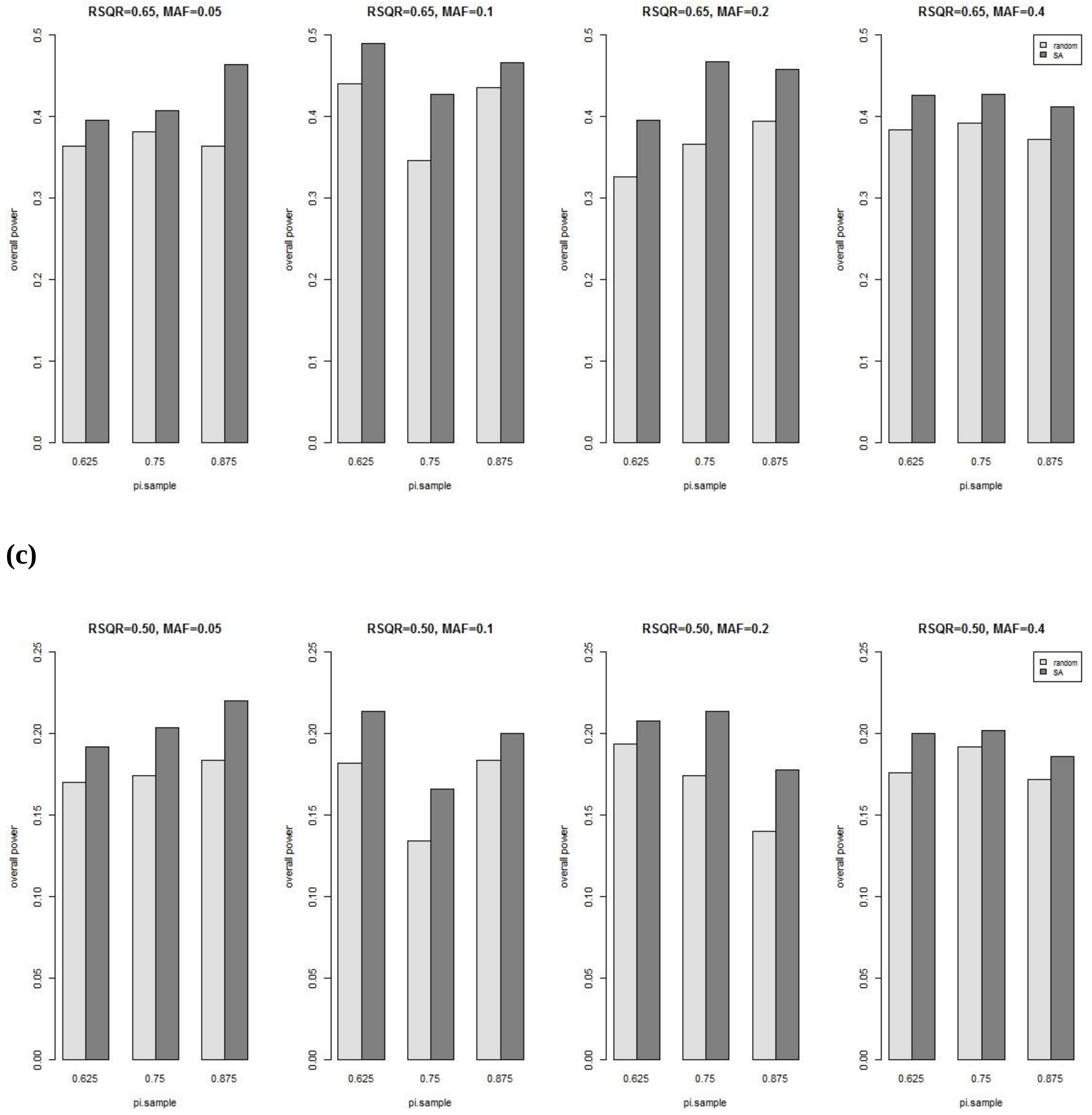
Figure 5.6 The relationship between relative power and $pi.marker$ for different MAF.


RSQR=0.95, MAF=0.05
RSQR=0.95, MAF=0.1
RSQR=0.95, MAF=0.2
RSQR=0.95, MAF=0.4
overall power
0.0
0.2
0.4
0.6
0.8
1.0
0.625
0.75
0.875
pi.sample
random
SA

(a)


RSQR=0.80, MAF=0.05
RSQR=0.80, MAF=0.1
RSQR=0.80, MAF=0.2
RSQR=0.80, MAF=0.4
overall power
0.0
0.2
0.4
0.6
0.8
0.625
0.75
0.875
pi.sample
random
SA

(b)



(c)

(d)

Figure 5.7 Power of a two-stage GWAS with a subset of subjects randomly selected or selected based on SA to be phenotyped in Stage 2 from a total sample of subjects with available imputed genotype. n.pheno=2000, n .geno=4000, pi.marker=0.000050 a) Imputation quality R^2=0.95; b) R^2=0.80; c) R^2=0.65; d) R^2=0.50.

CHAPTER VI: DISCUSSION

In the preceding chapters, I described a collection of findings that further our understanding of genetic architecture of two CVD related biomarkers in AAs who have a higher prevalence of CVD than other ethnic groups in the United States, and blood pressure measures in a Chinese population. I first conducted a genome-wide association study for iron-related phenotypes including serum iron, serum ferritin, SAT, and TIBC in 2347 AAs participating in the JHS. I observed a novel region on chromosome 3, *GAB3-G6PD*, significantly associated with ferritin levels and identified the putative causal variant in this region. I also observed multiple independent SNPs associated with TIBC in the *TF* region using conditional analyses. The two independent associations for TIBC at *TF* and the association for ferritin at *GAB3* were successfully replicated in HANDLS. Next, I conducted a genome-wide admixture and association study, and an exome-wide association study using Human Exome Beadchip for Lp(a) in 2895 AAs in JHS. I observed significant ($P<5x10^{-8}$) associations for hundreds of SNPs spanning an ~10Mb region on 6q surrounding the *LPA* gene. Interestingly, after adjusting for local ancestry, the region containing significantly associated SNPs became much narrower and was centered over the *LPA* gene (<1Mb). Gene-burden tests found significant associations between Lp(a) and aggregate collections of SNPs in *LPA* and *APOE*. Next, I tested the interaction between several environmental factors and four candidate genes in affecting blood pressure measures in 7319 Chinese in CHNS. I observed that the signal at rs1458038 in the *FGF5* region exhibited a significant genotype-by-BMI interaction affecting blood pressure, with the strongest variant effects in those with the highest BMI. This pattern of interaction was also observed in an independent sample of men from FAMHES. Finally, I described a selective phenotyping method that uses a simulated annealing (SA) algorithm to identify the optimal subset of subjects to be phenotyped in Stage 2 of a two-stage GWAS based on their available directly genotyped and/or imputed dosage data across multiple SNPs of interest, and showed that increased power could be achieved using this method.

Several aspects of this work are worth highlighting because they illustrated novel findings of genetic architecture of certain traits and shed some light on future directions. In Chapter II, we observed the estimated average (“global”) proportion of African ancestry was significantly associated with lower levels of TIBC, serum iron, and SAT, and nominally associated with higher level of serum ferritin - which are entirely consistent with previous findings reporting lower levels of TIBC, iron and SAT and higher level of ferritin, on average, in AAs compared to European Americans (96, 97). In Chapter III, I observed that a higher level of estimated global African ancestry was significantly associated with higher levels of Lp(a), which is also consistent with previous findings which reported that Lp(a) levels in populations of African ancestry are much higher (2~4-fold) than in populations of European ancestry (111). Interestingly, the observed higher global African ancestry associated with higher Lp(a) levels was entirely explained by higher African local ancestry surrounding the *LPA* gene. These results implicate novel genetic risk factors in AAs and underscore the importance of studying this population for genetic risk factors that uniquely/disproportionately impact them. The Jackson Heart Study (JHS) provides a unique opportunity to study the genetic basis of complex CVD related traits in AAs. The JHS is a longitudinal, community-based, observational study designed to identify risk factors for the development of CVD in more than 5000 AAs recruited from noninstitutionalized adults from urban and rural areas around the Jackson, Mississippi, metropolitan area (85). Jackson has the largest percentage (36.3%) of AAs in the United States, with the overall CVD mortality in AA men and women 12% and 22% higher, respectively, than in the rest of the nation. As the largest single-site, prospective, epidemiologic investigation of cardiovascular disease among AAs ever undertaken, JHS will continue show its value in exploring the reasons for this disparity of CVD prevalence between AAs and EA and lead to new approaches to reduce it.

Typically after establishing a GWA signal, an important follow-up analysis is to determine whether of additional independent signals exist in the associated region. The identification of multiple signals at individual loci could explain additional phenotypic variance ('missing heritability') of common traits, and help identify the causal genes or regulatory mechanisms. In my work from Chapter II, I identified a second significant independently associated SNP in the *TF* region for TIBC. Our report was the first to identify such a second signal. Interestingly, this second signal (index SNP rs9872999 which maps to an intergenic region approximately 10Kb proximal to *TF*) only became significant at the genome-wide level after conditioning on our top *TF* region SNP rs8177253. The allele associated with an increase in TIBC for rs8177253 is preferentially on the same haplotype with the allele associated with a decrease in TIBC for our top variant rs9872999. Thus, the mean effects for rs9872999 are shrunk towards the null when not factoring in genotype for rs8177253. The presence of multiple association signals at individual loci is an indicator of allelic heterogeneity, which is defined as the presence of multiple alleles that act through one gene to influence a trait. A recent study investigating the patterns of association consistent with allelic heterogeneity using *cis* gene expression phenotypes observed several patterns of association when analyzing single SNPs compared with multiple SNPs at the same loci (182). The frequency with which trait raising alleles segregated on the same or opposite haplotypes affected the degree to which association statistics ‘fell’ or ‘jumped’ in multivariable compared with univariable analyses. If two trait raising alleles segregated on the same haplotype, the association statistics tend to ‘fall’ in multivariable analyses because multivariable analysis will adjust for the correlated effect of the other primary SNP on the secondary SNP. If two trait raising alleles segregate on the opposite haplotype (i.e. one trait associated raising allele and the other trait associated lowering allele segregate on the same haplotype), the association statistics tend to ‘jump’ because a multivariable analysis will adjust for the cancelling out effect of the other SNP. The distinction between two independent signals and one partially tagged signal is rather important when trying to use association results to identify causal genes, or when choosing SNPs for functional studies.

Another important follow-up analysis after establishing a GWA signal is to identify the actual causal variants and to determine the underlying functional mechanisms. Through a variety of approaches, my work identified both common and uncommon variants which are potentially causal to the variation of CVD related phenotypes. In Chapter II, I observed a functional missense variant in *G6PD*, rs1050828 (MAF=0.13, leading to a Val68Met amino acid substitution), which was associated with ferritin but narrowly missed genome-wide significance ($p=9.1x10^{-8}$). This variant was imputed in the original GWAS data. The association between the top genome-wide significant variant, rs141555380, in this region and ferritin disappeared after adjustment for this functional variant (rs1050828) at *G6PD* (p = 0.55). This functional variant is implicated in malaria resistance, and the A- form of G6PD deficiency in Africa is under strong natural selection from the preferential protection it provides against life-threatening malaria (110). This is the first time *G6PD-GAB3* region has been reported to be significantly associated with ferritin, a specific iron-related measure in AA. In Chapter III, my work identified a splicing-altering variant rs41272114 in the *LPA* gene, a nonsynonymous variant rs7412 in the *APOE* gene and a second independent nonsynonymous variant rs769455 in the *APOE* gene that were significantly associated with Lp(a) level using data from the Ilumina Human Exome Beadchip. Although exome- and whole-genome sequencing is becoming increasingly affordable, genotyping arrays still remain to be a cost-effective approach to investigate rare variants in the human genome. The goal of this Human Exome Beadchip array was to enable an intermediate experiment between current genotyping arrays, which focus on relatively common variants, and exome sequencing of very large numbers of samples, which focus on rare and low-frequency coding variants. The Exome BeadChip contains 247,870 variants discovered through exome sequencing in ~12,000 individuals that are mostly protein-altering (nonsynonymous coding, splice-site and stop gain or loss codons) (64). So far, it has enabled identification of several rare and functional variants associated with fasting glucose, insulin processing, and type 2 diabetes susceptibility (65) (66). My work suggested custom genotyping array like Exome BeadChip can provide new insights for previously genotyped cohorts and enable the identification of functional variants in future genetic association studies. By following up thousands of targeted SNPs with prior evidence of association with related traits, custom targeted genotyping arrays will continue making contributions to discovering biologically functional variants and understanding potential mechanisms of disease pathogenesis.

A hot topic of debate in the field of genetic association studies is the “missing heritability” not explained by GWAS. It is believed that human complex traits are influenced by the interaction among genetic and environmental factors, thus investigation of such interactions will help explain some of the missing heritability and provide better insight into pathway mechanisms for complex diseases (67). In Chapter IV, I detected a significant interaction between rs1458038 5' of *FGF5* gene at 4q21 and BMI in affecting blood pressure measures in a Chinese population. Specifically, the effect of the rs1458038 risk genotype is considerably stronger in subjects in the heaviest quartile of the BMI distribution. This study implicates that the risk genotype at rs1458038 is particularly important in Chinese individuals with higher BMI. Candidate gene studies, such as the work I presented, to investigate hypothesized gene-environment interactions are quite common in human genetic research. With the development of large-scale high-throughput genotyping technologies, genome-wide association gene-gene interaction (GWAI) and genome-wide environmental interaction (GWEI) studies are beginning to emerge. GWAI studies have a number of challenges including high-dimensionality, computational complexity, the absence/presence of marginal effects, the high burden of multiple test correction, and genetic heterogeneity (183). Given the unavailability of sufficiently large sample sizes and the dramatically increased multiple testing burden, GWAI studies usually end with no statistically significant findings. Several statistical approaches have been proposed, including regression-based approaches and model-free approaches such as machine learning and pattern recognition (184). In general, the advantages of regression-based approaches are the clear interpretation of the model and the parameters that relate genotypes to phenotype. However, they have huge computational burden for testing higher-order interactions. In comparison, machine learning approaches are an alternative strategy to detect high-dimensional non-linear interactions. These latter approaches generally do not estimate parameters, but rather they find combinations of SNPs that can best separate cases and controls by epistatic interactions or joint effects. Some model-free approaches collapse high dimensional data into two dimensions; some try to detect differentially inherited SNP modules by hierarchically clustering SNPs that could be interactively associated with a disease; and some use a multi-locus Mann–Whitney statistic to evaluate the joint association of a SNP combination. All these machine-learning approaches do not have the problem of an increasing number of parameters when modeling high-order interactions, but it is often difficult to interpret how the detected SNP combinations affect a disease. Despite the abundance of statistical methods and tools for interaction analysis in recent years, only a few of them have demonstrated replicable results and there is a need for further development and extension of these methods to identify gene-gene and gene-environment interactions in the context of genome-wide association studies.

Another novel and interesting finding is the importance of adjustment for local ancestry when performing the genetic association analysis in admixed populations. Admixed populations, those that descended from more than one ancestral population, offer a unique opportunity for mapping disease genes that have large allele frequency differences between ancestral populations. In Chapter III, I observed significant ($P<5x10^{-8}$) associations for hundreds of SNPs spanning ~10Mb region on 6q surrounding the *LPA* gene after adjusting for global ancestry estimate. Interestingly, after adjusting for local ancestry at 6q25.3, the region containing significantly associated SNPs got much narrower (from 9.8Mb to 0.7Mb) and was centered on the three genes *SLC22A*, *LPL2* and *LPA*. This result suggests confounding between local ancestry and SNPs spanning the larger 6q region identified to be associated with Lp(a). Given the relatively recent admixture in the AA population, local ancestry can confound associations across a relatively large region surrounding the population-specific, or population-enriched, causal variant(s) (132, 133). The observation that the associations in and near *LPA* remains robust after adjustment for local ancestry at *LPA* while the evidence for association further away dramatically declines suggests that the ancestry-specific (or highly-enriched) causal risk variant(s) resides in or near

LPA and that most, if not all, of the observed associations outside this narrower region are spurious associations. My work suggested that formal genetic admixture analyses may point to the correct region containing the causal SNPs, but the region may not be precise enough for following up causal SNPs or even causal genes. In comparison, the local ancestry adjustment did control the background signals induced by association with local ancestry and effectively prevented false positive findings, which makes it a useful tool for fine mapping of regions identified from admixture mapping studies. Several association tests that adjust for local ancestry have been proposed. For example, one test is based on a conditional likelihood framework which models the distribution of the test SNP given disease status and flanking marker genotypes. This conditional likelihood makes it possible to explicitly model local ancestry differences among study subjects and thus it can eliminate the effect of population stratification at the test SNP. It is particularly useful when the directions of association are different in the ancestral populations. Another test, which is computationally simpler, is based on logistic regression, with adjustment for local ancestry proportion. These association tests directly evaluate the correlation between a phenotype and a SNP genotype, and they directly compare the allele frequencies between cases and controls. In comparison, admixture mapping has substantially lower resolution than direct SNP association tests, as it does not make full use of the actual genotypes at each SNP and SNPs falling within the same extended ancestry block (which will be considerable larger than conventionally defined LD blocks due to the recent admixture events) will share similar admixture mapping signal. Future studies on association tests which can appropriately control for local ancestry is still needed as a general tool for genetic association studies in admixed populations.

In addition to the application of genetic association studies for discovering risk factors for CVD related traits, my work also included investigation of the study design used to perform future genetic association studies. In Chapter V, I proposed a two-stage genetic association study design, where it is assumed that all subjects have available directly measured or imputed genome-wide genotype data but an unmeasured quantitative trait of interest. It is further assumed that cost constraints will limit the ability to measure the new phenotype on only a subset of study participants. I described a selective phenotyping method that uses simulated annealing (SA) to identify the optimal subset of subjects to be phenotyped in Stage 2 of a two-stage GWAS based on the identification of the most interesting SNPs from a GWAS study on a subset of unselected participants in Stage 1. Through both simulations and a real example, I showed that our SA algorithm-based two-stage approach achieves increased overall statistical power compared to a single stage study using a random selection approach for studies that have existing imputed or directly genotyped markers. While two-stage approaches are not novel for genetic association studies, the proposed two-stage phenotyping design is, to our knowledge, completely new. My study not only applied the SA-based two-stage phenotype selection approach to directly genotyped data, but to imputed dosage data as well. Since genotype imputation is widely used in GWAS, the extension of our approach to include these data is an important contribution to this two-stage study design. This approach was not sensitive to parameter choices and remained robust for a wide range of MAFs, proportional of markers to be followed up in Stage 2, and proportion of samples to be phenotyped in Stage 1. It is important to note that β estimates are unbiased by this procedure (other than winner's curse which effects all large-scale discovery study designs, including random sample selection). In fact, because our sampling design reduces variance of the estimates for the β's, the estimates are actually more precise, on average, than estimates based on random samples. The method would clearly benefit power and precision of estimated effects for markers included in the SA optimization in Stage 2, and it is important to note that there would be little impact on the parameter estimates or power for markers not included in the Stage 2 SA sample optimization versus what can be achieved from random sample selection, assuming that most non-included markers are independent of those used for optimization.

In general, the future of genetic studies of complex diseases will rely on a variety of statistical and biological approaches, including candidate gene studies, high-throughput high-dimensional genotyping arrays, custom genotyping arrays targeting certain trait clusters (e.g. the Immunochip, the HumanCVD, MetaboChip), genome-wide gene-gene and gene-environment interaction analyses based on machine learning, whole-exome sequencing (WES) and whole-genome sequencing (WGS), and replication of biological changes in animal models. In particular, as a result of the development of next-generation sequencing (NGS) technologies, WES and WGS have become considerably faster and more affordable over the past 5 years (62). In contrast to the first-generation sequencing, which was expensive in both time and money consuming to sequence the diploid human genome, NGS can now be used to sequence the same human genome within a few weeks for as little as US $4,000-5,000 (185). The prices and experimental time continue to decline, as the costs today for WGS are approaching $1,000 per sample. Common NGS applications include DNA-seq, RNA-seq, ChIP-seq, and methyl-seq (186). DNA-seq is to discover whether genomic variations including single nucleotide variants (SNVs), small DNA insertions or deletions (indels), copy number variations (CNVs), or other structural variants (SVs), are associated with human diseases. RNA-seq that measures gene expression changes is to discover new transcripts including noncoding RNAs and detect transcript splicing or gene fusion events. ChIP-seq is to discover genome-wide transcription factor binding sites and chromatin-associated modifications. Methyl-seq is to discover various types of DNA methylation such as 5-methylcytosine and 5-hydroxymethylcytosine at single nucleotide resolution. These above common NGS applications may be adopted as mainstream tools for human genomics and routine procedures as part of the clinical laboratory for disease treatment in near future. Today, many analyses integrate information from multiple different sequencing applications (e.g. to assess whether rare DNA variants are associated with gene expression). These types of analyses will become for common in the near future and our ability to integrate these different types of data will likely lead to many new exciting discoveries.

In conclusion, as CVD is the number one killer in the United States, understanding the contributing genetic risk factors as well as their interactions with environment risk factors is of critical importance for the prevention and treatment of the disease. The results of genetic association studies conducted in my dissertation further our understanding of the genetic architecture of several biomarkers related to CVD, and the two-stage phenotyping design provides a more powerful way to perform future GWAS studies when the cost of phenotyping prohibits the inclusion of all available subjects.

REFERENCES

1 Mozaffarian, D., Benjamin, E.J., Go, A.S., Arnett, D.K., Blaha, M.J., Cushman, M., de Ferranti, S., Despres, J.P., Fullerton, H.J., Howard, V.J. *et al.* (2015) Executive summary: heart disease and stroke statistics-2015 update: a report from the american heart association. *Circulation*, **131**, 434-441.

2 Go, A.S., Mozaffarian, D., Roger, V.L., Benjamin, E.J., Berry, J.D., Borden, W.B., Bravata, D.M., Dai, S., Ford, E.S., Fox, C.S. *et al.* (2013) Heart disease and stroke statistics--2013 update: a report from the American Heart Association. *Circulation*, **127**, e6-e245.

3 Lopez, A.D., Mathers, C.D., Ezzati, M., Jamison, D.T. and Murray, C.J. (2006) Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data. *Lancet*, **367**, 1747-1757.

4 Pagidipati, N.J. and Gaziano, T.A. (2013) Estimating Deaths From Cardiovascular Disease: A Review of Global Methodologies of Mortality Measurement. *Circulation*, **127**, 749-756.

5 Ritchey, M.D., Wall, H.K., Gillespie, C., George, M.G. and Jamal, A. (2014) Million hearts: prevalence of leading cardiovascular disease risk factors--United States, 2005-2012. *MMWR. Morbidity and mortality weekly report*, **63**, 462-467.

6 Murabito, J.M., Pencina, M.J., Nam, B.H., D'Agostino, R.B., Sr., Wang, T.J., Lloyd-Jones, D., Wilson, P.W. and O'Donnell, C.J. (2005) Sibling cardiovascular disease as a risk factor for cardiovascular disease in middle-aged adults. *Jama*, **294**, 3117-3123.

7 Vasan, R.S. (2006) Biomarkers of Cardiovascular Disease: Molecular Basis and Practical Considerations. *Circulation*, **113**, 2335-2362.

8 Ge, Y. and Wang, T.J. (2012) Identifying novel biomarkers for cardiovascular disease risk prediction. *Journal of internal medicine*, **272**, 430-439.

9 Weir, C.J. and Walley, R.J. (2006) Statistical evaluation of biomarkers as surrogate endpoints: a literature review. *Statistics in Medicine*, **25**, 183-203.

10 Andrews, N.C. and Schmidt, P.J. (2007) Iron homeostasis. *Annu Rev Physiol*, **69**, 69-85.

11 Lapice, E., Masulli, M. and Vaccaro, O. (2013) Iron deficiency and cardiovascular disease: an updated review of the evidence. *Current atherosclerosis reports*, **15**, 358.

12 Lipinski, B. and Pretorius, E. (2013) Iron-induced fibrin in cardiovascular disease. *Current neurovascular research*, **10**, 269-274.

13 McCord, J.M. (1991) Is iron sufficiency a risk factor in ischemic heart disease? *Circulation*, **83**, 1112-1114.

14 Salonen, J.T., Nyyssonen, K., Korpela, H., Tuomilehto, J., Seppanen, R. and Salonen, R. (1992) High stored iron levels are associated with excess risk of myocardial infarction in eastern Finnish men. *Circulation*, **86**, 803-811.

15 Tuomainen, T.P., Diczfalusy, U., Kaikkonen, J., Nyyssonen, K. and Salonen, J.T. (2003) Serum ferritin concentration is associated with plasma levels of cholesterol oxidation products in man. *Free radical biology & medicine*, **35**, 922-928.

16 Franchini, M., Targher, G., Montagnana, M. and Lippi, G. (2008) Iron and thrombosis. *Annals of hematology*, **87**, 167-173.

17 Gunawardena, S. and Dunlap, M.E. (2012) Anemia and iron deficiency in heart failure. *Current heart failure reports*, **9**, 319-327.

18 Njajou, O.T., Alizadeh, B.Z., Aulchenko, Y., Zillikens, M.C., Pols, H.A., Oostra, B.A., Swinkels, D.W. and van Duijn, C.M. (2006) Heritability of serum iron, ferritin and transferrin saturation in a genetically isolated population, the Erasmus Rucphen Family (ERF) Study. *Hum Hered*, **61**, 222-228.

19 Wilson, J.G., Lindquist, J.H., Grambow, S.C., Crook, E.D. and Maher, J.F. (2003) Potential role of increased iron stores in diabetes. *Am J Med Sci*, **325**, 332-339.

20 Benyamin, B., McRae, A.F., Zhu, G., Gordon, S., Henders, A.K., Palotie, A., Peltonen, L., Martin, N.G., Montgomery, G.W., Whitfield, J.B. *et al.* (2009) Variants in TF and HFE explain approximately 40% of genetic variation in serum-transferrin levels. *Am J Hum Genet*, **84**, 60-65.

21 Chambers, J.C., Zhang, W., Li, Y., Sehmi, J., Wass, M.N., Zabaneh, D., Hoggart, C., Bayele, H., McCarthy, M.I., Peltonen, L. *et al.* (2009) Genome-wide association study identifies variants in TMPRSS6 associated with hemoglobin levels. *Nature genetics*, **41**, 1170-1172.

22 Benyamin, B., Ferreira, M.A., Willemsen, G., Gordon, S., Middelberg, R.P., McEvoy, B.P., Hottenga, J.J., Henders, A.K., Campbell, M.J., Wallace, L. *et al.* (2009) Common variants in TMPRSS6 are associated with iron status and erythrocyte volume. *Nature genetics*, **41**, 1173-1175.

23 Oexle, K., Ried, J.S., Hicks, A.A., Tanaka, T., Hayward, C., Bruegel, M., Gögele, M., Lichtner, P., Müller-Myhsok, B., Döring, A. *et al.* (2011) Novel association to the proprotein convertase PCSK7 gene locus revealed by analysing soluble transferrin receptor (sTfR) levels. *Human Molecular Genetics*, **20**, 1042-1047.

24 Camaschella, C., Roetto, A., Cali, A., De Gobbi, M., Garozzo, G., Carella, M., Majorano, N., Totaro, A. and Gasparini, P. (2000) The gene TFR2 is mutated in a new type of haemochromatosis mapping to 7q22. *Nature genetics*, **25**, 14-15.

25 McLaren, C.E., Garner, C.P., Constantine, C.C., McLachlan, S., Vulpe, C.D., Snively, B.M., Gordeuk, V.R., Nickerson, D.A., Cook, J.D., Leiendecker-Foster, C. *et al.* (2011) Genome-wide association study identifies genetic loci associated with iron deficiency. *PloS one*, **6**, e17390.

26 Koschinsky, M.L., Cote, G.P., Gabel, B. and van der Hoek, Y.Y. (1993) Identification of the cysteine residue in apolipoprotein(a) that mediates extracellular coupling with apolipoprotein B-100. *The Journal of biological chemistry*, **268**, 19819-19825.

27 Nordestgaard, B.G., Chapman, M.J., Ray, K., Boren, J., Andreotti, F., Watts, G.F., Ginsberg, H., Amarenco, P., Catapano, A., Descamps, O.S. *et al.* (2010) Lipoprotein(a) as a cardiovascular risk factor: current status. *European heart journal*, **31**, 2844-2853.

28 Erqou, S., Kaptoge, S., Perry, P.L., Di Angelantonio, E., Thompson, A., White, I.R., Marcovina, S.M., Collins, R., Thompson, S.G. and Danesh, J. (2009) Lipoprotein(a) concentration and the risk of coronary heart disease, stroke, and nonvascular mortality. *JAMA : the journal of the American Medical Association*, **302**, 412-423.

29 Genest, J.J., Jr., Martin-Munley, S.S., McNamara, J.R., Ordovas, J.M., Jenner, J., Myers, R.H., Silberman, S.R., Wilson, P.W., Salem, D.N. and Schaefer, E.J. (1992) Familial lipoprotein disorders in patients with premature coronary artery disease. *Circulation*, **85**, 2025-2033.

30 Willeit, P., Kiechl, S., Kronenberg, F., Witztum, J.L., Santer, P., Mayr, M., Xu, Q., Mayr, A., Willeit, J. and Tsimikas, S. (2014) Discrimination and net reclassification of cardiovascular risk with lipoprotein(a): prospective 15-year outcomes in the Bruneck Study. *Journal of the American College of Cardiology*, **64**, 851-860.

31 Dangas, G., Mehran, R., Harpel, P.C., Sharma, S.K., Marcovina, S.M., Dube, G., Ambrose, J.A. and Fallon, J.T. (1998) Lipoprotein(a) and inflammation in human coronary atheroma: association with the severity of clinical presentation. *Journal of the American College of Cardiology*, **32**, 2035-2042.

32 Deb, A. and Caplice, N.M. (2004) Lipoprotein(a): new insights into mechanisms of atherogenesis and thrombosis. *Clinical cardiology*, **27**, 258-264.

33 Glader, C.A., Boman, J., Saikku, P., Stenlund, H., Weinehall, L., Hallmanns, G. and Dahlen, G.H. (2000) The proatherogenic properties of lipoprotein(a) may be enhanced through the formation of circulating immune complexes containing Chlamydia pneumoniae-specific IgG antibodies. *European heart journal*, **21**, 639-646.

34 Boerwinkle, E., Leffert, C.C., Lin, J., Lackner, C., Chiesa, G. and Hobbs, H.H. (1992) Apolipoprotein(a) gene accounts for greater than 90% of the variation in plasma lipoprotein(a) concentrations. *The Journal of clinical investigation*, **90**, 52-60.

35 Mooser, V., Scheer, D., Marcovina, S.M., Wang, J., Guerra, R., Cohen, J. and Hobbs, H.H. (1997) The Apo(a) gene is the major determinant of variation in plasma Lp(a) levels in African Americans. *American journal of human genetics*, **61**, 402-417.

36 Clarke, R., Peden, J.F., Hopewell, J.C., Kyriakou, T., Goel, A., Heath, S.C., Parish, S., Barlera, S., Franzosi, M.G., Rust, S. *et al.* (2009) Genetic variants associated with Lp(a) lipoprotein level and coronary disease. *The New England journal of medicine*, **361**, 2518-2528.

37 Ober C Fau - Nord, A.S., Nord As Fau - Thompson, E.E., Thompson Ee Fau - Pan, L., Pan L Fau - Tan, Z., Tan Z Fau - Cusanovich, D., Cusanovich D Fau - Sun, Y., Sun Y Fau - Nicolae, R., Nicolae R Fau - Edelstein, C., Edelstein C Fau - Schneider, D.H., Schneider Dh Fau - Billstrand, C. *et al.* Genome-wide association study of plasma lipoprotein(a) levels identifies multiple genes on chromosome 6q. in press.

38 Sandholzer, C., Hallman, D.M., Saha, N., Sigurdsson, G., Lackner, C., Csaszar, A., Boerwinkle, E. and Utermann, G. (1991) Effects of the apolipoprotein(a) size polymorphism on the lipoprotein(a) concentration in 7 ethnic groups. *Human genetics*, **86**, 607-614.

39 Lanktree, M.B., Anand, S.S., Yusuf, S. and Hegele, R.A. (2010) Comprehensive analysis of genomic variation in the LPA locus and its relationship to plasma lipoprotein(a) in South Asians, Chinese, and European Caucasians. *Circulation. Cardiovascular genetics*, **3**, 39-46.

40 Go, A.S., Mozaffarian, D., Roger, V.L., Benjamin, E.J., Berry, J.D., Borden, W.B., Bravata, D.M., Dai, S., Ford, E.S., Fox, C.S. *et al.* (2013) Executive summary: heart disease and stroke statistics--2013 update: a report from the American Heart Association. *Circulation*, **127**, 143-152.

41 Lim, S.S., Vos, T., Flaxman, A.D., Danaei, G., Shibuya, K., Adair-Rohani, H., Amann, M., Anderson, H.R., Andrews, K.G., Aryee, M. *et al.* (2012) A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet*, **380**, 2224-2260.

42 Alwan, A. (2011), in press., pp. 176-.

43 Drazner, M.H. (2011) The Progression of Hypertensive Heart Disease. *Circulation*, **123**, 327-334.

44 Diamond, J.A. and Phillips, R.A. (2005) Hypertensive heart disease. *Hypertension research : official journal of the Japanese Society of Hypertension*, **28**, 191-202.

45 Lawler, P.R., Hiremath, P. and Cheng, S. (2014) Cardiac target organ damage in hypertension: insights from epidemiology. *Current hypertension reports*, **16**, 446.

46 Chobanian, A.V., Bakris, G.L., Black, H.R., Cushman, W.C., Green, L.A., Izzo, J.L., Jr., Jones, D.W., Materson, B.J., Oparil, S., Wright, J.T., Jr. *et al.* (2003) Seventh report of the Joint

National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. *Hypertension*, **42**, 1206-1252.

47 Kunes, J. and Zicha, J. (2009) The interaction of genetic and environmental factors in the etiology of hypertension. *Physiological research / Academia Scientiarum Bohemoslovaca*, **58 Suppl 2**, S33-41.

48 Perusse, L., Moll, P.P. and Sing, C.F. (1991) Evidence that a single gene with gender- and age-dependent effects influences systolic blood pressure determination in a population-based sample. *American journal of human genetics*, **49**, 94-105.

49 Levy, D., Ehret, G.B., Rice, K., Verwoert, G.C., Launer, L.J., Dehghan, A., Glazer, N.L., Morrison, A.C., Johnson, A.D., Aspelund, T. *et al.* (2009) Genome-wide association study of blood pressure and hypertension. *Nature genetics*, **41**, 677-687.

50 Levy, D., DeStefano, A.L., Larson, M.G., O'Donnell, C.J., Lifton, R.P., Gavras, H., Cupples, L.A. and Myers, R.H. (2000) Evidence for a gene influencing blood pressure on chromosome 17. Genome scan linkage results for longitudinal blood pressure phenotypes in subjects from the framingham heart study. *Hypertension*, **36**, 477-483.

51 Rotimi, C.N., Cooper, R.S., Cao, G., Ogunbiyi, O., Ladipo, M., Owoaje, E. and Ward, R. (1999) Maximum-likelihood generalized heritability estimate for blood pressure in Nigerian families. *Hypertension*, **33**, 874-878.

52 Gu, D., Rice, T., Wang, S., Yang, W., Gu, C., Chen, C.S., Hixson, J.E., Jaquish, C.E., Yao, Z.J., Liu, D.P. *et al.* (2007) Heritability of blood pressure responses to dietary sodium and potassium intake in a Chinese population. *Hypertension*, **50**, 116-122.

53 Hirschhorn, J.N., Lohmueller, K., Byrne, E. and Hirschhorn, K. (2002) A comprehensive review of genetic association studies. *Genetics in medicine : official journal of the American College of Medical Genetics*, **4**, 45-61.

54 Patnala, R., Clements, J. and Batra, J. (2013) Candidate gene association studies: a comprehensive guide to useful in silico tools. *BMC genetics*, **14**, 39.

55 Risch, N.J. (2000) Searching for genetic determinants in the new millennium. *Nature*, **405**, 847-856.

56 Ganesh, S.K., Tragante, V., Guo, W., Guo, Y., Lanktree, M.B., Smith, E.N., Johnson, T., Castillo, B.A., Barnard, J., Baumert, J. *et al.* (2013) Loci influencing blood pressure identified using a cardiovascular gene-centric array. *Hum Mol Genet*, **22**, 1663-1678.

57 Johnson, T., Gaunt, T.R., Newhouse, S.J., Padmanabhan, S., Tomaszewski, M., Kumari, M., Morris, R.W., Tzoulaki, I., O'Brien, E.T., Poulter, N.R. *et al.* (2011) Blood pressure loci identified with a gene-centric array. *American journal of human genetics*, **89**, 688-700.

58 Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.-Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGiovanni, J.P., Mane, S.M., Mayne, S.T. *et al.* (2005) Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science*, **308**, 385-389.

59 Kato, N., Takeuchi, F., Tabara, Y., Kelly, T.N., Go, M.J., Sim, X., Tay, W.T., Chen, C.H., Zhang, Y., Yamamoto, K. *et al.* (2011) Meta-analysis of genome-wide association studies identifies common variants associated with blood pressure variation in east Asians. *Nature genetics*, **43**, 531-538.

60 Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747-753.

61 Cohen, J.C., Kiss, R.S., Pertsemlidis, A., Marcel, Y.L., McPherson, R. and Hobbs, H.H. (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science*, **305**, 869-872.

62 Metzker, M.L. (2010) Sequencing technologies - the next generation. *Nature reviews. Genetics*, **11**, 31-46.

63 Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G. *et al.* (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, **337**, 64-69.

64 Grove, M.L., Yu, B., Cochran, B.J., Haritunians, T., Bis, J.C., Taylor, K.D., Hansen, M., Borecki, I.B., Cupples, L.A., Fornage, M. *et al.* (2013) Best practices and joint calling of the HumanExome BeadChip: the CHARGE Consortium. *PloS one*, **8**, e68095.

65 Huyghe, J.R., Jackson, A.U., Fogarty, M.P., Buchkovich, M.L., Stancakova, A., Stringham, H.M., Sim, X., Yang, L., Fuchsberger, C., Cederberg, H. *et al.* (2013) Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nature genetics*, **45**, 197-201.

66 Wessel, J., Chu, A.Y., Willems, S.M., Wang, S., Yaghootkar, H., Brody, J.A., Dauriz, M., Hivert, M.F., Raghavan, S., Lipovich, L. *et al.* (2015) Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. *Nature communications*, **6**, 5897.

67 Murcray, C.E., Lewinger, J.P. and Gauderman, W.J. (2009) Gene-environment interaction in genome-wide association studies. *American journal of epidemiology*, **169**, 219-226.

68 Winham, S.J. and Biernacka, J.M. (2013) Gene-environment interactions in genome-wide association studies: current approaches and new directions. *Journal of child psychology and psychiatry, and allied disciplines*, **54**, 1120-1134.

69 Murcray, C.E., Lewinger, J.P., Conti, D.V., Thomas, D.C. and Gauderman, W.J. (2011) Sample size requirements to detect gene-environment interactions in genome-wide association studies. *Genetic epidemiology*, **35**, 201-210.

70 Zhao, Q., Kelly, T.N., Li, C. and He, J. (2013) Progress and future aspects in genetics of human hypertension. *Current hypertension reports*, **15**, 676-686.

71 Hoggart, C.J., Shriver, M.D., Kittles, R.A., Clayton, D.G. and McKeigue, P.M. (2004) Design and analysis of admixture mapping studies. *American journal of human genetics*, **74**, 965-978.

72 Patterson, N., Hattangadi, N., Lane, B., Lohmueller, K.E., Hafler, D.A., Oksenberg, J.R., Hauser, S.L., Smith, M.W., O'Brien, S.J., Altshuler, D. *et al.* (2004) Methods for high-density admixture mapping of disease genes. *American journal of human genetics*, **74**, 979-1000.

73 Zhu, X., Tang, H. and Risch, N. (2008) Admixture mapping and the role of population structure for localizing disease genes. *Advances in genetics*, **60**, 547-569.

74 Devlin, B. and Roeder, K. (1999) Genomic control for association studies. *Biometrics*, **55**, 997-1004.

75 Satten, G.A., Flanders, W.D. and Yang, Q. (2001) Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *American journal of human genetics*, **68**, 466-477.

76 Li, M., Reilly, M.P., Rader, D.J. and Wang, L.S. (2010) Correcting population stratification in genetic association studies using a phylogenetic approach. *Bioinformatics (Oxford, England)*, **26**, 798-806.

77 Emma Huang, B., Clifford, D. and Cavanagh, C. (2013) Selecting subsets of genotyped experimental populations for phenotyping to maximize genetic diversity. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*, **126**, 379-388.

78 Jannink, J.-L. (2005) Selective phenotyping to accurately map quantitative trait loci. *Crop Science*, **45**, 901-908.

79 Xu, Z., Zou, F. and Vision, T.J. (2005) Improving quantitative trait loci mapping resolution in experimental crosses by the use of genotypically selected samples. *Genetics*, **170**, 401-408.

80 Jin, C., Lan, H., Attie, A.D., Churchill, G.A., Bulutuglo, D. and Yandell, B.S. (2004) Selective phenotyping for increased efficiency in genetic mapping studies. *Genetics*, **168**, 2285-2293.

81 Sen, S., Johannes, F. and Broman, K.W. (2009) Selective genotyping and phenotyping strategies in a complex trait context. *Genetics*, **181**, 1613-1626.

82 Oexle, K., Ried, J.S., Hicks, A.A., Tanaka, T., Hayward, C., Bruegel, M., Gogele, M., Lichtner, P., Muller-Myhsok, B., Doring, A. *et al.* (2011) Novel association to the proprotein convertase PCSK7 gene locus revealed by analysing soluble transferrin receptor (sTfR) levels. *Human molecular genetics*, **20**, 1042-1047.

83 Pichler, I., Minelli, C., Sanna, S., Tanaka, T., Schwienbacher, C., Naitza, S., Porcu, E., Pattaro, C., Busonero, F., Zanon, A. *et al.* (2011) Identification of a common variant in the TFR2 gene implicated in the physiological regulation of serum iron levels. *Human molecular genetics*, **20**, 1232-1240.

84 Traglia, M., Girelli, D., Biino, G., Campostrini, N., Corbella, M., Sala, C., Masciullo, C., Vigano, F., Buetti, I., Pistis, G. *et al.* (2011) Association of HFE and TMPRSS6 genetic variants with iron and erythrocyte parameters is only in part dependent on serum hepcidin concentrations. *J Med Genet*, **48**, 629-634.

85 Taylor, H.A., Jr. (2005) The Jackson Heart Study: an overview. *Ethn Dis*, **15**, S6-1-3.

86 Wilson, J.G., Rotimi, C.N., Ekunwe, L., Royal, C.D., Crump, M.E., Wyatt, S.B., Steffes, M.W., Adeyemo, A., Zhou, J., Taylor, H.A., Jr. *et al.* (2005) Study design for genetic analysis in the Jackson Heart Study. *Ethn Dis*, **15**, S6-30-37.

87 Evans, M.K., Lepkowski, J.M., Powe, N.R., LaVeist, T., Kuczmarski, M.F. and Zonderman, A.B. (2010) Healthy aging in neighborhoods of diversity across the life span (HANDLS): overcoming barriers to implementing a longitudinal, epidemiologic, urban study of health, race, and socioeconomic status. *Ethn Dis*, **20**, 267-275.

88 Li, Y., Willer, C.J., Ding, J., Scheet, P. and Abecasis, G.R. (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol*, **34**, 816-834.

89 Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. and Abecasis, G.R. (2012) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet*, **44**, 955-959.

90 Alexander, D.H., Novembre, J. and Lange, K. (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, **19**, 1655-1664.

91 Deo, R.C., Reich, D., Tandon, A., Akylbekova, E., Patterson, N., Waliszewska, A., Kathiresan, S., Sarpong, D., Taylor, H.A., Jr. and Wilson, J.G. (2009) Genetic differences between the determinants of lipid profile phenotypes in African and European Americans: the Jackson Heart Study. *PLoS genetics*, **5**, e1000342.

92 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, **81**, 559-575.

93 Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, **38**, 904-909.

94 Whitfield, J.B., Treloar, S., Zhu, G., Powell, L.W. and Martin, N.G. (2003) Relative importance of female-specific and non-female-specific effects on variation in iron stores between women. *Br J Haematol*, **120**, 860-866.

95 Cade, J.E., Moreton, J.A., O'Hara, B., Greenwood, D.C., Moor, J., Burley, V.J., Kukalizch, K., Bishop, D.T. and Worwood, M. (2005) Diet and genetic factors associated with iron status in middle-aged women. *Am J Clin Nutr*, **82**, 813-820.

96 McLaren, C.E., McLachlan, S., Garner, C.P., Vulpe, C.D., Gordeuk, V.R., Eckfeldt, J.H., Adams, P.C., Acton, R.T., Murray, J.A., Leiendecker-Foster, C. *et al.* (2012) Associations between Single Nucleotide Polymorphisms in Iron-Related Genes and Iron Status in Multiethnic Populations. *PLoS ONE*, **7**, e38339.

97 Beutler, E. and West, C. (2005) Hematologic differences between African-Americans and whites: the roles of iron deficiency and α-thalassemia on hemoglobin levels and mean corpuscular volume. *Blood*, **106**, 740-745.

98 Kutalik, Z., Benyamin, B., Bergmann, S., Mooser, V., Waeber, G., Montgomery, G.W., Martin, N.G., Madden, P.A., Heath, A.C., Beckmann, J.S. *et al.* (2011) Genome-wide association study identifies two loci strongly affecting transferrin glycosylation. *Hum Mol Genet*, **20**, 3710-3717.

99 Paterson, A.D., Waggott, D., Boright, A.P., Hosseini, S.M., Shen, E., Sylvestre, M.P., Wong, I., Bharaj, B., Cleary, P.A., Lachin, J.M. *et al.* (2010) A genome-wide association study identifies a novel major locus for glycemic control in type 1 diabetes, as measured by both A1C and glucose. *Diabetes*, **59**, 539-549.

100 Kathiresan, S., Manning, A.K., Demissie, S., D'Agostino, R.B., Surti, A., Guiducci, C., Gianniny, L., Burtt, N.P., Melander, O., Orho-Melander, M. *et al.* (2007) A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study. *BMC medical genetics*, **8 Suppl 1**, S17.

101 O'Donnell, C.J., Cupples, L.A., D'Agostino, R.B., Fox, C.S., Hoffmann, U., Hwang, S.J., Ingellson, E., Liu, C., Murabito, J.M., Polak, J.F. *et al.* (2007) Genome-wide association study for subclinical atherosclerosis in major arterial territories in the NHLBI's Framingham Heart Study. *BMC Med Genet*, **8 Suppl 1**, S4.

102 Lowe, J.K., Maller, J.B., Pe'er, I., Neale, B.M., Salit, J., Kenny, E.E., Shea, J.L., Burkhardt, R., Smith, J.G., Ji, W. *et al.* (2009) Genome-wide association studies in an isolated founder population from the Pacific Island of Kosrae. *PLoS Genet*, **5**, e1000365.

103 Vanita, V., Singh, D., Robinson, P.N., Sperling, K. and Singh, J.R. (2006) A novel mutation in the DNA-binding domain of MAF at 16q23.1 associated with autosomal dominant "cerulean cataract" in an Indian family. *Am J Med Genet A*, **140**, 558-566.

104 Jamieson, R.V., Perveen, R., Kerr, B., Carette, M., Yardley, J., Heon, E., Wirth, M.G., van Heyningen, V., Donnai, D., Munier, F. *et al.* (2002) Domain disruption and mutation of the bZIP transcription factor, MAF, associated with cataract, ocular anterior segment dysgenesis and coloboma. *Hum Mol Genet*, **11**, 33-42.

105 Christiansen, G. and Mohney, B.G. (2007) Hereditary hyperferritinemia-cataract syndrome. *Journal of AAPOS : the official publication of the American Association for Pediatric Ophthalmology and Strabismus / American Association for Pediatric Ophthalmology and Strabismus*, **11**, 294-296.

106 Suga, Y., Nagita, A., Takesako, R., Tanaka, I., Kobayashi, K., Hirai, M. and Matsuoka, H. (2011) A new glucose-6-phosphate dehydrogenase deficiency variant, G6PD Mizushima, showing increases in serum ferritin and cytosol leucine aminopeptidase levels. *Journal of pediatric hematology/oncology*, **33**, 15-17.

107 Wong, C.T. and Saha, N. (1987) Haemoglobin, serum iron, transferrin, ferritin concentrations and total iron-binding capacity in erythrocyte glucose-6-phosphate dehydrogenase deficiency. *Tropical and geographical medicine*, **39**, 350-353.

108 Chen, Z., Tang, H., Qayyum, R., Schick, U.M., Nalls, M.A., Handsaker, R., Li, J., Lu, Y., Yanek, L.R., Keating, B. *et al.* (2013) Genome-wide association analysis of red blood cell traits in African Americans: the COGENT Network. *Hum Mol Genet*, **22**, 2529-2538.

109 Lo, K.S., Wilson, J.G., Lange, L.A., Folsom, A.R., Galarneau, G., Ganesh, S.K., Grant, S.F., Keating, B.J., McCarroll, S.A., Mohler, E.R., 3rd *et al.* (2011) Genetic association analysis highlights new loci that modulate hematological trait variation in Caucasians and African Americans. *Human genetics*, **129**, 307-317.

110 Guindo, A., Fairhurst, R.M., Doumbo, O.K., Wellems, T.E. and Diallo, D.A. (2007) X-linked G6PD deficiency protects hemizygous males but not heterozygous females against severe malaria. *PLoS medicine*, **4**, e66.

111 Marcovina Sm Fau - Albers, J.J., Albers Jj Fau - Wijsman, E., Wijsman E Fau - Zhang, Z., Zhang Z Fau - Chapman, N.H., Chapman Nh Fau - Kennedy, H. and Kennedy, H. Differences in Lp[a] concentrations and apo[a] polymorphs between black and white Americans. in press.

112 Deo, R.C., Wilson, J.G., Xing, C., Lawson, K., Kao, W.H., Reich, D., Tandon, A., Akylbekova, E., Patterson, N., Mosley, T.H., Jr. *et al.* (2011) Single-nucleotide polymorphisms in LPA explain most of the ancestry-specific variation in Lp(a) levels in African Americans. *PloS one*, **6**, e14581.

113 Payne, T.J., Wyatt, S.B., Mosley, T.H., Dubbert, P.M., Guiterrez-Mohammed, M.L., Calvin, R.L., Taylor, H.A., Jr. and Williams, D.R. (2005) Sociocultural methods in the Jackson Heart Study: conceptual and descriptive overview. *Ethn Dis*, **15**, S6-38-48.

114 Wilson Jg Fau - Rotimi, C.N., Rotimi Cn Fau - Ekunwe, L., Ekunwe L Fau - Royal, C.D.M., Royal Cd Fau - Crump, M.E., Crump Me Fau - Wyatt, S.B., Wyatt Sb Fau - Steffes, M.W., Steffes Mw Fau - Adeyemo, A., Adeyemo A Fau - Zhou, J., Zhou J Fau - Taylor, H.A., Jr., Taylor Ha Jr Fau - Jaquish, C. *et al.* Study design for genetic analysis in the Jackson Heart Study. in press.

115 Carpenter Ma Fau - Crow, R., Crow R Fau - Steffes, M., Steffes M Fau - Rock, W., Rock W Fau - Heilbraun, J., Heilbraun J Fau - Evans, G., Evans G Fau - Skelton, T., Skelton T Fau - Jensen, R., Jensen R Fau - Sarpong, D. and Sarpong, D. Laboratory, reading center, and coordinating center data management methods in the Jackson Heart Study. in press.

116 Asselbergs, Folkert W., Guo, Y., van Iperen, Erik P., Sivapalaratnam, S., Tragante, V., Lanktree, Matthew B., Lange, Leslie A., Almoguera, B., Appelman, Yolande E., Barnard, J. *et al.* (2012) Large-Scale Gene-Centric Meta-analysis across 32 Studies Identifies Multiple Lipid Loci. *American journal of human genetics*, **91**, 823-838.

117 Purcell S Fau - Neale, B., Neale B Fau - Todd-Brown, K., Todd-Brown K Fau - Thomas, L., Thomas L Fau - Ferreira, M.A.R., Ferreira Ma Fau - Bender, D., Bender D Fau - Maller, J., Maller J Fau - Sklar, P., Sklar P Fau - de Bakker, P.I.W., de Bakker Pi Fau - Daly, M.J., Daly Mj Fau - Sham, P.C. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. in press.

118 Li Y Fau - Willer, C.J., Willer Cj Fau - Ding, J., Ding J Fau - Scheet, P., Scheet P Fau - Abecasis, G.R. and Abecasis, G.R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. in press.

119 Alexander Dh Fau - Novembre, J., Novembre J Fau - Lange, K. and Lange, K. Fast model-based estimation of ancestry in unrelated individuals. in press.

120 Price Al Fau - Patterson, N.J., Patterson Nj Fau - Plenge, R.M., Plenge Rm Fau - Weinblatt, M.E., Weinblatt Me Fau - Shadick, N.A., Shadick Na Fau - Reich, D. and Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. in press.

121 Valdar, W., Sabourin, J., Nobel, A. and Holmes, C.C. (2012) Reprioritizing genetic associations in hit regions using LASSO-based resample model averaging. *Genetic epidemiology*, **36**, 451-462.

122 Scheet, P. and Stephens, M. (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *American journal of human genetics*, **78**, 629-644.

123 Abecasis, G.R. and Cookson, W.O. (2000) GOLD--graphical overview of linkage disequilibrium. *Bioinformatics*, **16**, 182-183.

124 Barrett, J.C., Fry, B., Maller, J. and Daly, M.J. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263-265.

125 Schaid, D.J. (2006) Power and sample size for testing associations of haplotypes with complex traits. *Annals of human genetics*, **70**, 116-130.

126 Madsen, B.E. and Browning, S.R. (2009) A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet*, **5**, e1000384.

127 Lee S Fau - Wu, M.C., Wu Mc Fau - Lin, X. and Lin, X. Optimal tests for rare variant effects in sequencing association studies. in press.

128 Wu Mc Fau - Lee, S., Lee S Fau - Cai, T., Cai T Fau - Li, Y., Li Y Fau - Boehnke, M., Boehnke M Fau - Lin, X. and Lin, X. Rare-variant association testing for sequencing data with the sequence kernel association test. in press.

129 Adzhubei, I., Jordan, D.M. and Sunyaev, S.R. (2013) Predicting functional effect of human missense mutations using PolyPhen-2. *Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.]*, **Chapter 7**, Unit7.20.

130 Ng, P.C. and Henikoff, S. (2003) SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research*, **31**, 3812-3814.

131 Steyrer, E., Durovic, S., Frank, S., Giessauf, W., Burger, A., Dieplinger, H., Zechner, R. and Kostner, G.M. (1994) The role of lecithin: cholesterol acyltransferase for lipoprotein (a) assembly. Structural integrity of low density lipoproteins is a prerequisite for Lp(a) formation in human plasma. *The Journal of clinical investigation*, **94**, 2330-2340.

132 Qin, H., Morris, N., Kang, S.J., Li, M., Tayo, B., Lyon, H., Hirschhorn, J., Cooper, R.S. and Zhu, X. (2010) Interrogating local population structure for fine mapping in genome-wide association studies. *Bioinformatics*, **26**, 2961-2968.

133 Wang, X., Zhu, X., Qin, H., Cooper, R.S., Ewens, W.J., Li, C. and Li, M. (2011) Adjustment for local ancestry in genetic association analysis of admixed populations. *Bioinformatics*, **27**, 670-677.

134 Heng, C.K., Saha, N. and Tay, J.S. (1995) Lack of association of apolipoprotein E polymorphism with plasma Lp(a) levels in the Chinese. *Clinical genetics*, **48**, 113-119.

135 Muros, M. and Rodriguez-Ferrer, C. (1996) Apolipoprotein E polymorphism influence on lipids, apolipoproteins and Lp(a) in a Spanish population underexpressing apo E4. *Atherosclerosis*, **121**, 13-21.

136 Muls, E., Kempen, K., Vansant, G., Cobbaert, C. and Saris, W. (1993) The effects of weight loss and apolipoprotein E polymorphism on serum lipids, apolipoproteins A-I and B, and lipoprotein(a). *International journal of obesity and related metabolic disorders : journal of the International Association for the Study of Obesity*, **17**, 711-716.

137 Schaefer, E.J., Lamon-Fava, S., Johnson, S., Ordovas, J.M., Schaefer, M.M., Castelli, W.P. and Wilson, P.W. (1994) Effects of gender and menopausal status on the association of apolipoprotein E phenotype with plasma lipoprotein levels. Results from the Framingham Offspring Study. *Arteriosclerosis and thrombosis : a journal of vascular biology / American Heart Association*, **14**, 1105-1113.

138 Bailleul, S., Couderc, R., Rossignol, C., Fermanian, J., Boutouchent, F., Farnier, M.A. and Etienne, J. (1995) Lipoprotein(a) in childhood: relation with other atherosclerosis risk factors and family history of atherosclerosis. *Clinical chemistry*, **41**, 241-245.

139 de Knijff, P., Kaptein, A., Boomsma, D., Princen, H.M., Frants, R.R. and Havekes, L.M. (1991) Apolipoprotein E polymorphism affects plasma levels of lipoprotein(a). *Atherosclerosis*, **90**, 169-174.

140 Routi, T., Ronnemaa, T., Salo, P., Seppanen, R., Marniemi, J., Viikari, J., Ehnholm, C. and Simell, O. (1996) Effects of prospective, randomized cholesterol-lowering dietary intervention and apolipoprotein E phenotype on serum lipoprotein(a) concentrations of infants aged 7-24 mo. *The American journal of clinical nutrition*, **63**, 386-391.

141 Klausen, I.C., Gerdes, L.U., Hansen, P.S., Lemming, L., Gerdes, C. and Faergeman, O. (1996) Effects of apoE gene polymorphism on Lp(a) concentrations depend on the size of apo(a): a study of 466 white men. *Journal of molecular medicine (Berlin, Germany)*, **74**, 685-690.

142 Horita, K., Eto, M., Saito, M., Nakata, H., Iwashima, Y., Ito, H., Takahashi, M., Kondo, A., Morikawa, A. and Makino, I. (1993) Effects of apolipoprotein E polymorphism on plasma lipoprotein(a) levels. *Artery*, **20**, 324-336.

143 Frikke-Schmidt, R., Nordestgaard, B.G., Agerholm-Larsen, B., Schnohr, P. and Tybjaerg-Hansen, A. (2000) Context-dependent and invariant associations between lipids, lipoproteins, and apolipoproteins and apolipoprotein E genotype. *Journal of lipid research*, **41**, 1812-1822.

144 Lindahl, G., Mailly, F., Humphries, S. and Seed, M. (1994) Apolipoprotein E phenotype and lipoprotein(a) in familial hypercholesterolaemia: implication for lipoprotein(a) metabolism. *The Clinical investigator*, **72**, 631-638.

145 Anuurad, E., Lu, G., Rubin, J., Pearson, T.A. and Berglund, L. (2007) ApoE genotype affects allele-specific apo[a] levels for large apo[a] sizes in African Americans: the Harlem-Basset Study. *Journal of lipid research,* **48**, 693-698.

146 Joffres, M., Falaschetti, E., Gillespie, C., Robitaille, C., Loustalot, F., Poulter, N., McAlister, F.A., Johansen, H., Baclic, O. and Campbell, N. (2013) Hypertension prevalence, awareness, treatment and control in national surveys from England, the USA and Canada, and correlation with stroke and ischaemic heart disease mortality: a cross-sectional study. *BMJ open*, **3**, e003423.

147 Gao, Y., Chen, G., Tian, H., Lin, L., Lu, J., Weng, J., Jia, W., Ji, L., Xiao, J., Zhou, Z. *et al.* (2013) Prevalence of hypertension in china: a cross-sectional study. *PloS one*, **8**, e65938.

148 Kearney, P.M., Whelton, M., Reynolds, K., Muntner, P., Whelton, P.K. and He, J. (2005) Global burden of hypertension: analysis of worldwide data. *Lancet*, **365**, 217-223.

149 Alcocer, L. and Cueto, L. (2008) Hypertension, a health economics perspective. *Therapeutic advances in cardiovascular disease*, **2**, 147-155.

150 Tabara, Y., Kohara, K., Kita, Y., Hirawa, N., Katsuya, T., Ohkubo, T., Hiura, Y., Tajima, A., Morisaki, T., Miyata, T. *et al.* (2010) Common variants in the ATP2B1 gene are associated with susceptibility to hypertension: the Japanese Millennium Genome Project. *Hypertension*, **56**, 973-980.

151 Popkin, B.M., Du, S., Zhai, F. and Zhang, B. (2010) Cohort Profile: The China Health and Nutrition Survey—monitoring and understanding socio-economic and health change in China, 1989–2011. *International Journal of Epidemiology*, **39**, 1435-1440.

152 Chio, S.S., Urbina, E.M., Lapointe, J., Tsai, J. and Berenson, G.S. (2011) Korotkoff sound versus oscillometric cuff sphygmomanometers: comparison between auscultatory and DynaPulse blood pressure measurements. *Journal of the American Society of Hypertension : JASH*, **5**, 12-20.

153 Johnson, A.D., Newton-Cheh, C., Chasman, D.I., Ehret, G.B., Johnson, T., Rose, L., Rice, K., Verwoert, G.C., Launer, L.J., Gudnason, V. *et al.* (2011) Association of hypertension drug target genes with blood pressure and hypertension in 86,588 individuals. *Hypertension*, **57**, 903-910.

154 Montasser, M.E., Shimmin, L.C., Hanis, C.L., Boerwinkle, E. and Hixson, J.E. (2009) Gene by smoking interaction in hypertension: identification of a major quantitative trait locus on chromosome 15q for systolic blood pressure in Mexican-Americans. *Journal of hypertension*, **27**, 491-501.

155 Batis, C., Gordon-Larsen, P., Cole, S.R., Du, S., Zhang, B. and Popkin, B. (2013) Sodium intake from various time frames and incident hypertension among Chinese adults. *Epidemiology (Cambridge, Mass.)*, **24**, 410-418.

156 Pinheiro, J.C. and Bates, D.M. (2000) *Mixed-Effects Models in S and SPLUS. New York*. Springer.

157 Yang, X., Sun, J., Gao, Y., Tan, A., Zhang, H., Hu, Y., Feng, J., Qin, X., Tao, S., Chen, Z. *et al.* (2012) Genome-wide association study for serum complement C3 and C4 levels in healthy Chinese subjects. *PLoS genetics*, **8**, e1002916.

158 Tan, A., Gao, Y., Yang, X., Zhang, H., Qin, X., Mo, L., Peng, T., Xia, N. and Mo, Z. (2011) Low serum osteocalcin level is a potential marker for metabolic syndrome: results from a Chinese male population survey. *Metabolism: clinical and experimental*, **60**, 1186-1192.

159 Reynolds, K., Gu, D., Muntner, P., Wu, X., Chen, J., Huang, G., Duan, X., Whelton, P.K. and He, J. (2003) Geographic variations in the prevalence, awareness, treatment and control of hypertension in China. *Journal of hypertension*, **21**, 1273-1281.

160 Ehret, G.B., Munroe, P.B., Rice, K.M., Bochud, M., Johnson, A.D., Chasman, D.I., Smith, A.V., Tobin, M.D., Verwoert, G.C., Hwang, S.J. *et al.* (2011) Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*, **478**, 103-109.

161 Newton-Cheh, C., Johnson, T., Gateva, V., Tobin, M.D., Bochud, M., Coin, L., Najjar, S.S., Zhao, J.H., Heath, S.C., Eyheramendy, S. *et al.* (2009) Genome-wide association study identifies eight loci associated with blood pressure. *Nature genetics*, **41**, 666-676.

162 Liu, C., Li, H., Qi, Q., Lu, L., Gan, W., Loos, R.J. and Lin, X. (2011) Common variants in or near FGF5, CYP17A1 and MTHFR genes are associated with blood pressure and hypertension in Chinese Hans. *Journal of hypertension*, **29**, 70-75.

163 Niu, W., Zhang, Y., Ji, K., Gu, M., Gao, P. and Zhu, D. (2010) Confirmation of top polymorphisms in hypertension genome wide association study among Han Chinese. *Clinica chimica acta; international journal of clinical chemistry*, **411**, 1491-1495.

164 Cho, Y.S., Go, M.J., Kim, Y.J., Heo, J.Y., Oh, J.H., Ban, H.J., Yoon, D., Lee, M.H., Kim, D.J., Park, M. *et al.* (2009) A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nature genetics*, **41**, 527-534.

165 Brown, C.D., Higgins, M., Donato, K.A., Rohde, F.C., Garrison, R., Obarzanek, E., Ernst, N.D. and Horan, M. (2000) Body mass index and the prevalence of hypertension and dyslipidemia. *Obesity research*, **8**, 605-619.

166 Kotsis, V., Stabouli, S., Bouldin, M., Low, A., Toumanidis, S. and Zakopoulos, N. (2005) Impact of obesity on 24-hour ambulatory blood pressure and hypertension. *Hypertension*, **45**, 602-607.

167 Bell, S.E., Mavila, A., Salazar, R., Bayless, K.J., Kanagala, S., Maxwell, S.A. and Davis, G.E. (2001) Differential gene expression during capillary morphogenesis in 3D collagen matrices: regulated expression of genes involved in basement membrane matrix assembly, cell cycle progression, cellular differentiation and G-protein signaling. *Journal of cell science*, **114**, 2755-2773.

168 Park, S.Y., Lee, H.J., Ji, S.M., Kim, M.E., Jigden, B., Lim, J.E. and Oh, B. (2014) ANTXR2 is a potential causative gene in the genome-wide association study of the blood pressure locus 4q21. *Hypertension research : official journal of the Japanese Society of Hypertension*, in press.

169 Irving, R.J., Walker, B.R., Noon, J.P., Watt, G.C., Webb, D.J. and Shore, A.C. (2002) Microvascular correlates of blood pressure, plasma glucose, and insulin resistance in health. *Cardiovascular research*, **53**, 271-276.

170 Frisbee, J.C. (2005) Hypertension-independent microvascular rarefaction in the obese Zucker rat model of the metabolic syndrome. *Microcirculation (New York, N.Y. : 1994)*, **12**, 383-392.

171 De Boer, M.P., Meijer, R.I., Wijnstok, N.J., Jonk, A.M., Houben, A.J., Stehouwer, C.D., Smulders, Y.M., Eringa, E.C. and Serne, E.H. (2012) Microvascular dysfunction: a potential mechanism in the pathogenesis of obesity-associated insulin resistance and hypertension. *Microcirculation (New York, N.Y. : 1994)*, **19**, 5-18.

172 Jonk, A.M., Houben, A.J., de Jongh, R.T., Serne, E.H., Schaper, N.C. and Stehouwer, C.D. (2007) Microvascular dysfunction in obesity: a potential mechanism in the pathogenesis of obesity-associated insulin resistance and hypertension. *Physiology (Bethesda, Md.)*, **22**, 252-260.

173 Coskun, T., Bina, H.A., Schneider, M.A., Dunbar, J.D., Hu, C.C., Chen, Y., Moller, D.E. and Kharitonenkov, A. (2008) Fibroblast growth factor 21 corrects obesity in mice. *Endocrinology*, **149**, 6018-6027.

174 Kharitonenkov, A., Wroblewski, V.J., Koester, A., Chen, Y.F., Clutinger, C.K., Tigno, X.T., Hansen, B.C., Shanafelt, A.B. and Etgen, G.J. (2007) The metabolic state of diabetic monkeys is regulated by fibroblast growth factor-21. *Endocrinology*, **148**, 774-781.

175 Tomlinson, E., Fu, L., John, L., Hultgren, B., Huang, X., Renz, M., Stephan, J.P., Tsai, S.P., Powell-Braxton, L., French, D. *et al.* (2002) Transgenic mice expressing human fibroblast growth factor-19 display increased metabolic rate and decreased adiposity. *Endocrinology*, **143**, 1741-1747.

176 Hirschhorn, J.N. and Daly, M.J. (2005) Genome-wide association studies for common diseases and complex traits. *Nature reviews. Genetics*, **6**, 95-108.

177 Kirkpatrick, S., Gelatt, C.D., Jr. and Vecchi, M.P. (1983) Optimization by simulated annealing. *Science (New York, N.Y.)*, **220**, 671-680.

178 Coghlan, A., Mac Donaill, D.A. and Buttimore, N.H. (2001) Representation of amino acids as five-bit or three-bit patterns for filtering protein databases. *Bioinformatics (Oxford, England)*, **17**, 676-685.

179 Rutenbar, R.A. (1989) Simulated annealing algorithms: An overview. *Circuits and Devices Magazine, IEEE*, **5**, 19-26.

180 Skol, A.D., Scott, L.J., Abecasis, G.R. and Boehnke, M. (2006) Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nature genetics*, **38**, 209-213.

181 Anderson, C.A., Pettersson, F.H., Barrett, J.C., Zhuang, J.J., Ragoussis, J., Cardon, L.R. and Morris, A.P. (2008) Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms. *American journal of human genetics*, **83**, 112-119.

182 Wood, A.R., Hernandez, D.G., Nalls, M.A., Yaghootkar, H., Gibbs, J.R., Harries, L.W., Chong, S., Moore, M., Weedon, M.N., Guralnik, J.M. *et al.* (2011) Allelic heterogeneity and more detailed analyses of known loci explain additional phenotypic variation and reveal complex patterns of association. *Hum Mol Genet*, **20**, 4082-4092.

183 Steen, K.V. (2012) Travelling the world of gene-gene interactions. *Briefings in bioinformatics*, **13**, 1-19.

184 Sun, X., Lu, Q., Mukherjee, S., Crane, P.K., Elston, R. and Ritchie, M.D. (2014) Analysis pipeline for the epistasis search - statistical versus biological filtering. *Frontiers in genetics*, **5**, 106.

185 Thompson, R., Drew, C.J. and Thomas, R.H. (2012) Next generation sequencing in the clinical domain: clinical advantages, practical, and ethical challenges. *Advances in protein chemistry and structural biology*, **89**, 27-63.

186 Bao, R., Huang, L., Andrade, J., Tan, W., Kibbe, W.A., Jiang, H. and Feng, G. (2014) Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing. *Cancer informatics*, **13**, 67-82.