THE GENETIC ARCHITECTURE OF ECHOCARDIOGRAPHICALLY DETERMINED MEASURES OF LEFT VENTRICULAR REMODELING IN AFRICAN-AMERICANS OF THE GENETIC EPIDEMIOLOGY NETWORK OF ARTERIOPATHY (GENOA) STUDY

by

Kristin Joy Meyers

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Epidemiological Sciences)
in The University of Michigan
2009

Doctoral Committee:

Professor Sharon Reilly Kardia, Chair
Professor Ana V. Diez Roux
Professor Patricia A. Peyser
Assistant Professor Bhramar Mukherjee

Acknowledgements

First and foremost, I would like to thank Sharon Kardia for being an incredibly inspiring and insightful mentor. I have learned many academic, research, and life lessons from you that will be taken with me throughout my career. Everyone in the "Kardia Lab" has also been an amazing support network during the last four years, this includes Yan Sun, Jian Chu, Tracy Fuller, Jennifer Smith, Reagan Kelly, Todd Greene, Linda Feldkamp, Doug Jacobsen, and Koji Yanagisawa. Thank you for making graduate school not only bearable, but enjoyable. Jen, Reagan and Todd, it is hard to imagine the world of genetic epidemiology without you.

I would also like to thank my doctoral committee members, Pat Peyser, Bhramar Mukherjee, and Ana Diez-Roux. You were all a joy to work with. Thank you for your comments, patience, time, and for shaping me into the epidemiologist I am today.

I would also like to thank my comp exam study group, I feel blessed to have been able to go through every step of the program with a great group of friends and colleagues. Specifically Eileen and Jen, I would also like to thank you both for being so supportive of me the past year. Providing a place to sleep, cars to borrow, food to eat, and a friendly face so that I still feel at "home". I also need to thank Sarah, Josh and Zoe for opening their doors for me as well over the past year and a half.

While they almost never understood exactly what I was doing or talking about, my family and friends have been tremendously supportive of me over the years as well.

Mom, because of graduate school, I will never forget your birthday ever again. Sorry about that. Brent, thank you for supporting me through this endeavor, helping me through the stressful times, sharing the joy of my accomplishments, and helping me to keep all of this in perspective. Believe it or not – I'm finally done with school!

Table of Contents

List of Tables

Table Page

List of Figures

List of Abbreviations

AIMs	ancestry informative markers
ANCOVA	analysis of covariance
ANOVA	analysis of variance
ARIC	Atherosclerosis Risk in Communities
BMI	body mass index (kg/m^2)
CAD	coronary artery disease
CART	classification and regression trees
CPM	combinatorial partitioning method
CDCV	common disease common variant hypothesis
CV	cross-validation
DBP	diastolic blood pressure (mmHg)
FBPP	Family Blood Pressure Program
FDR	False Discovery Rate
GC	genomic control
GENOA	Genetic Epidemiology Network of Arteriopathy
GWAS	genome-wide association study
HGDP	Human Genome Diversity Project
HR	hazard ratio
HWE	Hardy-Weinberg equilibrium

HyperGEN	Hypertension Genetics study cohort of FBPP
LAMP	Local Ancestry in adMixed Populations
LME	linear mixed effects model
LV	left ventricle
LVH	left ventricular hypertrophy
LVIDD	left ventricular internal diameter
LVM	left ventricular mass (g)
LVMI	left ventricular mass index (g/m$^{2.7}$)
MAF	minor allele frequency
MI	myocardial infarction
MDR	Multifactor Dimensionality Reduction
NHLBI	National Heart Lung and Blood Institute
PCA	principal component analysis
PCR	polymerase chain reaction
PWT	posterior wall thickness
RF	Random Forests
RR	relative risk
RWT	relative wall thickness
SBP	systolic blood pressure (mmHg)
SNP	single nucleotide polymorphism
SOLAR	sequential oligogenic linkage analysis routine

Chapter 1

Introduction

Achievements in public health research and practice over the past 50 years have led to substantial decreases in mortality from heart disease, defined as myocardial infarction (MI), hypertensive and ischemic heart disease, and heart failure. In spite of this, heart disease remains the number one cause of mortality and morbidity in the United States[1], racial/ethnic disparities still exist, and there is extensive variation in individual level responses to both pharmaceutical and behavioral intervention on traditional risk factors. Genetics has often been considered a major component of the "black-box" explanation underlying this variation and the persistence of heart disease in the population.

In this thesis, I will focus on studying the genetic architecture of two quantitative measures of left ventricular (LV) remodeling – left ventricular mass (LVM) and relative wall thickness (RWT). Both LVM and RWT have been repeatedly documented as strong predictors of heart disease incidence, heart disease mortality, and all-cause mortality.[2-4]

LV remodeling is a compensatory response of the heart to increased volume and pressure load in the left ventricle from any number of sources, one of which is high blood pressure. However, the relationship between hypertension, LV remodeling, and clinical outcomes is hardly straightforward. LVM and RWT are quantitative, complex traits that exhibit familial correlations. While these traits have been examined in numerous candidate gene studies, the genetic underpinnings of LVM and RWT have yet to be fully elucidated.

Genetic Mapping of Complex Traits

Heart disease, LVM and RWT are complex traits that arise from interactions among many genetic and environmental factors. To date, most published genetic studies for complex traits in humans have been conducted using one of two approaches: linkage analysis or candidate-gene association studies.[5] Linkage analysis is an approach used to identify chromosomal regions that may contain a gene locus co-segregating with a trait of interest. Linkage analysis is conducted by scanning a set of genome-wide markers for co-segregation with a trait of interest in families. Linkage analysis proved very successful for mapping “Mendelian” diseases (ex. Hunginton’s disease and small subsets of commonly occurring diseases such as Alzheimer’s disease and breast cancer).[6] One of the lessons learned from linkage analysis was that even seemingly “Mendelian” diseases exhibit a lot of complexity including allelic heterogeneity, locus heterogeneity, and incomplete penetrance.[7] While successful for certain traits, linkage analysis left certain questions unanswered, such as, what is the exact location of genetic mutations associated with these diseases; linkage only provided rough estimates of genomic regions.

As implied by the name, candidate gene association studies limit the number of genes investigated to either 1) genes suspected to be associated with the outcome of interest based on pathophysiological knowledge of the disease process or 2) genes located in genomic regions identified through linkage studies.[5] Candidate gene association studies became popular because unlike linkage analysis, they could be conducted in epidemiological population-based samples (instead of family data), had greater power to identify associations for complex traits, and further localized the genomic region where the causal variant may lie.[5] While grounded in *a priori* knowledge, the candidate gene approach does not fully examine the genetic architecture of complex traits and is limited by our incomplete understanding of disease processes. The genetic architecture of complex traits is likely to involve many loci from across the genome (not only in candidate genes) operating independently (with differing magnitudes of effect), through epistasis (gene x gene interactions), or via environmentally context dependent effects (gene x environment interaction).

The "common disease – common variant" (CDCV) hypothesis was born in the mid 1990's as a result of the following reflections: 1) linkage analysis was identifying rare variants associated with Mendelian traits, not common diseases, 2) most variation in the genome is common, and 3) the *a priori* knowledge informing candidate gene studies was clearly insufficient.[7, 8] Because of the CDCV hypothesis, the door opened to conduct genome-wide association studies (GWAS) in large epidemiological cohorts. GWAS systematically search hundreds of thousands of common variants throughout the genome for association with a given trait.

GWAS have many advantages compared to candidate gene association studies. First, scanning a dense set of markers across the entire genome provides an agnostic approach to finding genetic variation contributing to complex traits. Second, because of the immense multiple testing issues resulting from testing thousands of genetic markers, the research community has set high standards for conducting, replicating, and publishing GWAS,[9] which is not found in all areas of epidemiological research. Third, in order to gain statistical power and also to meet the stringent requirements for publication, collaboration among numerous large studies has been encouraged and materialized (examples are efforts by McPherson et al.[10] and Samani et al.[11]).

Heart Disease and Left Ventricular Traits

Heart disease, including MI, hypertensive and ischemic heart disease, and heart failure, accounts for 27% of the 2.4 million deaths in the U.S. each year.[1] Understanding predictors of cardiac events is particularly important for the African-American population. In 2004, the age-adjusted heart disease death rate for African-Americans was higher compared to non-Hispanic whites, 281 vs. 213 per 100,000 persons per year.[1] Furthermore, hypertension (defined as either systolic blood pressure (SBP) ≥140mmHg or diastolic blood pressure (DBP) ≥90mmHg) occurs in 40-60% of African-American adults compared to 28-38% of non-Hispanic white adults[12, 13] and the prevalence of resulting cardiac target organ damage is nearly twice that in the non-Hispanic white population.[14]

Hypertension is one of the primary risk factors for heart disease. However, because persons with similar blood pressure levels have varied cardiac outcomes, and due to the complex physiology involved in blood pressure and heart disease, hypertension is not a necessary nor sufficient predictor of cardiac risk.[2, 3] Stronger, independent predictors of heart disease outcomes are measurements of LV remodeling, such as increased LVM and RWT.[2, 3, 15]

Increased LVM and RWT are repeatedly identified as independent predictors of sudden death, ventricular arrhythmias, myocardial ischemia, congestive heart failure, stroke, and angina.[16, 17] The Framingham Heart Study found that for each 50 g/m increase in LVM, the adjusted relative risk (RR) of incident cardiovascular disease was 1.49 (95% CI, 1.20-1.85) in men and 1.57 (95% CI, 1.20-2.04) in women.[2] In a largely hypertensive, African-American cohort from the Cook County Heart Disease Registry in Chicago, the RR of all-cause mortality associated with LV hypertrophy (defined as LVM >131 g/m^2 in men and >100 g/m^2 in women) was 2.14 (95% CI, 1.24-3.68) among those with coronary artery disease (CAD) (diagnosed if at least one vessel showed ≥70% reduction in diameter determined via cardiac catheterization) and 4.14 (95% CI, 1.77-9.71) among those without CAD after adjusting for age, sex, and hypertension.[18] More recently, the Atherosclerosis Risk in Communities (ARIC) study reported hazard ratios (HR) of all-cause mortality based on geometric patterns of LV hypertrophy.[19] The highest risk of death for men was in those with a specific geometric pattern of LV remodeling known as concentric hypertrophy (increased RWT and LVM) (HR=1.75, 95% CI 0.71-4.33) and the highest risk in women was for those with eccentric hypertrophy (normal RWT and increased LVM) (HR=1.23, 95% CI 0.46-3.28).[19] Therefore, the risk associated with increased LVM is present irrespective of gender, race, or CAD status. However, the risk may depend on geometric patterns of hypertrophy (ie. RWT) and the gender and racial differences have yet to be definitively isolated.

Left Ventricular Physiology

For over twenty years the increased risks of incident heart disease, heart disease mortality, and all-cause mortality associated with LV remodeling have been apparent. However, the reasons for this increased risk are still not understood, as outlined by Benjamin and Levy in a journal article titled "Why is left ventricular hypertrophy so predictive of morbidity and mortality?"[4] In order to hypothesize answers to this question, it is necessary to understand the physiology of the left ventricle.

The left ventricle experiences excess wall stress resulting from two mechanical factors: increased pressure load and increased volume load. Increased pressure load is marked by high end-systolic pressure in the left ventricle, while increased volume load is marked by high end-diastolic pressure. In response to these stressors, the heart undergoes two compensatory adaptations: 1) the myocardium in the left ventricle hypertrophies, to better adapt to increased pressure load and 2) the left ventricle passively stretches, altering in dimension in order to lessen stress from volume load. These two adaptations of the heart to sustained hypertension are known collectively as LV remodeling and are measured by LVM and RWT, respectively. Physiological processes of LV remodeling include mechanical stressors such as increased volume and pressure loads from hypertension[16, 17, 20], as well as neurohumoral factors such as increased catecholamine, growth hormone, or thyroxine, and a stimulated renin-angiontensinogen system.[20, 21]

The physiology of LV remodeling is often described as a result of increased hemodynamic pressure and load, as it was described above. In support of this, prospective studies have confirmed that hypertension is a predictor of LV remodeling. However, the converse has been found as well, increased LVM is a predictor of incident hypertension.[22, 23] This bi-directional association implies that both increased hypertension and increased LVM and RWT may have a common cause beyond the known risk factors of male sex, increased age, body mass index (BMI) (kg/m^2), and dietary salt intake (Figure 1.1).[4]

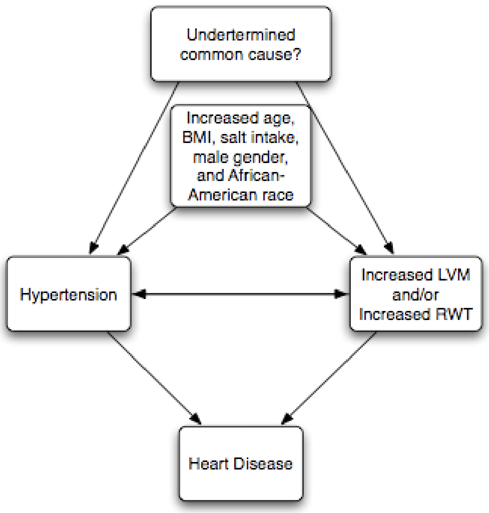

Undertermined common cause?
Increased age, BMI, salt intake, male gender, and African-American race
Hypertension
Increased LVM and/or Increased RWT
Heart Disease

Figure 1.1. Causal diagram depicting relationship between risk factors, hypertension, left ventricular traits, and heart disease.

Clinical Measurement and Classification of Left Ventricular Remodeling

LVM and RWT can be measured using electrocardiograms, echocardiography, or magnetic resonance imaging. Echocardiography is considered the gold standard in large epidemiological cohorts because it is non-invasive, cost-effective, and measures of LVM and RWT via echocardiography have been validated.[24-26] In addition to LVM and RWT, echocardiography provides other measures of left ventricular structure and function such as ventricular dimensions, aortic root diameter, ejection fraction, and stroke volume.

The quantitative trait LVM is used clinically to diagnose left ventricular hypertrophy (LVH). Because heart size is highly dependent on body size, it is standard practice in clinical settings to index LVM to body size (body surface area) or height (ex. meters, meters^2, meters$^{2.7}$) in order to more appropriately identify increased LVM relative to the size of an individual. When measured using two-dimensionally guided M-mode echocardiography, the correlation between different LVM indices ranged between 0.90-0.99 in a cohort of largely hypertensive, African-American subjects.[27] LVM index using meters$^{2.7}$ is a common indexing method because it corrects for height without artificially altering the relation of LVM to overweight[28], and also negates gender differences in LVM that are expected when using a naïve LVM measure or other indices.[29] Thresholds for LVH have been largely determined in non-Hispanic white populations; however a study conducted in the African-American cohort of the ARIC study confirmed that a LVM index ≥51 g/m$^{2.7}$ threshold is appropriate for classifying LVH in both whites and African-Americans.[29]

Because geometric patterns of LVH have been shown to stratify risk further than by presence of LVH alone,[19] a threshold value for RWT (>0.43) is used to classify concentric geometric remodeling.[21] The two classifications are considered in conjunction to define four geometric patterns of left ventricular remodeling: concentric LVH (high LVM and RWT), eccentric LVH (high LVM and normal RWT), concentric remodeling (normal LVM with high RWT), and normal LV geometry. The highest risk of heart disease outcomes is associated with concentric LVH.[3, 18]

Risk Factors for Left Ventricular Remodeling

Risk factors for increased LVM and RWT are well documented in multiethnic studies and are similar to the risk factors for heart disease, including older age, male sex, diabetes and increased BMI, SBP, and dietary sodium intake.[14, 30-32] Depending on the study population, these risk factors in combination account for only 25-50% of the variation in LVM.[4] In the African-American cohort of GENOA used in this study, age, gender, BMI, diabetes status and SBP only account for 33% of the variability of LVM and 14% of the variability in RWT. Therefore, up to 67% of the inter-individual variation of LVM and 86% of RWT remains unexplained in this study cohort.

Genetic Studies of Left Ventricular Traits

The unexplained variation in LVM and RWT has prompted researchers to consider genetic contributions to LVM and RWT that may act independently of, or in conjunction with, already established risk factors. Heritability studies have been conducted in order to first determine the proportion of variance in the crude measure of LVM (ie. not indexed to any body size measurement) that is attributable to additive genetic effects. In non-Hispanic white populations, heritability estimates for LVM have ranged between 0.36-0.59, after adjustment for known risk factors.[33-36] Heritability of LVM within African-American populations is estimated at approximately 0.46.[36] Heritability of RWT was found to range between 0.17-0.22 in an American Indian population when adjusted for risk factors.[37] Arnett and colleagues found that the adjusted sibling correlations in African-Americans ranged from 0.29-0.44 for LVM and 0.04-0.12 for RWT.[38] However, the same study found the opposite relationship in non-Hispanic whites; sibling correlations were higher for RWT than LVM, which indicates that genetic contributions to LVM and RWT may vary across racial groups.[38]

Historically, linkage studies are conducted as a follow-up to heritability studies in order to identify chromosomal locations for loci contributing to disease. Interestingly, until recently, there were no published linkage studies for LVM or RWT in humans. As part of a larger investigation of genome-wide contributions to numerous echocardiographic traits, the Framingham Heart Study reported that LVM had the highest linkage peak of all the echocardiogram traits on chromosome 5 with a LOD score of 4.38.[35] The first, and only, published linkage study for LVM index ($g/m^{2.7}$) in African-Americans is currently under peer review for publication.[39]

An alternative strategy to linkage analysis is the gene-disease association study, which provides greater statistical power for identifying genetic variation underlying diseases of multi-factorial etiology.[5, 40] Numerous candidate gene association studies have been conducted for LVM and various LVM indexed measures (LVMI). These studies have concentrated on genes in pathways known to be involved in hypertension or LV remodeling (ex. the renin-angiotensin system). Table 1.1 summarizes a sample of findings from candidate gene association studies for LVM and LVMI. It is clear that consistency across studies and validation of results in independent datasets is lacking. Inconsistent results across studies could be due to analytical aspects such as different indices used for LVM, different adjustment variables used in statistical models, or variability in sampling protocols for subject recruitment across studies. On the other hand, inconsistent results may be expected because the distribution of genetic and environmental risk factors contributing to the architecture of complex diseases is likely to be population specific.[41] Therefore, genetic loci that are identified as significantly associated with an outcome in one population may not show a similar effect in a genetically or environmentally different population. This replication issue is not unique to studies of LV remodeling, as most published gene-disease associations using the candidate gene approach have not been replicated.[42]

Study Population

Each analytical chapter of this dissertation describes the study subjects used in analyses; however here is an overview of the study population used for this dissertation. The National Heart Lung and Blood Institute established the Family Blood Pressure Program (FBPP) in 1996, which joined existing research networks that were investigating hypertension and cardiac diseases.[43] One of the four networks in FBPP is the Genetic Epidemiology Network of Arteriopathy (GENOA), which recruited hypertensive African-American (Jackson, MS) and non-Hispanic white (Rochester, MN) sibships and diabetic Hispanic (Starr County, TX) sibships for linkage analysis and family-based association studies to investigate genetic contributions to hypertension and hypertensive target organ damage.

Sibships containing at least two individuals with clinically diagnosed essential hypertension before age 60 were recruited in Jackson, MS. Written informed consent was obtained from all subjects and approval was granted by participating institutional review boards. Participants were diagnosed with hypertension if they had either 1) a previous clinical diagnosis of hypertension by a physician with current anti-hypertensive treatment, or 2) an average SBP ≥140 mmHg or DBP ≥90 mmHg on the second and third clinic visit. Exclusion criteria were secondary hypertension, alcoholism or drug abuse, pregnancy, insulin-dependent diabetes mellitus, or active malignancy.

GENOA data was collected in two phases. Phase I (1996-1999) and Phase II (2000-2004) data consist of demographic information, medical history, clinical characteristics, lifestyle factors, and blood samples for genotyping and biomarker assays. Approximately 80% of Phase I Jackson, MS participants were successfully re-recruited for Phase II at which time echocardiograms were performed to measure target organ damage. Of 1,482 Phase II Jackson, MS participants, 1,440 from 620 sibships have echocardiogram data and will be used for this dissertation.

Aims of Dissertation

The overall goal of this proposed dissertation is to investigate the genetic architecture of clinical measures for LV remodeling (LVM and RWT) in the African-American cohort of the Genetic Epidemiology Network of Arteriopathy (GENOA) study. Variation in both LVM and RWT have been shown to have substantial genetic contributions, have been studied extensively using the candidate gene approach, and are the result of complex processes that likely involve many genetic loci with small or modest effects. In an attempt to develop a deeper understanding of the genetic architecture of these complex traits, I will use both candidate gene and genome-wide approaches to test for main single nucleotide polymorphism (SNP) effects and SNP-environment interactions associated with LVM and RWT variation.

Aim1: Apply a candidate gene approach to identify SNPs associated with the quantitative measures LVM and RWT.

The pathways contributing to the genetic architecture of LV remodeling may involve genes related to LV structure, cell signal transduction, hormones, growth factors, calcium homeostasis, and blood pressure.[21] Three hundred and ninety-five SNPs were already genotyped in 80 candidate genes in the African-American cohort of GENOA. Therefore, as a first approach to understanding the genetics of LV remodeling, I took a candidate gene approach using these 395 SNPs to identify SNPs with main effects associated with LVM and RWT that replicated within sub-samples of the African-American GENOA cohort.[44]

Aim 2: Evaluate 1,878 SNPs from 268 candidate genes for gene x environment and gene x gene interactions associated with variation in LVM.

The genetic architecture of quantitative, complex traits such as LVM will likely involve interactions between genetic and environmental factors. While context dependent genetic effects underlying complex traits are likely commonplace, methods to best detect interactions analytically are constantly under development and debate.[45, 46] If a genetic effect is dependent upon a certain environmental characteristic being present or absent, the genetic marker would not necessarily exhibit marginal statistical effects.[45] Therefore, we screened a total of 1,878 SNPs from 268 genes for potential interactions with environmental factors and all possible pairwise combinations of SNP-SNP interactions, regardless of the presence or absence of marginal SNP effects. "Environments" that will be considered for context dependent SNP effects on LVM include age, sex, height, weight, SBP, DBP, hypertension status, diabetes status, history of myocardial infarction, taking anti-hypertensive medications, duration of anti-hypertensive medication use, smoking status, total cholesterol, low density lipoprotein cholesterol, and triglycerides.

<u>Aim 3</u>: Conduct a GWAS to identify SNPs throughout the genome associated with LVM and RWT.

The candidate gene approach is grounded in *a priori* knowledge underlying the physiological mechanisms for LV remodeling; however candidate gene association studies for many common, complex traits have not been replicated.[42] In addition, candidate gene studies by definition will fail to identify any genetic contributions that are outside of established pathways.[5] This is especially important because recent genome-wide research for diseases such as prostate cancer and coronary heart disease has confirmed that replicable genetic effects can be found in chromosomal regions that are not within candidate genes.[10, 11, 47, 48] Not only are these replicated genetic effects outside of candidate genes, but they are in non-coding sequences of the genome. Another new finding supporting the notion of looking for genetic effects outside candidate genes is that regions of the genome previously thought to be "noise" were recently identified as "pervasively transcribed" and may therefore have a large role in function.[49]

As part of an NIH funded grant (NIH - HL087660), the African-American cohort of GENOA was genotyped for 1.8 million genome-wide markers, almost 900,000 of which are SNPs, on the Affymetrix® Genome-Wide Human SNP Array 6.0.[50] For Aim 3, these genome-wide SNPs were tested for statistical associations with variation in LVM and RWT.

Overview of Dissertation

This dissertation consists of a total of seven chapters, five of which contain analyses and inferences. Chapter 1 has provided background information regarding the traits of interest and their public health significance. Chapter 2 is an investigation into the heritability and genetic correlations of LVM and RWT. Chapter 3 is a candidate gene association study of main SNP effects associated with LVM and RWT and has already been published in the journal *Hypertension* (**Aim 1**).[44] Chapter 4 investigates population substructure in the African-American cohort of GENOA and is a necessity for analyses conducted in Chapters 5 and 6. Chapter 5 investigates candidate gene SNPs for main and interaction effects in association with LVM with special attention to population substructure and multiple testing issues (**Aim 2**). The final analytical chapter is Chapter 6, a GWAS to identify SNPs associated with LVM and RWT with replicated effects in an independent study cohort (**Aim 3**). Chapter 7 provides insight into, and integrates, this extensive body of research into the genetic architecture of LVM and RWT.

Table 1.1. Summary of select candidate-gene association studies for increased left ventricular mass. Note: these references are not found in the "References" section of this dissertation but are available upon request

Gene	Statistically significant association found	No evidence for statistically significant association
ACE insertion/deletion	Iwai et al. 1994	Gomez-Angelats E et al. 2000
	Saeed M et al. 2005	Lindpaintner K et al. 1996
		Shlyakhto et al. 2001
		Kauma H et al. 1998
Angiotensin II Type-2 Receptor (AT_2-R)	Schmider et al. 2001	
Angiotensinogen M235T	Tang W et al. 2002	Kauma et al. 1998
		Rasumussen-Torvik et al. 2005
Aldosterone Synthase (*CYP11B2*)	Kupari M et al. 1998	Delles C et al. 2001
	Stella P et al. 2004	Schunkert H et al. 1999
		Swan L et al. 2002
ANP	Rubattu et al. 2006	
APOE	Yilmaz et al. 2001	
IGF-1	Bleumink GS et al. 2005	
eNOS	Takaoka M et al. (abstract)	Karvonen J et al. 2002
GNB3		Shlyakhto EV et al. 2002
		Sedlácek K et al. 2002
		Swan L et al. 2002
ESR1 (repeat polymorphism)	Leibowitz D et al. 2006	
ESR2	Peter I et al. 2005	

References

1. National Center for Health Statistics. Health, 2006: with chartbook on trends in the health of Americans. Hyattsville, MD: 2006.

2. Levy D, Garrison RJ, Savage DD, Kannel WB, Castelli WP. Prognostic implications of echocardiographically determined left ventricular mass in the Framingham Heart Study. N Engl J Med. 1990; 322(22):1561-1566.

3. Koren MJ, Devereux RB, Casale PN, Savage DD, Laragh JH. Relation of left ventricular mass and geometry to morbidity and mortality in uncomplicated essential hypertension. Ann Intern Med. 1991; 114(5):345-352.

4. Benjamin EJ, Levy D. Why is left ventricular hypertrophy so predictive of morbidity and mortality? Am J Med Sci. 1999; 317(3):168-175.

5. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. Nat Rev Genet. 2005; 6(2):95-108.

6. Risch NJ. Searching for genetic determinants in the new millennium. Nature. 2000; 405(6788):847-856.

7. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. Science. 2008; 322(5903):881-888.

8. Collins FS, Guyer MS, Charkravarti A. Variations on a theme: Cataloging human DNA sequence variation. Science. 1997; 278(5343):1580-1581.

9. NCI-NHGRI Working Group on Replication in Association Studies, Chanock SJ, Manolio T, et al. Replicating genotype-phenotype associations. Nature. 2007; 447(7145):655-660.

10. McPherson R, Pertsemlidis A, Kavaslar N, et al. A common allele on chromosome 9 associated with coronary heart disease. Science. 2007; 316(5830):1488-1491.

11. Samani NJ, Erdmann J, Hall AS, et al. Genomewide association analysis of coronary artery disease. N Engl J Med. 2007; 357(5):443-453.

12. Hertz RP, Unger AN, Cornell JA, Saunders E. Racial disparities in hypertension prevalence, awareness, and management. Arch Intern Med. 2005; 165(18):2098-2104.

13. Kramer H, Han C, Post W, et al. Racial/ethnic differences in hypertension and hypertension treatment and control in the Multi-Ethnic Study of Atherosclerosis (MESA). Am J Hypertens. 2004; 17(10):963-970.

14. Kizer JR, Arnett DK, Bella JN, et al. Differences in left ventricular structure between black and white hypertensive adults: The Hypertension Genetic Epidemiology Network study. Hypertension. 2004; 43(6):1182-1188.

15. Liao Y, Cooper RS, McGee DL, Mensah GA, Ghali JK. The relative effects of left ventricular hypertrophy, coronary artery disease, and ventricular dysfunction on survival among black adults. JAMA. 1995; 273(20):1592-1597.

16. Devereux RB, de Simone G, Ganau A, Roman MJ. Left ventricular hypertrophy and geometric remodeling in hypertension: Stimuli, functional consequences and prognostic implications. J Hypertens Suppl. 1994; 12(10):S117-27.

17. van der Wall EE, van der Laarse A, Pluim BM, Bruschke AVG (eds). Left Ventricular Hypertrophy: Physiology Versus Pathology. Dordrecht, The Netherlands: Kluwer Academic Publishers, 1999.

18. Ghali JK, Liao Y, Simmons B, Castaner A, Cao G, Cooper RS. The prognostic role of left ventricular hypertrophy in patients with or without coronary artery disease. Ann Intern Med. 1992; 117(10):831-836.

19. Taylor HA, Penman AD, Han H, et al. Left ventricular architecture and survival in African-Americans free of coronary heart disease (from the Atherosclerosis Risk in Communities [ARIC] study). Am J Cardiol. 2007; 99(10):1413-1420.

20. Opie LH. Heart Physiology: From Cell to Circulation. Fourth Edition ed. Philadelphia: Lippincott Williams & Wilkins, 2004.

21. Arnett DK. Genetic contributions to left ventricular hypertrophy. Curr Hypertens Rep. 2000; 2(1):50-55.

22. Post WS, Larson MG, Levy D. Impact of left ventricular structure on the incidence of hypertension. The Framingham Heart Study. Circulation. 1994; 90(1):179-185.

23. de Simone G, Devereux RB, Roman MJ, Schlussel Y, Alderman MH, Laragh JH. Echocardiographic left ventricular mass and electrolyte intake predict arterial hypertension. Ann Intern Med. 1991; 114(3):202-209.

24. Devereux RB, Alonso DR, Lutas EM, et al. Echocardiographic assessment of left ventricular hypertrophy: Comparison to necropsy findings. Am J Cardiol. 1986; 57(6):450-458.

25. Arnett DK, de las Fuentes L, Broeckel U. Genes for left ventricular hypertrophy. Curr Hypertens Rep. 2004; 6(1):36-41.

26. Palmieri V, Dahlof B, DeQuattro V, et al. Reliability of echocardiographic assessment of left ventricular structure and function: The PRESERVE study.

prospective randomized study evaluating regression of ventricular enlargement. J Am Coll Cardiol. 1999; 34(5):1625-1632.

27. Liao Y, Cooper RS, Durazo-Arvizu R, Mensah GA, Ghali JK. Prediction of mortality risk by different methods of indexation for left ventricular mass. J Am Coll Cardiol. 1997; 29(3):641-647.

28. de Simone G, Daniels SR, Devereux RB, et al. Left ventricular mass and body size in normotensive children and adults: Assessment of allometric relations and impact of overweight. J Am Coll Cardiol. 1992; 20(5):1251-1260.

29. Nunez E, Arnett DK, Benjamin EJ, et al. Optimal threshold value for left ventricular hypertrophy in blacks: The Atherosclerosis Risk in Communities study. Hypertension. 2005; 45(1):58-63.

30. Galderisi M, Anderson KM, Wilson PW, Levy D. Echocardiographic evidence for the existence of a distinct diabetic cardiomyopathy (the Framingham Heart Study). Am J Cardiol. 1991; 68(1):85-89.

31. Bella JN, Wachtell K, Palmieri V, et al. Relation of left ventricular geometry and function to systemic hemodynamics in hypertension: The LIFE study. Losartan intervention for endpoint reduction in hypertension study. J Hypertens. 2001; 19(1):127-134.

32. du Cailar G, Ribstein J, Mimran A. Dietary sodium and target organ damage in essential hypertension. Am J Hypertens. 2002; 15(3):222-229.

33. Swan L, Birnie DH, Padmanabhan S, Inglis G, Connell JM, Hillis WS. The genetic determination of left ventricular mass in healthy adults. Eur Heart J. 2003; 24(6):577-582.

34. Sharma P, Middelberg RP, Andrew T, Johnson MR, Christley H, Brown MJ. Heritability of left ventricular mass in a large cohort of twins. J Hypertens. 2006; 24(2):321-324.

35. Vasan RS, Larson MG, Aragam J, et al. Genome-wide association of echocardiographic dimensions, brachial artery endothelial function and treadmill exercise responses in the Framingham Heart Study. BMC Med Genet. 2007; 8 Suppl 1:S2.

36. de Simone G, Tang W, Devereux RB, et al. Assessment of the interaction of heritability of volume load and left ventricular mass: The HyperGEN offspring study. J Hypertens. 2007; 25(7):1397-1402.

37. Bella JN, MacCluer JW, Roman MJ, et al. Heritability of left ventricular dimensions and mass in American Indians: The Strong Heart Study. J Hypertens. 2004;

22(2):281-286.

38. Arnett DK, Hong Y, Bella JN, et al. Sibling correlation of left ventricular mass and geometry in hypertensive African Americans and whites: The HyperGEN study. Hypertension genetic epidemiology network. Am J Hypertens. 2001; 14(12):1226-1230.

39. Arnett DK. Personal communication.

40. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science. 1996; 273(5281):1516-1517.

41. Sing CF, Stengard JH, Kardia SL. Dynamic relationships between the genome and exposures to environments as causes of common human diseases. World Rev Nutr Diet. 2004; 93:77-91.

42. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. Genet Med. 2002; 4(2):45-61.

43. FBPP Investigators. Multi-center genetic study of hypertension: The Family Blood Pressure Program (FBPP). Hypertension. 2002; 39(1):3-9.

44. Meyers KJ, Mosley TH, Fox E, et al. Genetic variations associated with echocardiographic left ventricular traits in hypertensive blacks. Hypertension. 2007; 49(5):992-999.

45. Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ. Exploiting gene-environment interaction to detect genetic associations. Hum Hered. 2007; 63(2):111-119.

46. Thornton-Wells TA, Moore JH, Haines JL. Genetics, statistics and human disease: Analytical retooling for complexity. Trends Genet. 2004; 20(12):640-647.

47. Yeager M, Orr N, Hayes RB, et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nat Genet. 2007; 39(5):645-649.

48. Gudmundsson J, Sulem P, Manolescu A, et al. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. Nat Genet. 2007; 39(5):631-637.

49. ENCODE Project Consortium, Birney E, Stamatoyannopoulos JA, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007; 447(7146):799-816.

50. Affymetrix. Affymetrix ® Genome-Wide Human SNP Array 6.0 2007.

Chapter 2

Heritability and Genetic Correlations of Left Ventricular Mass and Relative Wall Thickness in the African-American Cohort of GENOA

Abstract

The heritability of LVM and RWT has been estimated in various populations, but not specifically in the African-American cohort of GENOA. Here we explore the heritability of LVM and RWT of this population and conduct a bivariate analysis to test for evidence of pleiotropy. The heritability of LVM after adjusting for age, sex, height, weight, SBP, and diabetes is 0.416 (SE=0.072, p-value=2.29x10^{-10}) and the heritability of RWT adjusted for age, sex, height, weight, SBP, and diabetes is 0.235 (SE=0.07, p-value=0.0001). The genetic correlation of covariate adjusted LVM and RWT is 0.181 (SE=0.162), which is not statistically different from zero (p-value=0.291). There is significant statistical evidence that genes influence variation in LVM and RWT; however there is no significant statistical evidence for shared genes influencing variation in LVM and RWT.

Introduction

Central to biology, medicine, and public health is to understand causes of observed variation in a trait, and whether it is due to environmental factors and/or genetic factors. The population specific parameter that estimates the contribution of genetic variation to variation in a trait is called “heritability”. Heritability is often one of the first parameters estimated when conducting genetic epidemiological research because it allows us to test the null hypothesis that genes are not involved in the observed variation. Knowing the heritability of traits can aid in understanding the relative impact of genes on the outcome of interest and perhaps prioritize genetic association and linkage studies based on traits with higher heritability. As was described in Chapter 1 of this dissertation, many studies have estimated the heritability of LVM and RWT with varying estimates. Since heritability is a population specific parameter, it is likely to vary across populations and over time. Heritability is influenced by allele frequencies in a population, effect sizes of the variants, mode of genetic action, variation in environmental forces, pedigree size and structure, measurement error of trait, and the total phenotypic variation observed.[1, 2] Below, I outline what heritability is, how it is estimated, how to interpret it, and report heritability estimates of LVM and RWT from the GENOA cohort. I also discuss genetic correlation between LVM and RWT and whether there is evidence for shared genetic effects between the two traits.

Methods

Heritability

Heritability is grounded in the principles of quantitative genetics and polygenic inheritance and is defined as the proportion of variance in a trait due to variability in genetic factors.[1, 3] Traditionally, the resemblance of a given trait between relatives is used to estimate the heritability. No actual measures of genetic variation are needed because Mendelian laws of inheritance indicate the levels of genetic sharing between relatives, which will affect the covariance of a trait between relatives who share a known proportion of genes in common.[4]

Variation in a quantitative trait can be modeled as a function of genetic and environmental factors using family members and the relative importance of each component (genetic or environmental) towards the total variation observed in a given population.[1, 2] This type of modeling is a reflection of the "nature vs. nurture" debate focused on the simple question: how much variation in a trait is determined by "nature" (genes) and how much is determined by "nurture" (environment).

In traditional quantitative genetic modeling, the mean value of a quantitative phenotype (P) is considered to be a function of the genotypic value of genes underlying the trait (G) and the environment (E): $P = G + E$. Because of the additive property of variances of independent variables, the total variance (V) of a quantitative trait can be partitioned into variance due to genetic factors and variance due to environment: $V_P = V_G + V_E$, as long as there is no evidence of GxE covariation. If genetic factors are modeled to reflect the different modes of gene action (i.e. additive, dominant, and interaction effects), a more detailed variance component decomposition is $V_G = V_A+V_D+V_I$. Historically, the V_G component has been estimated as a part of animal and plant breeding programs focused on the selection of agriculturally relevant traits. In plants and animals, neither the dominance variance nor the gene-gene interaction variance can respond to breeding selection, the selection response is determined only by the additive effects. The reason for this is because only one allele is passed on from parent to child (i.e. "inherited") and polygenic inheritance assumes these alleles combine in a purely additive manner. The dominance action of alleles on a phenotype only occurs once the two alleles from each parent have been combined in the child, therefore dominance is not directly "inherited".[4] Thus, the additive genetic variance (V_A) is the portion of genetic effects of interest in most heritability studies and has been likewise the focus of many human studies.

This focus on the heritability of additive genetic effects is referred to as the narrow sense heritability (h^2), which is defined as the proportion of total phenotypic variance due to purely additive genetic effects: $h^2 = V_A/V_P$.[1] Alternatively, when considering all types of genetic effects, this is referred to as broad sense heritability (H^2) is estimated by V_G/V_P. For the most part, in human populations, broad sense heritability is very difficult to estimate and is more of a theoretical concept.[1, 4] An approximation of H^2 can be estimated from studies comparing the correlation of monozygotic (MZ) and dizygotic (DZ) twins ($H^2=2(r_{MZ}-r_{DZ})$). However this estimate is based on the unrealistic assumption that the two types of twins share the same degree of environment.[5] From here on, when I refer to "heritability", it is in the narrow sense.

Estimating heritability is relatively straightforward and can be done with only observed phenotypic data from families (provided the genetic relationship between family members is known), making it an appealing, inexpensive first step of genetic epidemiological analyses. Using analysis of variance (ANOVA), the variance of an observed phenotype is partitioned into the variance between families ($\sigma^2_{\mathbf{B}}$) and the variance within families ($\sigma^2_{\mathbf{W}}$).[6] Because these are observed variances from a study population, we will use σ^2 to represent the true variance parameter, V, of each component. The degree of phenotypic resemblance in families can then be expressed as the proportion of total observed variance ($\sigma^2_{\mathbf{B}}+\sigma^2_{W}$) due to the between family component of variance, $\sigma^2_{\mathbf{B}}$. This is the intra-class correlation coefficient: $r = \sigma^2_{\mathbf{B}}/(\sigma^2_{\mathbf{B}}+\sigma^2_{W})$.[6] Multiplying the intra-class correlation by the reciprocal of the proportion of genes in common for that kinship type (i.e. full siblings, half siblings) provides an estimate of heritability.[1] For example, when using full-siblings, the intra-class correlation is multiplied by the reciprocal of the proportion of genes shared between full siblings on average. For full siblings, $h^2 \approx 2r$, because on average full siblings share ½ of their genes.

Estimating heritability in samples with mixed kinships can quickly become complicated using ANOVA because ANOVA assumes all families have the same genetic structure. Therefore, software packages are available that estimate heritability using regression techniques and can easily account for varying kinships and family size within a single sample. One such program is SOLAR (Sequential Oligogenic Linkage Analysis Routines) [7], which implements a variance component regression method where the outcome, y, is modeled for each individual i as follows:

$$y_i = \mu + \Sigma\beta_j X_{ij} + g_i + e_i$$

where μ is the mean of y in the population, X_{ij} is the j-th covariate with associated regression coefficient β_j, g_i is the additive genetic effect (normally distributed with mean 0, variance σ^2_g), and e_i is a random residual effect (normally distributed with mean 0, variance σ^2_e). The sum of variances $\sigma^2_g+\sigma^2_e$ is assumed to equal 1. Variance due to any non-additive or unmeasured environmental component of variance is included in the e_i term. Measurement error is also included in this term. Based on this equation, σ^2_g provides the estimate of heritability after accounting for adjustment covariates. Statistical significance of the heritability is then estimated using likelihood ratio test statistics comparing the full model to a model with heritability restricted to zero.

Genetic Correlations and Pleiotropy

Another concept central to quantitative genetics is that correlated phenotypic traits may have shared genetic causes.[8] The action of one gene on two or more traits is referred to as pleiotropy. The presence and degree of pleiotropy can be estimated with genetic correlation. Similar to the description above where the variance of a trait is partitioned to estimate heritability, the genetic correlation for two traits can be partitioned into the covariance (cov) due to genetic effects and environmental effects: $cov_P = cov_G + cov_E$, assuming no G-E correlation.[1]

Tests for genetic correlations can also be conducted in SOLAR using maximum likelihood estimation methods.[9, 10] In SOLAR, the phenotypic correlation (ρ) between two traits is derived based on the genetic correlation (ψ_g), environmental correlation (ψ_e), and the heritability of the two traits.

$$\rho = \left[\sqrt{h_1^2 h_2^2} \times \psi_g\right] + \left[\sqrt{1 - h_1^2} \times \sqrt{1 - h_2^2} \times \psi_e\right]$$

Testing for pleiotropy is a two-step process. The first statistical test has a null hypothesis that the genetic correlation is equal to zero. A likelihood ratio test statistic calculated as the difference in -2 x ln likelihoods between a restricted model (where genetic correlation = 0) and an unrestricted model (where all parameters are estimated) provides the statistical significance for pleiotropy. If the genetic correlation is significantly different from zero, there is statistical evidence for pleiotropy. The second statistical test is to test if the genetic correlation is different from one. This determines the magnitude of shared genetic effects. If the genetic correlation is not statistically different from one, then all genes influencing one trait are assumed to influence the other trait as well. Similarly, if the environmental correlation is different from zero, there is statistical evidence of shared environmental factors influencing the traits other than those adjusted for.

Results

Heritability of LVM and RWT in GENOA

There are 1,440 African-Americans from 620 families with LVM and RWT measures. The distribution of siblings in families can be found in Table 2.1. The most frequently occurring sibship size was two (232/620 = 37% of the families had 2 siblings). Because the distribution of LVM and RWT are right skewed (Figures 2.1 and 2.2), both were transformed using the natural logarithm in order to approximate a normal distribution. Heritability was calculated for LVM and RWT without any adjustment for covariates and then adjusted for potentially confounding risk factors including age, sex, systolic blood pressure (SBP), height, weight, and diabetes status.

Analyses were conducted using a variance components approach implemented in SOLAR version 4.1.9.[7] The estimate of h^2 for unadjusted LVM was 0.550 (SE=0.070, p-value=7.9×10^{-17}). After including age, sex, height, weight, SBP, and diabetes into the SOLAR model, the estimate of h^2 for adjusted LVM was reduced, but remained statistically significantly different from zero with an estimate of 0.416 (SE=0.072, p-value=2.29×10^{-10}).

The estimate of h^2 for unadjusted RWT was 0.469 (SE=0.070, p-value=1.72×10^{-13}). After including age, sex, height, weight, SBP, and diabetes into the SOLAR model, the estimate of h^2 for adjusted RWT was also reduced, but remained statistically significantly different from zero with an estimate of 0.235 (SE=0.070, p-value=0.0001).

Genetic Correlations between LVM and RWT in GENOA

LVM and RWT are biologically related traits and are increased in response to mechanical stimuli on the left ventricle.[11, 12] Furthermore, LVM and RWT are used jointly to classify left ventricular remodeling patterns and predict risks of heart failure and sudden death.[13, 14] The Pearson correlation coefficient for logLVM and logRWT was estimated in the statistical analysis software R to be 0.244 in the African-American cohort of GENOA after adjusting both outcome variables for age, sex, height, weight, SBP and diabetes status (Figure 2.3). Given the biological rationale for genetic correlation and evidence for positive phenotypic correlation based on the Pearson correlation coefficient, we hypothesize that there are likely shared genes influencing both LVM and RWT. In order to test this hypothesis, we estimated bivariate genetic correlations between covariate-adjusted LVM and RWT using maximum likelihood methods as implemented in SOLAR.[9]

After adjusting both LVM and RWT for age, sex, height, weight, SBP, and diabetes, the genetic correlation between LVM and RWT was 0.181 (SE= 0.162). The likelihood ratio test statistic showed this estimate for genetic correlation was not statistically significantly different from zero (p-value=0.291). The estimate for environmental correlation was 0.278 (SE=0.069). The environmental correlation was statistically significantly different from zero (p-value=0.00025) indicating shared environmental influences of LVM and RWT beyond the adjustment covariates.

Discussion

Epidemiological studies have shown the heritability of LVM to be in the range of 0.36 – 0.59, depending on the population and adjustment covariates.[15-19] Only one other study to our knowledge has looked at the heritability of RWT; the Strong Heart Study of

American Indians estimated the heritability of RWT to be approximately 0.20.[19] In our study of African-Americans in GENOA, the estimates of heritability for LVM of 0.416 and 0.235 for RWT after adjusting for known risk factors are consistent with estimates reported by others. This consistency is to be expected because even though heritability estimates are population specific, they have been found to be relatively reliable across populations for numerous traits.[2]

A limitation of heritability is that it is estimated under a set of unrealistic assumptions. These assumptions include no gene-environment correlation, no gene-environment or gene-gene interactions, random mating, no selection, and no environmental transmission from parent to offspring.[4] Furthermore, heritability assumes a polygenic model, meaning there are many genetic loci contributing to variation in a trait, each loci with small, equal, and additive effects.[4] With complex, quantitative traits, it is well accepted that the phenotype is the result of genes, environment, and interactions between genes, between genes and environments and between environments; completely contrary to assumptions made when estimating heritability. While it might seem irrelevant to even estimate heritability because of this, Visscher and colleagues make an argument in a recent review paper that as long as researchers have an appropriate understanding of what heritability means.[2] Heritability is still a very useful tool for understanding response to selection in evolutionary biology, prediction of disease risk in medicine, prioritizing epidemiological studies, and will allow us to better unravel the interplay between genes and environment.[2] Furthermore, a recent review paper outlining the future of genetic mapping studies in the post genome-wide association era even recommended future sequencing studies be prioritized based on higher heritabilities because of the large cost of sequencing.[20]

Despite being biologically and clinically related traits, our results indicate no evidence of pleiotropic genetic effects for LVM and RWT in the GENOA sample of African Americans with a very high prevalence of hypertension. We repeated the analysis limiting the sample to those between the ages of 45 and 80 years and repeated the analysis excluding the most extreme values of LVM and RWT as outliers. These modifications to the sample did not substantially alter the estimates of genetic correlations or heritability. While it is possible that LVM and RWT do not share genes, there are also limitations to using this approach. The genetic correlation is based on the "net" effect of all segregating genes influencing both phenotypes.[1] If a single gene increases LVM and decreases RWT, these effects may cancel each other and result in a genetic correlation of zero.[1]

Although there was no evidence for shared genetic effects, the heritability of LVM and RWT indicate that each trait has substantial genetic effects underlying the variation in the GENOA African-American cohort. Heritability does not provide any insight into the genetic architecture of specific genes involved in a trait; it only provides a relative contribution of variability in genes to variation in a trait. The remainder of this dissertation will focus on uncovering the genetic architecture of these traits via genetic association studies.

Table 2.1. Distribution of sibship size in the African-American participants of GENOA with echocardiography data.

# of Siblings in Sibship	# of Sibships	Total number of individuals	% of Total Number of Individuals
1	176	176	12.2%
2	232	464	32.2%
3	121	363	25.2%
4	50	200	13.9%
5	17	85	5.9%
6	19	114	7.9%
7	4	28	1.9%
8	0	0	0%
9	0	0	0%
10	1	10	0.7%
Totals	**620 sibships**	**1,440 individuals**	**100.0%**

Figure 2.1. Distribution of left ventricular mass (raw and transformed using the natural logarithm) in a sample of 1,440 African-Americans of the GENOA study.

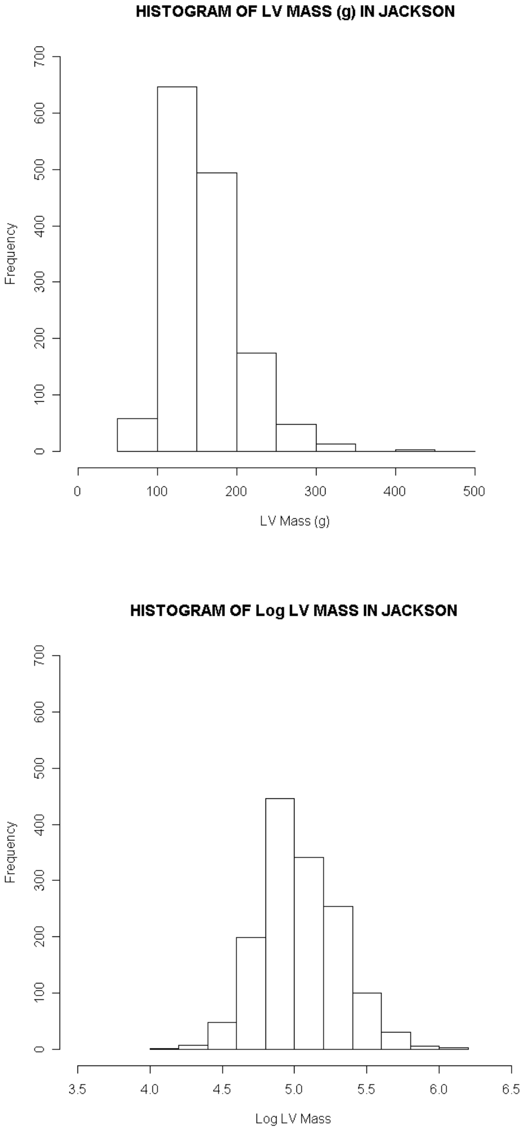


Figure 2.2. Distribution of relative wall thickness (raw and transformed using the natural logarithm) in a sample of 1,440 African-Americans of the GENOA study.

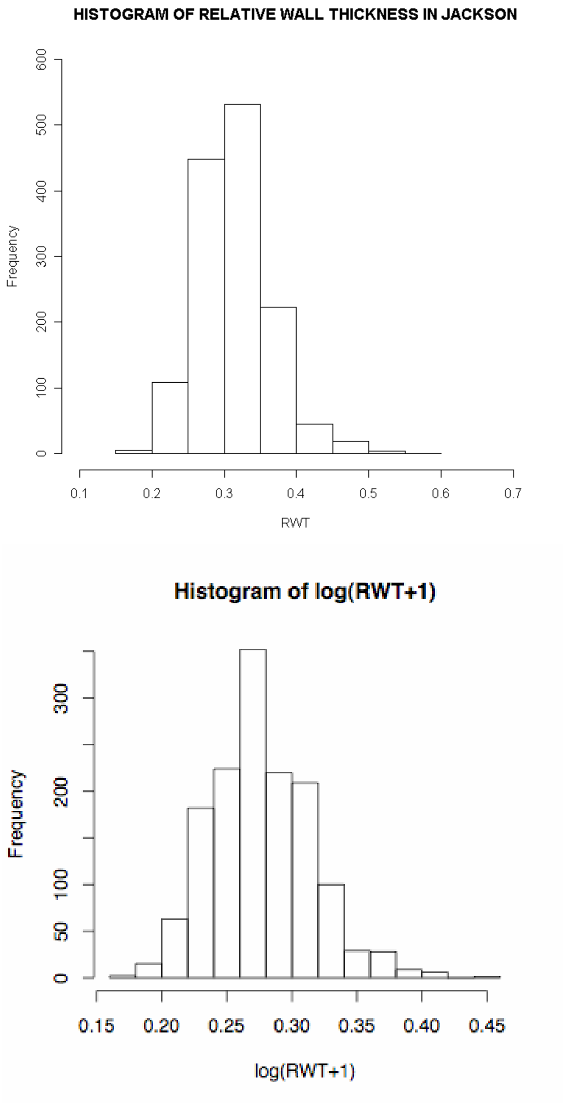
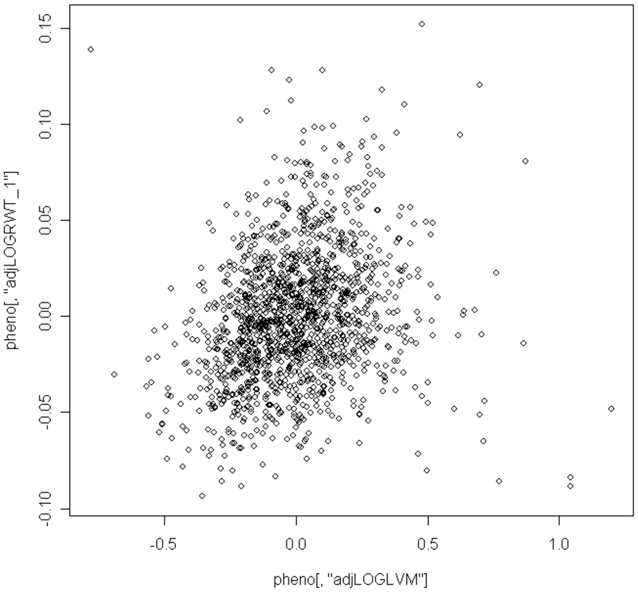


Figure 2.3. Correlation of logLVM and logRWT in the African-American cohort of GENOA (n=1,440) after adjusting both variables for age, sex, height, weight, SBP, and diabetes status. Pearson correlation coefficient, r = 0.244.



References

1. Falconer DS, Mackay TFC. Introduction to Quantitative Genetics. Fourth Edition ed. England: Pearson Prentice Hall, 1996.

2. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era--concepts and misconceptions. Nat Rev Genet. 2008; 9(4):255-266.

3. Lange K, Westlake J, Spence MA. Extensions to pedigree analysis. III. Variance components by the scoring method. Ann Hum Genet. 1976; 39(4):485-491.

4. Lynch M, Walsh B. Genetics and Analysis of Quantitative Traits. Sunderland, MA: Sinauer Associates, Inc., 1998.

5. Pierce B. Genetics: A Conceptual Approach. 2nd ed. W.H. Freeman, 2004.

6. Mather WB. Principles of Quantitative Genetics. Minneapolis, Minnesota: Burgess Publishing Company, 1964.

7. Almasy L, Blangero J. Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet. 1998; 62(5):1198-1211.

8. Falconer DS, Mackay TFC. Introduction to Quantitative Genetics. Fourth Edition ed. England: Pearson Prentice Hall, 1996.

9. Lange K, Boehnke M. Extensions to pedigree analysis. IV. Covariance components models for multivariate traits. Am J Med Genet. 1983; 14(3):513-524.

10. Boehnke M, Moll PP, Lange K, Weidman WH, Kottke BA. Univariate and bivariate analyses of cholesterol and triglyceride levels in pedigrees. Am J Med Genet. 1986; 23(3):775-792.

11. Devereux RB, de Simone G, Ganau A, Roman MJ. Left ventricular hypertrophy and geometric remodeling in hypertension: Stimuli, functional consequences and prognostic implications. J Hypertens Suppl. 1994; 12(10):S117-27.

12. van der Wall EE, van der Laarse A, Pluim BM, Bruschke AVG (eds). Left Ventricular Hypertrophy: Physiology Versus Pathology. Dordrecht, The Netherlands: Kluwer Academic Publishers, 1999.

13. Taylor HA, Penman AD, Han H, et al. Left ventricular architecture and survival in African-Americans free of coronary heart disease (from the Atherosclerosis Risk in Communities [ARIC] study). Am J Cardiol. 2007; 99(10):1413-1420.

14. Ghali JK, Liao Y, Simmons B, Castaner A, Cao G, Cooper RS. The prognostic role of left ventricular hypertrophy in patients with or without coronary artery disease. Ann Intern Med. 1992; 117(10):831-836.

15. Swan L, Birnie DH, Padmanabhan S, Inglis G, Connell JM, Hillis WS. The genetic determination of left ventricular mass in healthy adults. Eur Heart J. 2003; 24(6):577-582.

16. Sharma P, Middelberg RP, Andrew T, Johnson MR, Christley H, Brown MJ. Heritability of left ventricular mass in a large cohort of twins. J Hypertens. 2006; 24(2):321-324.

17. Vasan RS, Larson MG, Aragam J, et al. Genome-wide association of echocardiographic dimensions, brachial artery endothelial function and treadmill exercise responses in the framingham heart study. BMC Med Genet. 2007; 8 Suppl 1:S2.

18. de Simone G, Tang W, Devereux RB, et al. Assessment of the interaction of heritability of volume load and left ventricular mass: The HyperGEN offspring study. J Hypertens. 2007; 25(7):1397-1402.

19. Bella JN, MacCluer JW, Roman MJ, et al. Heritability of left ventricular dimensions and mass in American Indians: The Strong Heart Study. J Hypertens. 2004; 22(2):281-286.

20. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. Science. 2008; 322(5903):881-888.

Chapter 3

Genetic Variations Associated with Echocardiographic Left Ventricular Traits in Hypertenstive African-Americans[*]

Abstract

Echocardiographic measures of cardiac target organ damage – including left ventricular mass and relative wall thickness – are powerful predictors of heart disease morbidity and mortality. The aim of this study is to investigate whether single nucleotide polymorphisms in candidate genes for hypertension and heart disease have effects on quantitative measures of hypertensive cardiac target organ damage, independent of their actions on blood pressure levels, in a cohort of hypertensive, African-American sibships. In order to detect replication of genetic effects across samples, this study took advantage of the affected sib pair design and created two samples, each with 448 unrelated individuals. As part of the Genetic Epidemiology Network of Arteriopathy study, subjects were screened using two-dimensional echocardiography and 395 single nucleotide polymorphisms in 80 candidate genes were genotyped. Linear regression was used to test for single nucleotide polymorphisms significantly associated with left ventricular mass index (g/m$^{2.7}$) or relative wall thickness after adjusting for associated

[*] This work was previously published: Meyers KJ, Mosley TH, Fox E, Bowerwinkle E, Arnett DK, Devereux RB, Kardia SLR. Genetic variations associated with echocardiographic left ventricular traits in hypertensive blacks. *Hypertension.* 2007; 49:992.

covariates. Significant single nucleotide polymorphisms were subsequently tested for consistent directionality in genotype-phenotype relationships across samples. Three single nucleotide polymorphisms – one each in the *APOE*, *SCN7A*, and *SLC20A1* genes – were significantly associated in both samples with left ventricular mass index and had replicate genotype-phenotype relationships. One in the *ADRB1* gene was significantly associated with relative wall thickness with replicate effects in both samples. We identified genetic variation that significantly influences left ventricular traits with replicable effects in a cohort of hypertensive, African-American siblings.

Introduction

Cardiac disease, including myocardial infarction, hypertensive and ischemic heart disease, and heart failure, is the leading cause of mortality and morbidity, accounting for 28% of the 2.4 million deaths in the U.S. each year.[1] Understanding predictors of cardiac events is particularly important for the African-American population. In 2003, the age-adjusted cardiac death rate for African-Americans was higher compared to the entire population, 364 vs. 232 per 100,000 persons per year.[1] Furthermore, hypertension occurs in 40-60% of African-Americans[2, 3] and prevalence of cardiac target organ damage may be nearly twice that in the non-Hispanic white population.[4]

Hypertension is a primary risk factor for cardiac disease.[5] However, because persons with similar blood pressure levels have varied cardiac outcomes, hypertension is not an exclusive predictor of cardiac risk or of damage to cardiac muscle.[6, 7] Echocardiographic measures of cardiac target organ damage – increased left ventricular mass (LVM) and high relative wall thickness (RWT) – have been well documented as powerful, independent predictors of cardiac morbidity and mortality.[6, 7]

Echocardiography is a validated, non-invasive method for measuring cardiac function and structure to identify preclinical cardiac damage.[8-10]

Echocardiographically measured LVM and RWT are complex quantitative traits with both genetic and environmental components. Correlates of LVM and RWT are well documented in multiethnic studies and include older age, male sex, diabetes and increased body mass index (BMI), systolic blood pressure (SBP), stroke volume, and dietary sodium intake.[4, 11-13] However, up to 75% of inter-individual variation in LVM and RWT remains unexplained by these established risk factors,[14] prompting studies of genetic influences on echocardiographic measures of preclinical cardiac disease. The Framingham Heart Study estimated the heritability of LVM at 0.24 to 0.32[15] and a study involving 110 twin pairs, obtained an LVM heritability estimate of 0.69.[16] Identification of genetic loci involved in determining LVM and RWT, and therefore predisposing to cardiac disease, are highly warranted.

Linkage studies conducted in rats have identified several chromosomal regions associated with increased LVM.[17, 18] An alternative strategy to linkage analysis are gene-disease association studies, which provide greater statistical power for identifying genetic variation underlying diseases of multifactorial etiology.[19] Although replication of results in independent data sets is critical for validation, most published gene-disease associations have not been replicated.[20, 21]To examine the effects of single nucleotide polymorphisms (SNPs) on echocardiographic measures of cardiac target organ damage, we examined associations of SNPs with LVM and RWT in hypertensive African-American adults, using two samples of unrelated individuals in order to replicate results.

Methods

Study Population

The National Heart Lung and Blood Institute established the Family Blood Pressure Program (FBPP) in 1996, joining established research networks investigating hypertension and cardiac diseases. One of the four networks in FBPP is the Genetic Epidemiology Network of Arteriopathy (GENOA), which recruited hypertensive African-Americans, Hispanic, and non-Hispanic white sibships for linkage and family-based association studies to investigate genetic contributions to hypertension and hypertensive target organ damage. Subject recruitment for GENOA was population-based in three geographic locations: Jackson, Mississippi (MS), Starr County, Texas, and Rochester, Minnesota. Subjects for this particular GENOA sub-study were African-Americans from Jackson, MS. GENOA recruited sibships containing at least two individuals with clinically diagnosed essential hypertension before age 60. Participants were diagnosed with hypertension if they had a previous clinical diagnosis of hypertension by a physician with current anti-hypertensive treatment, or an average SBP≥140 or diastolic blood pressure (DBP)≥90 on the second and third clinic visit. Exclusion criteria included secondary hypertension, alcoholism or drug abuse, pregnancy, insulin-dependent diabetes mellitus, or active malignancy.

Data for GENOA was collected over two phases. Phase I began in 1996, collecting blood pressure readings, information regarding family history, hypertension risk factors, and blood samples for genotyping and laboratory tests. Study visits were conducted in the morning after an overnight fast of at least eight hours. Blood pressure was measured with random zero sphygmomanometers and cuffs appropriate for arm size. Three readings were taken in the right arm after the participant rested in the sitting position for at least five minutes; the last two readings were averaged for the analyses. In Jackson, MS this was done for 1,854 people coming from 683 sibships. Phase II began in 2001 with the goal of measuring target organ damage, specifically via echocardiography. Approximately 80% of Phase I participants were successfully re-recruited for Phase II. Blood pressure, hypertension risk factor information, and blood samples were reassessed. Informed consent was obtained from all subjects and approval was granted by participating institutional review boards.

In order to detect SNP effects that replicated in another sample, this study took advantage of the affected sib pair design, originally used for linkage analysis, and created two samples, each with 448 unrelated individuals. This was done by randomly sampling one sib from each hypertensive sibship without replacement to create the first sample. From the remaining people, we randomly sampled a second sib from each sibship to establish the second sample. Singletons were equally divided between the two samples. These two large samples enabled us to test for replication of SNP associations in samples with similar genetic and environmental backgrounds.

Echocardiography

The left ventricular phenotypes of interest, LVM and RWT, were derived using phased-array echocardiographs with M-mode, two-dimensional and pulsed, continuous wave, and colorflow Doppler capabilities. Standardized methods, along with training and certification, were used by field-center technicians to achieve high-quality recordings. Readings were performed at the New York Presbyterian Hospital-Weill Cornell Medical Center and verified by a single highly experienced investigator. To measure LVM and RWT, the parasternal acoustic window was used to record at least 10 consecutive beats of two-dimensional and M-mode recordings of the left ventricular internal diameter (LVID) and wall thicknesses at, or just below, the tips of the anterior mitral leaflet in long- and short-axis views. Correct orientation of planes for imaging and Doppler recordings was verified using standardized protocols. Measurements were made using a computerized review station equipped with digitizing tablet and monitor screen overlay for calibration and performance of each measurement. LVID and interventricular septal and posterior wall thicknesses (PWT) were measured at end-diastole and end-systole according to the recommendations of the American Society of Echocardiography in up to three cardiac cycles.[22] Calculations of LVM were made using a necropsy-validated formula,[9] RWT was calculated as 2*(PWT)/LVID. Left ventricular phenotypes have excellent reliability when measured through echocardiography; e.g., the correlation between repeated measures of LVM was 0.93 between paired echocardiograms in hypertensive adults.[8]

Genotyping

Genotyping of 395 SNPs in 80 candidate genes for subjects from Jackson, MS was conducted at the GENOA central genotyping center at the University of Texas-Houston (see http://hyper.ahajournals.org for list of SNPs). Genes were selected to represent biological pathways or positional candidate genes from systems known to be associated with hypertension and heart disease including ion transport, inflammation, vascular wall biology, renin-angiotensin system, and lipid metabolism. SNPs provide a measure of genetic variation occurring between individuals of a population because they measure single nucleotide base changes in a gene. SNPs were selected with a minor allele frequency >0.05 if non-synonymous and >0.1 if marking 5’, 3’, or intronic regions of the gene. On average five SNPs per gene (range 1-18) were selected to represent variation in the region. This was done using the public NCBI database (http://www.ncbi.nlm.nih.gov) and the private Celera database (http://www.celeradiscoverysystem.com). SNP genotyping was obtained using a combination of two genotyping platforms: mass spectrometer-based detection system implemented on a Sequenom MassARRAY system, and the fluorogenic TaqMan assay implemented on an ABI Prism 7900 Sequence Detection System. Primer and probe sequences are available upon request.

Statistical Methods

High throughput data analyses were conducted using the statistical software R. Descriptive statistics for covariates, outcome variables, and SNPs were generated. Continuous variables are presented as mean±standard deviation. Student's t-test was used to confirm that the covariates and outcome variables in the two samples were not significantly different. Left ventricular mass index (g/m$^{2.7}$) (LVMI) and RWT were transformed using the natural logarithm. Linear regression was used to identify covariates associated with LVMI and RWT. Multivariable models for both LVMI and RWT were then constructed based on the univariate modeling results and known risk factors for the outcome. Genetic descriptive statistics were calculated, including either a chi-square test or the Fisher exact test for Hardy-Weinberg equilibrium (HWE).

We applied a multi-stage analysis strategy to identify and validate SNP associations replicating in two samples and to reduce the implications of false positives on our inferences. First, linear regression was used in both samples to identify SNPs that were significantly associated (α=0.10) with the respective outcome after adjustment for covariates. Only SNPs statistically significant in both samples were eligible for further analysis. Second, the genotype-specific means of LVMI and RWT associated with each SNP were determined in both samples. Two-sample t-tests were applied to all pairwise comparisons of genotype-specific means within the SNP to determine if the mean outcome differed between genotypes.[23, 24] SNPs with the same pairwise results in both samples were further analyzed. Third, analysis of covariance (ANCOVA) was utilized in the pooled sample to determine if the genotype-specific mean of LVMI or RWT within each SNP was parallel and coincident across samples, thus ensuring homogeneity of effect across samples.[24]

Results

General descriptive statistics of the covariate and outcome variables for the two samples are shown in Table 3.1. No significant differences were found between the two samples for any of the covariates or outcomes. Using the partition value of 51 g/m$^{2.7}$,[25] 15% of participants had LVH in sample 1 and 16% in sample 2. Using the threshold of 0.43 for RWT to detect concentric LV geometry,[26, 27] 3% of participants had concentric LV geometric patterns in sample 1 and 4% in sample 2.

Tables 3.2a-b outline the covariate associations in the univariate and final multivariable model of each outcome in detail. LVMI was adjusted for age, BMI, gender, SBP, diabetes and stroke volume. RWT was adjusted for age, BMI, SBP, and stroke volume. The residuals from the adjusted models provided the dependent variable for the SNP associations.

SNP Associations

In sample 1, 34 SNPs in 23 genes were significantly associated (α=0.10) with LVMI and 19 SNPs in 14 genes were associated with RWT. In sample 2, 33 SNPs in 20 genes were significantly associated with LVMI and 34 SNPs in 23 genes were associated with RWT. Of these, 3 SNPs were significantly associated with LVMI in both samples: APOE_rs449647 (p-value=0.014 and 0.055 respectively), SCN7A_cv356952 (p-value=0.089 and 0.024 respectively), and SLC20A1_cv9546580 (p-value=0.095 and

0.001 respectively) (Table 3.3). One SNP was significantly associated with RWT in both samples: ADRB1_Arg389Gly (p-value=0.019 and 0.065 respectively) (Table 3.3).

Genotype-Phenotype Relationships for LVMI

The three SNPs replicating associations with LVMI exhibited significant and consistent genotype-phenotype relationships in both samples (Table 3.4). There was a significant difference in mean LVMI between the APOE_rs449647 AA genotype, and the combined AT/TT genotypes, in samples 1 and 2 (p-value=0.010 and 0.004, respectively). The AT and TT genotypes were combined because t-tests revealed no significant differences between mean LVMI at these genotypes in either sample (p-value=0.644 and 0.876, respectively). The major allele, T, of APOE_rs449647 was significantly associated with higher LVMI by a mean of 4.1g/m$^{2.7}$ in sample 1 and 3.4 g/m$^{2.7}$ in sample 2.

A significant difference in mean LVMI was also found between the AA homozygote in SCN7A_cv356952 and the combined AT/TT genotypes in samples 1 and 2 (p-value=0.041 and 0.025, respectively). The AT and TT genotypes were combined because mean LVMI between those genotypes was not significantly different in either sample (p-value=0.427 and 0.055). The major allele, T, of SCN7A_cv356952 was significantly associated with higher LVMI by a mean of 1.9 g/m$^{2.7}$ in sample 1 and 2.3 g/m$^{2.7}$ in sample 2.

The AA homozygote in SLC20A1_ cv9546580 had a significantly lower mean LVMI compared to the heterozygote, AG in samples 1 and 2 (p-value=0.037 and 0.005, respectively). The difference in mean LVMI between genotypes AG and GG was only significant in sample 2 (p-value=0.162 and 0.0003). The AG genotype of SLC20A1_ cv9546580 was associated with the highest values of LVMI in both samples, exhibiting a pattern of overdominance.

In addition, all three SNPs were separately tested using ANCOVA for genotype-sample interactions. There was no statistical evidence (α=0.05) that sample influenced the genotype-specific means of LVMI (data not shown). Therefore, the consistency of directionality in the genotype-phenotype relationships for LVMI was confirmed across samples (shown without log transformations in Figures 3.1-3.3).

The three SNPs found to be significantly associated with LVMI were added to the multivariable model and the R^2 significantly increased in both samples. In sample 1, the R^2 increased from 0.316 to 0.381 (p-value=0.003). In sample 2, the R^2 increased from 0.316 to 0.368 (p-value<0.001). Jointly, these SNPs explain 5-7% additional variability in LVMI beyond traditional risk factors.

Genotype-Phenotype Relationship for RWT

The Arg389Gly SNP in *ADRB1* was significantly associated with RWT in both samples and exhibited significant, consistent genotype-phenotype relationships within and across samples (Table 3.4). The CC homozygote in ADRB1_Arg389Gly had a significantly lower mean RWT compared to the combined CG and GG genotypes in samples 1 and 2 (p-value=0.028 and 0.019, respectively). The CG and GG genotypes were combined because the mean RWT between those genotypes was not significantly different in either sample (p-value=0.079 and 0.869). The major allele, G, of ADRB1_Arg389Gly was significantly associated with higher RWT by a mean of 0.01 in sample 1 and 0.009 in sample 2.

ADRB1_Arg389Gly genotypes were also tested using ANCOVA for genotype-sample interactions. There was no statistical evidence (α=0.05) that sample influenced the genotype-specific means of RWT (data not shown). Therefore, the consistency of directionality in the genotype-phenotype relationships for RWT was confirmed across samples (Figure 3.4).

The *ADRB1* SNP found to be significantly associated with RWT was added to the multivariable model and the R^2 increased in both samples. In sample 1, the R^2 increased from 0.124 to 0.142 (p-value=0.019). In sample 2, the R^2 increased from 0.122 to 0.132 (p-value=0.067). This indicates that ADRB1_Arg389Gly explains 1-2% additional variability in RWT beyond traditional risk factors.

Genetic Descriptives for Significant SNPs

All 395 SNPs were tested for HWE using either the chi-square test or the Fisher exact test. Of the four SNPs identified as significantly associated with LVMI or RWT, ADRB1_Arg389Gly violated HWE in both samples (p-value=0.015 & 0.038 respectively) and SLC20A1_cv9546580 violated HWE in sample 1 (p-value=0.007). This could be due to various reasons including population substructure,[28] type I error, or because selection is working against one or more genotypes, thereby indicating a true association.[29] Furthermore, we would not expect this hypertensive sample to follow HWE distributions if the SNP is associated with hypertension.[30] None of the four SNPs were in linkage disequilibrium and more information specific to individual SNPs can be found on the supplemental online table (see http://hyper.ahajournals.org).

Replication of Results

In order to assess whether the effects of these genetic polymorphisms replicated beyond the African-American Jackson, MS cohort, we utilized the already existing Hispanic cohort of GENOA from Starr County, Texas (n=1,228). Of the four SNPs reported above, two also showed significant associations with LVMI in the Hispanic cohort; SCN7A_cv356952 (AA vs. AT&TT: p-value=0.029) and SLC20A1_cv9546580 (AA/AG vs. GG: p-value=0.074). We noted that the allele frequency distribution was different (α=0.05, χ^2df=1) between the African-American cohort from Jackson, MS and the Hispanic cohort from Starr County, TX in three of the four SNPs, SCL20A1_cv9546580 being the only SNP with similar allele frequencies. The minor allele frequencies for APOE_rs449647, SCN7A_cv356952, and ADRB1_Arg389Gly in African-Americans are 0.29, 0.46, and 0.38 (respectively) compared to 0.23, 0.24, and 0.21 in Hispanics. These differences in allele frequencies indicate that we had different power to detect the same SNP effect in these two ethnic groups. Allele frequencies affect the power of single gene association tests since they are related the number of individuals in a particular genotype class.

Discussion

In this study of hypertensive, African-American siblings, we identified three SNPs (APOE_rs449647, SCN7A_cv356952, and SLC20A1_cv9546580) that replicated association with LVMI in two African-American samples and showed consistent phenotypic effects. Additionally, we identified one SNP (ADRB1_Arg389Gly) that replicated association with RWT in two African-American samples and also demonstrated consistent phenotypic effects. Furthermore, two of the SNPs showed significant associations with LVMI in an independent population-based sample of Hispanic participants in GENOA (SCN7A_cv356952 and SLC20A1_cv9546580).

The process by which we identified these results carefully addressed two main issues of conducting gene-disease association studies: 1) replication of results and 2) the issues of multiple hypothesis testing. Both of these issues will require increasing attention in the near future as the field of genetic epidemiology transitions towards genome-wide association (GWA) studies. GWA studies will increase the number of non-replicable results since the number of potentially positive results will increase exponentially. A meta-analysis of over 600 reported gene-disease associations found few replications.[31] Of the 600 associations, 166 had been studied three or more times, and only 6 were consistently replicated.[31] While this meta-analysis included studies of many outcomes, the same issue applies to gene association studies for left ventricular traits. It is possible that the lack of replicated results in the published literature is due to differences among study populations in distributions of genetic and environmental factors[32] as well as in sampling techniques and analysis strategies. Therefore, creating two samples from a single population provides an opportunity to compare results across samples with similar genetic backgrounds and environmental exposures.

The second issue of association studies we addressed is that of multiple hypothesis testing. A variety of methods are available to correct for multiple hypothesis testing: the conservative Bonferroni adjustment of a single p-value, controlling for false discovery rates, cross-validation techniques, or replicating results in an independent sample. Most of these methods operate on the principle of lowering the alpha level necessary for rejecting the null hypotheses. Alpha levels and the power to detect associations inherently take into account factors such as allele frequencies and the size of each allele effect. It is highly likely that alleles conferring an effect in complex, common diseases will have small effect sizes. Therefore overly conservative adjustments of the alpha level may miss a large portion of true associations. GWA studies magnify the problems of correcting for multiple hypothesis testing since thousands, instead of hundreds, of hypothesis tests are conducted – making conservative correction techniques even more certain to exclude many true associations. To address the issue of multiple testing and to reduce the number of false positives, we implemented a multi-stage analysis strategy to identify and validate SNP associations replicating in both samples. Ultimately, all four SNPs reported in this paper: 1) were significantly associated with LVMI or RWT in both samples, 2) exhibited consistent, significant differences between the same genotype-specific mean outcome in both samples and 3) showed consistent directionality and magnitude of genotype-phenotype relationships across samples.

Hypertension, cardiac target organ damage, and overt cardiac disease all occur at increased rates in the African-American population. Of the four SNPs identified through the multi-stage testing, ADRB1_Arg389Gly associated with RWT is of particular interest for the African-American population. *ADRB1* is the gene for the β_1-adrenergic receptor, which is the receptor for norepinephrine on the cardiomyocyte. Arg389Gly SNP is a non-synonymous mutation resulting in a missense substitution within amino acid 389. This amino acid change modifies protein function by increasing contractile response at the myocyte.[33] The G allele has been found to be more frequent in the African-American population and may contribute to increased hypertension and heart failure rates in African-Americans.[34] We identified the G allele of the Arg389Gly SNP to be significantly associated with increased RWT, which is a predictor of cardiac outcomes such as heart failure. Our current results support the hypothesis of the Arg389Gly mutation playing a role in increased heart disease in African-Americans.

The direct functional implications of the 3 SNPs found to be associated with LVMI are not well studied in terms of their impact on LVM, but the genes are etiologic candidates for greater LVM and heart disease. ApoE is a lipoprotein implicated in dyslipidemia and associated with a variety of heart disease outcomes. SCN7A is a voltage gated sodium ion channel and SLC20A1 is the gene for a solute carrier protein, both of which may play a role in LVMI through cell signaling and regulation of hemodynamic load.

LVMI and RWT are complex quantitative traits that are influenced by both genetics and the environment. This study identified SNPs with replicated main effects for LVMI and RWT while adjusting for other risk factors. However, context dependency of the main effects within the genome and environment are critical in order to fully understand these complex traits. Therefore, future directions for our research are to further explore these SNP main effects along with potential interactions with other known risk factors and/or with other SNPs. In addition, we feel it is worthwhile to further explore the *ADRB1* SNP given that it is a nonsynonymous SNP, which has been previously implicated in heart disease in African-Americans.

Perspectives

The field of genetic epidemiology is growing rapidly with the increasing availability of genomic data. This paper highlights statistical and methodological issues that are emerging from this exponential growth and presents a novel approach for replicating gene-disease association results. This study provides insight into genetic contributions of quantitative measures of hypertensive cardiac organ damage. Consideration of context dependency and interactive genetic effects is an obvious progression for this research.

Table 3.1. Descriptive statistics of covariates and outcome variables.

	Sample 1 (N=439)		Sample 2 (N=439)	
Variable	**Mean ± SD**	**Range**	**Mean ± SD**	**Range**
Age, years	64.4 ± 9.0	28.4 - 86.0	64.7 ± 8.5	35.4 - 95.7
BMI, kg/m^2	32.15 ± 6.78	16.4 - 58.6	31.6 ± 6.3	18.9 - 55.1
Systolic BP, mmHg	141.8 ± 20.7	89 - 220	143.9 ± 22.6	79 - 243
Diastolic BP, mmHg	80.1 ± 11.5	46 - 121	80.4 ± 11.7	45 - 121
Pulse BP, mmHg	61.7 ± 17.4	30 - 141	63.5 ± 19.1	19 - 132
Stroke Volume	78.1 ± 13.7	40.8 - 136.6	77.8 ± 16.4	29.8 - 190.6
Male, N(%)	115(26%)		146(33%)	
Smoker, N(%)	61(14%)		57(13%)	
Diabetes, N(%)	161(36%)		140(31%)	
LV Mass Index, g/m$^{2.7}$	40.6 ± 10.3	16.2 - 90.0	41.5 ± 11.9	21.4 - 114.8
Log LVMI	3.67 ± 0.25	2.8 - 4.5	3.69 ± 0.26	3.1 - 4.7
Relative Wall Thickness	0.32 ± 0.05	0.20 - 0.52	0.33 ± 0.05	0.19 - 0.57
Log RWT	-1.14 ± 0.15	-1.61 - (-0.65)	-1.13 ± 0.15	-1.66 - (-0.56)

Table 3.2.a. Univariate linear modeling of covariates with outcome variables LVMI and RWT.

	Sample 1		Sample 2	
LVMI	**β**	**P-value**	**β**	**P-value**
Age, years	0.0044	<0.001	0.0056	0.001
BMI, kg/m^2	0.0111	<0.001	0.0144	<0.001
Male Gender	0.0008	0.977	-0.0255	0.373
SBP, mmHg	0.0026	<0.001	0.0030	<0.001
DBP, mmHg	0.0004	0.705	0.0008	0.450
Pressure Pulse	0.0035	<0.001	0.0039	<0.001
Diabetes	0.1032	<0.001	0.0637	0.018
Smoking	-0.0357	0.293	-0.0174	0.640
Stroke Volume	0.0078	<0.001	0.0065	<0.001

	Sample 1		Sample 2	
RWT	**β**	**P-value**	**β**	**P-value**
Age, years	0.0043	<0.001	0.0052	<0.001
BMI, kg/m^2	-0.0014	0.202	-0.0001	0.901
Male Gender	0.0277	0.084	-0.0165	0.324
SBP, mmHg	0.0012	<0.001	0.0011	<0.001
DBP, mmHg	0.0007	0.269	0.0001	0.829
Pressure Pulse	0.0014	<0.001	0.0015	<0.001
Diabetes	0.0274	0.062	0.0203	0.201
Smoking	-0.0105	0.608	0.0052	0.813
Stroke Volume	-0.0017	0.001	-0.0014	0.002

Table 3.2b. Multivariable linear modeling of covariates with outcome variables.

	Sample 1		Sample 2	
LVMI[*]	**β**	**P-value**	**β**	**P-value**
(Intercept)	2.28	<0.001	2.20	<0.001
Age, years	0.0032	0.0081	0.0061	<0.001
BMI, kg/m^2	0.0096	<0.001	0.0126	<0.001
Male Gender	-0.0051	0.836	-0.0152	0.558
SBP, mmHg	0.0025	<0.001	0.0022	<0.001
Diabetes	0.0852	<0.001	0.0173	0.455
Stroke Volume	0.0065	<0.001	0.0049	<0.001
	R^2=0.316		R^2=0.316	

	Sample 1		Sample 2	
RWT[*]	**β**	**P-value**	**β**	**P-value**
(Intercept)	-1.45	<0.001	-1.52	<0.001
Age, years	0.0045	<0.001	0.0050	<0.001
BMI, kg/m^2	0.0018	0.111	0.0026	0.030
SBP, mmHg	0.0010	0.003	0.0010	0.002
Stroke Volume	-0.0023	<0.001	-0.0020	<0.001
	R^2=0.124		R^2=0.122	

* LVMI and RWT were log transformed for analysis

Table 3.3. SNPs significantly associated with LVMI or RWT in both samples after adjustment for covariates.

LVMI[*]		**Sample 1**		**Sample 2**	
Gene	**SNP**	**F(df1,df2)**	**P-value**	**F(df1,df2)**	**P-value**
APOE	rs449647	4.37(2,413)	0.014	2.84(2,409)	0.055
SCN7A	cv356952	2.38(2,413)	0.089	3.78(2,406)	0.024
SLC20A1	cv9546580	2.57(2,387)	0.095	7.07(2,394)	0.001

RWT[†]		**Sample 1**		**Sample 2**	
Gene	**SNP**	**F(df1,df2)**	**P-value**	**F(df1,df2)**	**P-value**
ADRB1	Arg389Gly	3.83(2,414)	0.019	2.91(2,406)	0.065

[*] LVMI was log transformed and adjusted for age, sex, BMI, SBP, diabetes, and stroke volume

[†] RWT was log transformed and adjusted for age, BMI, SBP, and stroke volume

Table 3.4. Mean values of logLVMI or logRWT based on genotypes in samples 1 and 2.

SNP	Genotype (N1/N2)	Sample 1 Mean[*] log LVMI ± SD	P-value[†]	Sample 2 Mean[*] log LVMI ± SD	P-value[†]
APOE_rs449647	AA (32/37)	3.57 ± 0.22		3.61 ± 0.17	
	AT/TT (384/375)	3.68 ± 0.20	0.010	3.70 ± 0.22	0.004
SCN7A_cv356952	AA (92/77)	3.63 ± 0.20		3.64 ± 0.18	
	AT/TT (324/332)	3.68 ± 0.20	0.041	3.70 ± 0.22	0.025
SLC20A1_cv9546580	AA (218/231)	3.66 ± 0.20		3.67 ± 0.22	
	AG (131/135)	3.71 ± 0.20	0.037	3.74 ± 0.22	0.005
	GG (41/31)	3.66 ± 0.16	0.871	3.60 ± 0.17	0.041
SNP	**Genotype (N1/N2)**	**Sample 1 Mean[‡] log RWT ± SD**	**P-value[†]**	**Sample 2 Mean[‡] log RWT ± SD**	**P-value[†]**
ARB1A_Arg389Gly	CC (170/161)	-1.16 ± 0.15		-1.15 ± 0.14	
	GC/GG (249/253)	-1.13 ± 0.16	0.028	-1.12 ± 0.15	0.019

[*] mean log LVMI is adjusted for age, BMI, gender, SBP, diabetes, and stroke volume

[†] P-values are from two-sample t-tests comparing the respective genotype to the AA genotype for the LVMI SNPs and to the CC genotype for the RWT SNP.

[‡] mean RWT is adjusted for age, BMI, SBP, and stroke volume

Figure 3.1: Genotype specific means of LVMI in Samples 1 and 2 for the APOE_rs449467 SNP. Standard errors for each genotype-specific mean are represented by the vertical bars at each point.

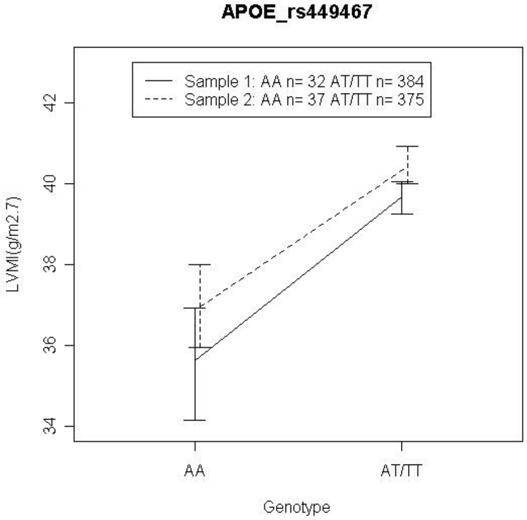


Figure 3.2: Genotype specific means, with standard errors, of LVMI in Samples 1 and 2 for the SCN7A_356952 SNP.

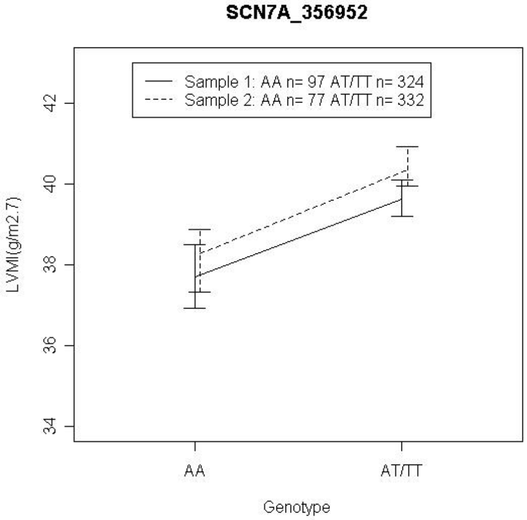


Figure 3.3: Genotype specific means, with standard errors, of LVMI in Samples 1 and 2 for the SLC20A1_cs9546580 SNP.

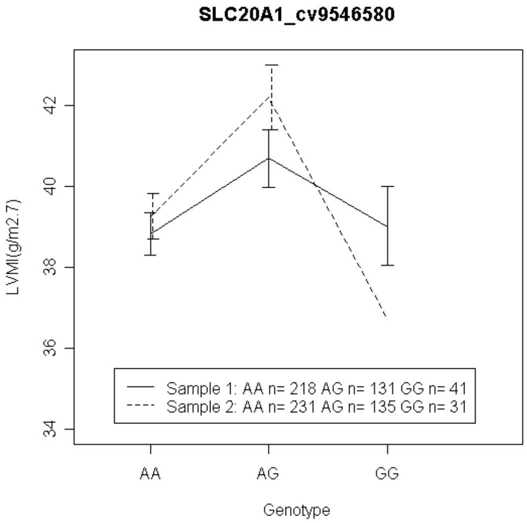
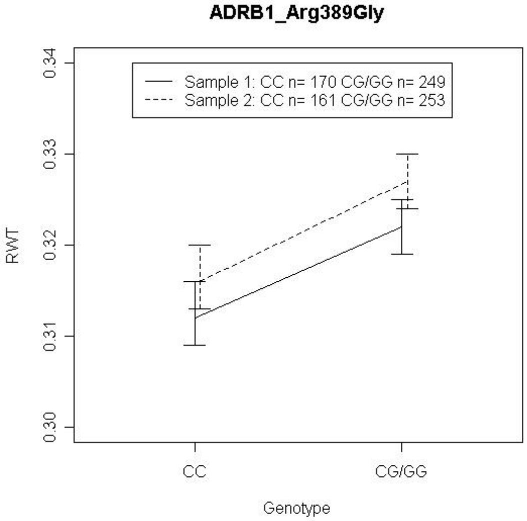


Figure 3.4: Genotype specific means, with standard errors, of RWT in Samples 1 and 2 for the ADRB1_Arg389Gly SNP.



References

1. National Center for Health Statistics. CDC Health, United States, 2005. Hyattsville, MD: US Government printing office, 2005.

2. Hertz RP, Unger AN, Cornell JA, Saunders E. Racial disparities in hypertension prevalence, awareness, and management. Arch Intern Med. 2005; 165(18):2098-2104.

3. Kramer H, Han C, Post W, et al. Racial/ethnic differences in hypertension and hypertension treatment and control in the Multi-Ethnic Study of Atherosclerosis (MESA). Am J Hypertens. 2004; 17(10):963-970.

4. Kizer JR, Arnett DK, Bella JN, et al. Differences in left ventricular structure between black and white hypertensive adults: The Hypertension Genetic Epidemiology Network study. Hypertension. 2004; 43(6):1182-1188.

5. [Anonymous]. The impact of cardiovascular risk factors on the age-related excess risk of coronary heart disease. Int J Epidemiol. 2006; .

6. Koren MJ, Devereux RB, Casale PN, Savage DD, Laragh JH. Relation of left ventricular mass and geometry to morbidity and mortality in uncomplicated essential hypertension. Ann Intern Med. 1991; 114(5):345-352.

7. Levy D, Garrison RJ, Savage DD, Kannel WB, Castelli WP. Prognostic implications of echocardiographically determined left ventricular mass in the Framingham Heart Study. N Engl J Med. 1990; 322(22):1561-1566.

8. Palmieri V, Dahlof B, DeQuattro V, et al. Reliability of echocardiographic assessment of left ventricular structure and function: The PRESERVE study. Prospective randomized study evaluating regression of ventricular enlargement. J Am Coll Cardiol. 1999; 34(5):1625-1632.

9. Devereux RB, Alonso DR, Lutas EM, et al. Echocardiographic assessment of left ventricular hypertrophy: Comparison to necropsy findings. Am J Cardiol. 1986; 57(6):450-458.

10. Arnett DK, de las Fuentes L, Broeckel U. Genes for left ventricular hypertrophy. Curr Hypertens Rep. 2004; 6(1):36-41.

11. Bella JN, Wachtell K, Palmieri V, et al. Relation of left ventricular geometry and function to systemic hemodynamics in hypertension: The LIFE study. Losartan intervention for endpoint reduction in hypertension study. J Hypertens. 2001; 19(1):127-134.

12. du Cailar G, Ribstein J, Mimran A. Dietary sodium and target organ damage in essential hypertension. Am J Hypertens. 2002; 15(3):222-229.

13. Galderisi M, Anderson KM, Wilson PW, Levy D. Echocardiographic evidence for the existence of a distinct diabetic cardiomyopathy (the Framingham Heart Study). Am J Cardiol. 1991; 68(1):85-89.

14. Benjamin EJ, Levy D. Why is left ventricular hypertrophy so predictive of morbidity and mortality? Am J Med Sci. 1999; 317(3):168-175.

15. Post WS, Larson MG, Myers RH, Galderisi M, Levy D. Heritability of left ventricular mass: The Framingham Heart Study. Hypertension. 1997; 30(5):1025-1028.

16. Swan L, Birnie DH, Padmanabhan S, Inglis G, Connell JM, Hillis WS. The genetic determination of left ventricular mass in healthy adults. Eur Heart J. 2003; 24(6):577-582.

17. Di Nicolantonio R, Kostka V, Kwitek A, Jacob H, Thomas WG, Harrap SB. Fine mapping of Lvm1: A quantitative trait locus controlling heart size independently of blood pressure. Pulm Pharmacol Ther. 2006; 19(1):70-73.

18. Innes BA, McLaughlin MG, Kapuscinski MK, Jacob HJ, Harrap SB. Independent genetic susceptibility to cardiac hypertrophy in inherited hypertension. Hypertension. 1998; 31(3):741-746.

19. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science. 1996; 273(5281):1516-1517.

20. Hattersley AT, McCarthy MI. What makes a good genetic association study? Lancet. 2005; 366(9493):1315-1323.

21. Moore JH. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Hum Hered. 2003; 56(1-3):73-82.

22. Lang RM, Bierig M, Devereux RB, et al. Recommendations for chamber quantification: A report from the American society of Echocardiography's Guidelines and Standards committee and the Chamber Quantification writing group, developed in conjunction with the European Association of Echocardiography, a branch of the European Society of Cardiology. J Am Soc Echocardiogr. 2005; 18(12):1440-1463.

23. Falconer DS, Mackay TFC. Introduction to Quantitative Genetics. Fourth Edition ed. England: Pearson Prentice Hall, 1996.

24. Sokal RR, Rohlf FJ. Biometry: The Principles and Practice of Statistics in Biological Research. 3rd ed. New York, NY: WH Freeman and Comany, 1994.

25. Nunez E, Arnett DK, Benjamin EJ, et al. Optimal threshold value for left ventricular hypertrophy in blacks: The Atherosclerosis Risk in Communities Study. Hypertension. 2005; 45(1):58-63.

26. Wachtell K, Bella JN, Liebson PR, et al. Impact of different partition values on prevalences of left ventricular hypertrophy and concentric geometry in a large hypertensive population : The LIFE study. Hypertension. 2000; 35(1 Pt 1):6-12.

27. Arnett DK. Genetic contributions to left ventricular hypertrophy. Curr Hypertens Rep. 2000; 2(1):50-55.

28. Deng HW. Population admixture may appear to mask, change or reverse genetic effects of genes underlying complex traits. Genetics. 2001; 159(3):1319-1323.

29. Lee WC. Searching for disease-susceptibility loci by testing for Hardy-Weinberg disequilibrium in a gene bank of affected individuals. Am J Epidemiol. 2003; 158(5):397-400.

30. Nielsen DM, Ehm MG, Weir BS. Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus. Am J Hum Genet. 1998; 63(5):1531-1540.

31. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. Genet Med. 2002; 4(2):45-61.

32. Hoffjan S, Nicolae D, Ober C. Association studies for asthma and atopic diseases: A comprehensive review of the literature. Respir Res. 2003; 4:14.

33. Small KM, Wagoner LE, Levin AM, Kardia SL, Liggett SB. Synergistic polymorphisms of beta1- and alpha2C-adrenergic receptors and the risk of congestive heart failure. N Engl J Med. 2002; 347(15):1135-1142.

34. Moore JD, Mason DA, Green SA, Hsu J, Liggett SB. Racial differences in the frequencies of cardiac beta(1)-adrenergic receptor polymorphisms: Analysis of c145A>G and c1165G>C. Hum Mutat. 1999; 14(3):271.

Chapter 4

Exploring Population Substructure in the African-American Cohort of GENOA

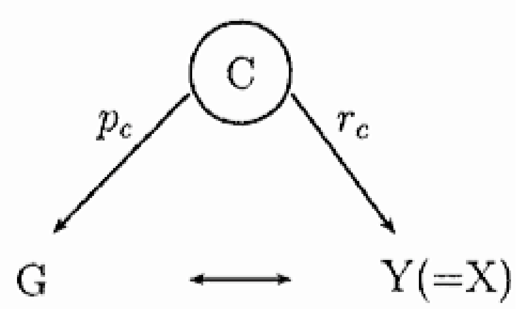
Abstract

Population substructure is a source of confounding in genetic association studies, particularly when studies are conducted in admixed populations such as African-Americans. This chapter describes the general topic of population substructure and reviews three methods commonly used to adjust for population substructure in population-based association studies: genomic control, structured association, and principal component analysis. Based on 385 ancestry informative SNPs in the program *Structure*, we identified substructure in the form of admixture within the African-American cohort of GENOA. The mean percent Caucasian ancestry was 16.5% (median: 14.8%, range: 0.16 – 60.3%). After the literature review presented here, principal component analysis was the method chosen as most appropriate for adjusting population substructure in the African-American cohort of GENOA. The methods used to implement principal component analysis in GENOA, and the results, are discussed.

Introduction

The main goal of genetic association studies is to identify genetic variants associated with an outcome of interest (either quantitative or qualitative). The underlying assumption when estimating the association between a genetic variant and outcome of interest is the absence of confounding. Unfortunately there is a pervasive confounding issue in genetic association studies referred to as population substructure. [1, 2]

Population substructure is defined by one of two possible scenarios: either the study population is composed of two genetically distinct subpopulations or the study population contains a single population of individuals admixed from two or more "parent" populations. This latter scenario of admixture is present in genetic association studies of African-American populations. The presence of population substructure will result in confounding of the gene-disease association if the following three conditions are satisfied: 1) the admixture proportions vary between individuals in the study population, 2) the risk of disease (or mean value of quantitative trait) varies with admixture proportions, and 3) the allele frequencies at the locus of interest vary with admixture.[3] The bias from population substructure can result in either false positive results or false negative results.[4] The diagram below is taken from a review article written by Devlin and pictorially demonstrates this confounding.[5]



Where C is the confounding variable (proportion of admixture in each individual), G is the genotype at a specific locus being tested for association with the outcome X. Confounding by C occurs when the presence of a specific genotype, G, is conditional on admixture proportion (p_c) and when the risk (Y) of X is conditional on admixture proportion (r_c).

While population substructure can be present in any study sample, including European samples, it is a particularly pertinent issue in highly admixed populations such as African-Americans. African-Americans are a relatively recently admixed group with allelic contributions from European and African ancestors. The sample being used in this research consists of African-Americans recruited from Jackson, Mississippi for the Genetic Epidemiology Network of Arteriopathy (GENOA) study. Based on anthropological evidence, the majority of African-Americans in the Mississippi delta area who came as a result of the trans-Atlantic slave trade are descendents of West African populations.[6] Therefore the African-Americans in our sample are likely admixed between West Africans and Northern Europeans. Given the large documented differences in allele frequencies between these ancestral populations,[7] we hypothesize that our study sample will have a distribution of genomic contributions from both West Africa and Northern Europe, satisfying the first and third criteria for population substructure. The second criterion above, risk of disease varying with admixture proportions, is also likely to be a concern for this research. African-Americans are known to experience higher rates of hypertension, left ventricular remodeling, and heart disease compared to non-Hispanic white populations.[8-11]

The goal of the following work is to explore the ancestral genomic distribution in our study sample and explore which methods are best suited for adjusting population substructure in the African-American cohort of GENOA.

Review of Population Substructure Adjustment Methods

There are various methods that can be used to adjust for population substructure. Conducting family-based association studies is considered to be the gold standard to control for population substructure.[3] However, population-based study designs have higher power to detect genetic associations and are more common study designs when investigating diseases with late-onset.[12] Given the surge in population-based genetic association studies, new population substructure adjustment methods have also been surging, with little agreement on a gold standard method for adjustment. I will discuss three of the most commonly used methods to control for confounding due to population substructure in population-based cohorts: genomic control, structured association testing (using *Structure*), and principal component analysis.

Genomic Control

Genomic control (GC) was one of the earliest methods developed to control for population substructure in association studies. GC uses the genome to provide control for relatedness within population-based study samples[13] achieving power similar to that of family-based methods.[14] At its inception, GC was aimed at adjusting for any type of cryptic relatedness, whether it be from population substructure, poorly chosen controls, or some other source of dependence within samples.[13] GC is grounded in the hypothesis that population substructure results in a relatively uniform inflation of test statistics across the genome, and therefore an increase in type I errors.[13]

Implementing GC in practice is very straightforward. GC utilizes null loci (ie. variants not proposed to have any effect on the outcome) from across the genome to calculate a genomic “inflation factor”, λ. This inflation factor quantifies the degree of inflation assumed to be due to population substructure and is calculated in a two-step process. First, the test for association with the outcome is conducted for all of the null GC loci. For example, in a case-control study a chi-square statistic, χ^2, would be calculated for each null locus. Second, λ is estimated by dividing the empirical median of the chi-square tests by the expected median test statistic assuming no association (median χ^2_{observed} / median χ^2_{expected}). If $\lambda=1$, then no confounding due to population substructure is present. If $\lambda<1$, λ is set to 1 and no inflation is assumed. If $\lambda>1$, λ represents the degree of inflation in all test statistics calculated for candidate loci.

Once estimated, λ is used to universally adjust all empirical test statistics for the candidate loci. Continuing with the chi-square example, each genetic variant being tested has a respective χ^2. Each one of those statistics is adjusted for population substructure by dividing by λ, χ^2/λ. The adjusted statistic has a chi-square distribution with one degree of freedom.[13] GC is also applicable for quantitative traits, the method is simply applied to the T statistic.[15]

GC is easily implemented on a genome-wide scale making it an appealing option when controlling for population substructure in genome-wide association studies. However, there are two glaring reasons why GC is not appropriate for this particular GENOA study. First of all, the assumption that over-dispersion of the test statistic due to population substructure is constant across the genome is unrealistic. In an admixed population many alleles show strong differentiation across parental populations and will by definition not show uniform admixture proportions. Ignoring this distribution of admixture and inappropriately treating the inflation as constant across the genome would lead to a loss in power.[16] In simulation studies, Zhang and colleagues showed that GC was the lowest performing option for population substructure adjustment based on multiple performance indices: power, type I error rates, accuracy and positive predictive value, further arguing against the use of GC.[17]

Structure

A cluster-based algorithm developed by Pritchard, *Structure*, can be used to test for and adjust for population substructure in association tests, this is referred to as "structured association testing".[18] In order to run *Structure*, genotype data must be available from uncorrelated loci for the study sample of interest and also for populations that are hypothesized to be the genomic ancestors. *Structure* then uses these multi-locus genotypes to detect the presence of *K* genetically distinct subpopulations, where *K* can be unknown. Individuals from the study sample of interest are then assigned to the *K* subpopulations with a given probability. If very distinct genomic subgroups are identified by *Structure*, the association analysis can be stratified based on those clusters; thereby conducting association tests within homogeneous subgroups.

Individuals from admixed populations such as African-American populations cannot be divided into distinct strata because of the continuum of admixture between the *K* subpopulations. In this situation, the probability of membership to *K* parent population becomes each individuals estimated genomic average of percent ancestry from the *K* parent population.[18] Therefore, another option for structured association testing is to use the percent ancestry variable as a quantitative adjustment variable in the association testing regression models. Although previous research has used this methodologically straightforward approach,[19, 20] there are a few drawbacks to using percent ancestry estimates as adjustment variables. First of all, estimation of the percent ancestry variable is done in a step completely separate from using it to adjust association tests in regression models. Therefore uncertainty in the estimate of ancestry is not easily incorporated into the model, which may underestimate the effect of confounding.[3] This uncertainty would not be a concern if the estimate of ancestry was very robust. However, estimates of ancestry in *Structure* are optimized only if the markers used are ancestry informative markers (AIMs).[17] The term AIMs aptly refers to a set of genome-wide markers (SNPs or microsatellites) that have highly differentiated allele frequencies between two populations. These markers can identify the ancestral origin of a chromosomal region and give accurate summaries of admixture proportions within a population. Panels of AIMs have been developed that are most appropriate for African-American samples. A recent panel of 2,000 SNPs was found to be optimal in determining ancestral contributions in an African-American study sample.[21] If the markers used in *Structure* are not highly informative, the estimate of percent ancestry would underestimate population substructure and therefore not sufficiently control for confounding. Another dissuading factor in the use of *Structure* is the exponential increase in computational time needed to run as the dimension of input data increases. Given the amount of time already needed to run many of the genome-wide analyses, methods that optimize time use are appealing, and *Structure* does not provide this.

Structure was originally developed in order to infer the presence of distinct subpopulations, and assign individuals to those populations.[18] This is at odds with the *a priori* hypothesis that an African-American population does not necessarily contain distinct genomic strata. *Structure* is commonly used to detect substructure because it is easy to implement and interpret and is well suited for exploratory analysis of substructure within our African-American study sample; however, it does not provide an optimal method to adjust for confounding due to admixture.

Principal Component Analysis

In an effort to address some of the concerns and limitations of GC and *Structure*, Price and colleagues present a method to detect and adjust for population substructure called EIGENSTRAT.[16] EIGENSTRAT is a statistically powerful approach and computationally very fast making it an appealing option for high dimensional datasets and genome-wide association studies. EIGENSTRAT uses principal component analysis (PCA) to infer continuous axes of genetic variation, as was first done in 1978 by Cavalli-Sforza and colleagues.[22] PCA is a data reduction procedure that can be used to reduce high-dimensional genetic marker data by creating orthogonal principal components which reveal underlying structure in the dataset with a smaller number of variables.[23] Each principal component is a linear combination of optimally weighted input variables (i.e. SNPs). Each SNP contributes to each component, only with different weights. The general formula for a principal component is as follows:

$$PC_i = b_{i1}(X_1) + b_{i2}(X_2) + \ldots + b_{ip}(X_p)$$

Where p is the number of observed input variables X, and b_{ip} is the weight assigned to each of the p input variables for the i^{th} principal component. There are two main characteristics of principal components.[23] The first principal component explains the largest amount of variation in the set of input variables with each successive component explaining less. The second characteristic is that each principal component is uncorrelated with the previous components, they are orthogonal. Therefore, all components are uncorrelated with each other. A chosen number of retained principal components can be used to adjust tests of genetic association for genomic variation in the dataset (i.e. population substructure).[16]

Implementing PCA with EIGENSTRAT is done using a software package called EIGENSOFT (http://genepath.med.harvard.edu/~reich/Software.htm). Features of EIGENSOFT include quantifying the degree of admixture with PCA, testing the significance of principal components, and adjusting association statistics for admixture. EINGENSTRAT is a three-step process. The first step is to apply PCA using some type of genetic marker (i.e. SNPs, microsatellites, haplotype frequencies). The second step involves adjustment of genotypes and phenotypes by amounts attributable to ancestry along each axis. Price and colleagues recommend a default number of 10 principal components be used to adjust for ancestry in association analyses.[16] Once the specific number of components to include has been decided, the third and final step is to compute the association test statistics for each candidate genetic locus within EIGENSTRAT using ancestry adjusted genotypes and phenotypes.[16]

One of the most appealing and unique features of EIGENSOFT, compared to implementing PCA outside of this software, is the ability to formally test the statistical significance of principal components. Patterson describes that the normalization process applied to the genotypes in EIGENSTRAT results in eigenvalues for the components that approximately follow a Tracy-Widom distribution.[24] EIGENSOFT computes the Tracy-Widom test statistic and p-value for each principal component. This is the first application with a formal test statistic for population substructure.[24] However, eigenvalues do not approximate the Tracy-Widom distribution in high admixed populations.[24] This does not mean principal components do not appropriately model ancestry in admixed populations, it only means the test statistic for significance of the components is not appropriate in admixed populations.

Comparisons of EIGENSTRAT and other population substructure adjustment methods have been done by Price and Zhang.[16, 17] Price and colleagues specifically compared EIGENSTRAT to GC under three different scenarios of substructure and three categories of candidate SNPs. The population substructure options were: 1) stratification arises from discrete subpopulations and the cases and controls have only modest ancestry differences, 2) stratification by discrete subpopulations with cases and controls having more extreme ancestry differences, and 3) an admixed population with ancestry differences between cases and controls being based on ancestry risk. The three categories of candidate SNPs used were: 1) completely random SNPs with no association with disease, 2) SNPs with varying frequencies between subpopulations that were not associated with outcome, and 3) SNPs casually related to the outcome. The results show that EIGENSTRAT is more effective in correcting for population substructure and has higher power to detect associations compared to GC in both simulated and real datasets.[16] This was particularly true for SNPs that were highly differentiated between ancestral populations, confirming the inappropriate uniform adjustment GC implements and subsequent loss of power.[16] Another important result of the simulation done by Price is that EIGENSTRAT is not sensitive to the number of principal components adjusted for assuming that the components included were sufficient to capture the ancestry differences.[16] For example, assuming the first 5 components fully captured the population substructure, adjusting the analysis for 10 components did not alter the results.

The population substructure adjustment method comparison performed by Zhang and colleagues compared GC, *Structure*, and PCA using EIGENSTRAT based on four indices: type I error, power, accuracy, and positive predictive value.[17] They compared the three methods considering that sample size, minor allele frequencies (MAF) of the candidate loci, and degrees of substructure all may impact the performance of each respective adjustment method. In short, what they found was that *Structure* and PCA always outperformed GC regardless of sample size, MAF, or degree of substructure. In addition, PCA almost always outperformed *Structure*. When many AIMs are used, *Structure* achieved similar performance to PCA, but the computational burden imposed with increasing numbers of markers limits the utility of *Structure*.

In summary, PCA is a robust statistical method for detecting and adjusting for population substructure. EIGENSOFT is one software program that implements PCA and has many advantages. This package provides a formal statistical test for the significance of principal components that is appropriate under certain circumstances, runs very quickly on large datasets, works well in admixed populations, allows a variety of input genetic markers, and does not force individuals into discrete populations as *Structure* does.[24] However, there are limitations to using the EIGENSOFT program to run PCA. Primarily, user control is lost when relying on a program developed by others. For example, when using EIGENSOFT, it only allows the user to choose up to 10 components to adjust for. The same PCA procedure can be conducted using different statistical analysis platforms (ie. R or Helix Tree http://goldenhelix.com/) and the user limitations of EIGENSOFT are avoided. When conducting PCA in a program other than EIGENSOFT, the user does lose the ability to test for the "significance" of components, a very attractive feature of EIGENSOFT. However, given that the tests for significance of the principal components are not applicable for African-American populations,[24] the dependence on using EIGENSOFT directly is further obviated.

For this dissertation, I implemented PCA using programs other than EIGENSOFT. By implementing PCA in R and Helix Tree, I have more flexibility with the PCA output while retaining the analytic advantages of PCA in EIGNESTRAT. This is similar to what was done by Stokowski and colleagues in their GWAS of skin pigmentation in Southeast Asians.[25]

Population Substructure in GENOA

Prior to adjusting for confounding due to population substructure, it is important to test for the presence of population substructure. One of the most well-known methods used to test for the presence of population substructure is *Structure.*[18] As mentioned, *Structure* uses genetic marker information (microsatellites or SNPs) to assign individuals to a pre-specified number of subpopulations (ie. clusters). A posterior probability for assignment to those subpopulations is then given for each individual and the number of clusters with the highest posterior probability is determined to be most likely. I have tested for the presence of population substructure in the Jackson cohort of GENOA using both microsatellite markers, and then again using SNPs when the genome-wide SNP measures were available. Below I describe the methods and results that were obtained using both types of genetic markers.

Microsatellite Markers in Structure

Using *Structure*, I have tested for the presence of population substructure within the African-American cohort of GENOA. For this analysis, I used 76 microsatellite marker genotypes that were measured in the Jackson and Rochester GENOA cohorts in addition to the Human Genome Diversity Project (HGDP).[26] The HGDP genotyped microsatellites from the Yoruba and Mandenka populations of West Africa. Therefore, when testing for substructure, the populations that served as "parents" to the Jackson, MS cohort of GENOA were African Yoruba and Mandenka populations from the HGDP and the Caucasian GENOA population from Rochester, MN. The average heterozygosity of the 76 microsatellites was 0.215.

Running numerous values of *K* clusters, *Structure* revealed that K=2 was optimal (Figure 4.1). This indicates no distinct subpopulations (ie. clusters) within the GENOA Jackson cohort that descend from populations other than the designated parental populations (ie. Caucasian and West African). In order to further confirm this, I repeated the analysis using various different parental populations. First, I used European HGDP individuals as the Caucasian parental group in place of the Rochester cohort of GENOA. The clustering results were essentially the same indicating that the Rochester GENOA and European HGDP populations have similar genotype pools and either can be used as an appropriate "parent" population for our African-American sample. Second, the analysis was repeated using various additional HGDP parental populations such as numerous African samples (South Africa, Kenya), Asian samples (North and South China), and Native American samples (Colombia, Pima, and Maya). The inclusion of these other samples always resulted in a third, distinct cluster that had no admixture with the African-American GENOA sample. The above analysis confirms that the African-Americans in GENOA are an admixed population with genomic contributions from West Africans and Northern Europeans. The next step was to determine the distribution of admixture for each individual in the sample.

From the analysis described above using 76 microsatellite markers, the African-American cohort of GENOA had a mean Caucasian ancestry of 4% (median 1%, range 0.1%-77.1%) (Figure 4.2). This is concerning since a recent paper of admixture mapping in the Family Blood Pressure Program (FBPP) cohorts found African-Americans of GENOA to have a mean European ancestry of 20%, using 269 microsatellite markers and individuals from Nigeria as the African parental population.[27] The reason for this discrepancy might be because the microsatellites used in the currently described analysis are too few or not informative enough for ancestry purposes. While it is most efficient to use AIMs for determining ancestry, it is not required. These 76 microsatellites were not specifically chosen as AIMs and might not provide enough information to get precise estimates of admixture proportions in our GENOA study sample. Therefore, in order to reconcile our seemingly underestimated admixture proportion estimates from *Structure*, the analysis was redone with a larger number of markers and a different type of genetic marker, SNPs from the Affymetrix 6.0 chip.

SNPs in Structure

Structure was run using SNPs from the Affymetrix 6.0 chip with the goal of obtaining of percent ancestry in the GENOA Jackson cohort consistent with previous reports. In order to do this, we began by identifying a list of 1,509 ancestry informative SNPs included on an African-American admixture mapping panel provided by Illumina (http://www.illumina.com/pages.ilmn?ID=235). From this list, there were 486 SNPs overlapping on the Affymetrix 6.0 panel. In order to include African ancestral information in *Structure*, we also had to identify which of these SNPs were available from HapMap by searching the latest build of HapMap Yoruba data (rel23a_NCBI_Build36) using the online HapMart data-mining tool (http://hapmart.hapmap.org/BioMart/martview). After determining which of these 486 SNPs were successfully genotyped in GENOA after quality control and were also genotyped in HapMap, we had a list of 385 ancestry informative SNPs to run in

Structure. *Structure* was run with burin length of 10,000 and a run length of 10,000 was used to estimate percent ancestry after the burnin time. This burnin and run length was chosen because it was determined to be "more than adequate" by Pritchard in the *Structure* documentation (http://pritch.bsd.uchicago.edu/software/structure2_2.html). Once again, no distinct substructure was identified, only admixture between African and Caucasian ancestors (K=2) (Figure 4.3). The mean percent of Caucasian ancestry in the Jackson cohort of GENOA was estimated to be 16.5% (median 14.8%, range 0.16-60.3%) (Figure 4.4).

There are two main conclusions to be drawn from the above work. First, the SNPs are clearly more informative for ancestry purposes given that the results are more reasonable and in line with previous research. Second, it is clear that a large degree of admixture is present within the African-American cohort of GENOA and methods to adjust for the potential confounding from population substructure are needed.

Adjusting for Population Substructure in GENOA

As mentioned earlier, PCA can be conducted using a variety of genetic marker data including microsatellites and SNPs. I used both types of markers to adjust for population substructure at various points in this dissertation. The work outlined in Chapter 5 focuses on the architecture of candidate genes potentially associated with LVM via main or context dependent effects. For that analysis, I ran PCA in R using microsatellite markers that had been previously genotyped in the GENOA cohorts for linkage analysis purposes. The work outlined in Chapter 6 is a genome-wide SNP association study for main effects associated with LVM and RWT. That analysis uses

SNPs from the Affymetrix 6.0 chip to run PCA in Helix Tree. Below I outline the detailed methods for implementing PCA using both types of genetic markers.

PCA using Microsatellite Markers

Microsatellite genotyping was conducted in GENOA during Phase I. For some of the Jackson participants of GENOA, genotyping occurred at the University of Texas – Houston and for others at the Marshfield Clinic Research Foundation in Marshfield, Wisconsin. Genotyping shifted to Marshfield when the GENOA's partnership in FBPP was solidified by NHLBI and all genotyping for the collaborative was to be done in a central location. In order to account for potential differences in genotypes by location, PCA was conducted stratified by location. There were 993 Jackson, MS individuals with 453 microsatelittes genotyped at Marshfield and 706 individuals with the same 453 markers genotyped at Houston. PCA was conducted using the function prcomp( ) within the "stats" package of R.[28] A requirement of this function is to have a square dataset, in other words, no missing data. In accordance, we first removed all markers with call rates <80% (174 markers in Houston, 93 markers in Marshfield). Of the remaining markers, we replaced any individual missing values with the most frequently occurring genotype at that locus. Ultimately, PCA was run in the Houston sample using 706 individuals with 279 markers and then in the Marshfield sample with 993 individuals and 360 markers. For PCA, each microsatellite allele length was treated as a quantitative variable.[24]

In the Houston sample, the first 20 principal components explained 25.2% of the genomic variation in the sample, 30 components explained 33.2%, and 234 components were needed to explain 90% of the variation. The distribution of proportion of variation explained by each component in the Houston sample can be seen in Figure 4.5. The first five principal components are plotted against each other in Figure 4.6. In the Marshfield sample, the first 20 principal components explained 21.3% of the genomic variation, 30 components explained 28.1%, and 312 components explained 90% of the variation. The distribution of proportion of variation explained by each principal component in the Marshfield sample can be found in Figure 4.7. The first five principal components from Marshfield are plotted in Figure 4.8.

After conducting PCA, the microsatellite datasets from Houston and Marshfield were each merged with files containing outcome data (LVM) and exposure data (candidate gene SNPs) for the respective individuals, resulting in a total sample of 1,328 individuals available for analysis.

Based on the theoretical discussion of confounding due to population substructure earlier in this chapter, the principal components (a measure of genomic admixture) would be associated with the outcome of interest in order to satisfy the third criteria for confounding. Therefore, we tested for statistical association of each of the first 20 principal components with the outcome. After adjustment for age, sex, SBP, height and weight, three of the principal components were significantly associated with LVM in the Houston cohort and two of the components were associated with LVM in the Marshfield cohort (See Table 4.1.).

Given the work done by Price showing that adjusting for additional components does not significantly alter the results of association testing[16], we decided to conservatively adjust for 20 components in order to sufficiently capture genomic variation within the dataset. The first 20 principal components were used as the quantitative adjustment variables in a least squares linear regression model in order to adjust the outcome of interest for population substructure. The adjustment was done separately in the Houston and Marshfield cohorts. The residuals from these two adjustment models were then combined and used as the single outcome variable to conduct the genetic association analyses discussed in Chapter 5.

PCA using Genome-Wide SNPs

Genome-wide SNPs were genotyped in the GENOA cohort beginning in 2008 using the Affymetrix 6.0 chip. Given the drastically increased number of genetic markers and genomic coverage these SNPs provide compared to the previously described microsatellite markers, these markers are a more appealing option for future PCA analyses. The genome-wide analysis outlined in Chapter 6 uses the genome-wide SNPs in PCA to adjust for population substructure. A similar procedure was applied to conduct PCA using the genome-wide SNPs as was described above with the microsatellites. However, because of the increased data dimensions (76 microsatellites increased to ~700,000 SNPs) and computational time needed, we used the commercially available software Helix Tree (http://goldenhelix.com/) to run PCA for genome-wide SNPs. In addition to increased computational efficiency, another advantage of Helix Tree over R is that SNPs with missing data are not completely omitted, as they are in R. When Helix Tree (and EIGENSTRAT) conducts PCA, an additive model is assumed for the SNPs and SNPs are standardized with mean 0 and variance 1. A missing SNP for an individual is set to zero and not included in the estimate of the mean for that SNP.

A total of 738,451 SNPs from 537 African-American GENOA participants with echocardiography were available for this analysis. Plots of the first five principal components in pairwise combinations can be seen in Figure 4.9. Because of collaborations with the HyperGen cohort of FBPP for the genome-wide association study of echocardiography traits, our analytic strategy is dictated by the work HyperGen already completed. To be consistent and increase comparability across study cohorts, we adjusted the outcomes of interest (LVM and RWT) for the first 30 principal components, as was done by HyperGen. These 30 components explained approximately 11.90% of the genomic variability. Statistical associations of these components with the two outcomes of interest were also considered and can be found in Tables 4.2a-b. After using least-squares linear regression to adjust LVM and RWT for the first 30 components, the residual phenotypes were used as the outcomes of interest for genome-wide association studies of Chapter 6.

Summary

The past few years have seen a surge in population-based genetic association studies involving very large cohorts from all over the world. Subsequently, the debate on whether or not it is important to consider population substructure and how to best adjust for it is very active in the literature. We have outlined three of the most popular and well-accepted methods for detecting and adjusting for population substructure: genomic control, structured association, and principal component analysis. We have determined that population substructure is present in the African-American cohort of GENOA in the form of admixture. Based on the literature review, we believe that principal component analysis is the most appropriate method (relative to *Structure* and genomic control) to use to adjust for population substructure in a highly admixed population of African-Americans.

Table 4.1 Significace of principal components estimated using microsatellites in association with logLVM after adjusted for age, sex, SBP, height, and weight.

Subjects genotyped at University of Texas-Houston				
Principal Component	**N**	**β Estimate**	**β St Error**	**P-value**
PC1	561	0.0007	0.0005	0.1745
PC2	561	-0.0005	0.0005	0.3788
PC3	561	0.0008	0.0006	0.1549
PC4	561	-0.0004	0.0006	0.4565
PC5	561	0.0011	0.0006	0.0566
PC6	561	0.0007	0.0006	0.2361
PC7	561	0.0009	0.0006	0.1435
PC8	561	-0.0014	0.0006	0.0256
PC9	561	-0.0008	0.0007	0.2289
PC10	561	0.0002	0.0007	0.7498
PC11	561	0.0003	0.0007	0.6049
PC12	561	-0.0007	0.0007	0.3378
PC13	561	-0.0013	0.0007	0.0646
PC14	561	-0.0008	0.0007	0.2844
PC15	561	0.0000	0.0007	0.9837
PC16	561	-0.0010	0.0007	0.1723
PC17	561	0.0012	0.0008	0.1075
PC18	561	0.0003	0.0007	0.6691
PC19	561	0.0000	0.0008	0.9704
PC20	561	-0.0007	0.0007	0.3834
Subjects genotyped at Marshfield Clinic				
Principal Component	**N**	**β Estimate**	**β St Error**	**P-value**
PC1	765	-0.0010	0.0003	0.0054
PC2	765	-0.0001	0.0004	0.8502
PC3	765	0.0003	0.0004	0.4053
PC4	765	0.0001	0.0004	0.8234
PC5	765	-0.0004	0.0005	0.3626
PC6	765	0.0007	0.0005	0.1293
PC7	765	0.0007	0.0005	0.1588
PC8	765	0.0014	0.0005	0.0052
PC9	765	-0.0004	0.0005	0.4193
PC10	765	-0.0006	0.0005	0.2537
PC11	765	-0.0002	0.0005	0.6909
PC12	765	-0.0010	0.0005	0.0418
PC13	765	0.0006	0.0005	0.2363
PC14	765	-0.0001	0.0005	0.7990
PC15	765	0.0000	0.0005	0.9734
PC16	765	-0.0007	0.0006	0.2135
PC17	765	-0.0002	0.0006	0.7253
PC18	765	-0.0002	0.0006	0.7853
PC19	765	-0.0009	0.0006	0.1157
PC20	765	0.0001	0.0006	0.8459

Table 4.2a. Significance of principal components estimated using SNPs in association with logLVM after adjustment for age, age^2 sex, height, weight, waist, diabetes, and number of anti-hypertensive medications.

Outcome	Principal Component	N	β Estimate	β St Error	P-value
logLVM	PC1	537	0.2226	0.2399	0.3539
	PC2	537	0.4177	0.2183	0.0562
	PC3	537	0.2584	0.2450	0.2920
	PC4	537	0.0674	0.2221	0.7615
	PC5	537	-0.0694	0.2316	0.7645
	PC6	537	-0.4767	0.2212	0.0316
	PC7	537	0.1159	0.2291	0.6132
	PC8	537	-0.0917	0.2380	0.7002
	PC9	537	-0.4032	0.2272	0.0766
	PC10	537	-0.1513	0.2249	0.5013
	PC11	537	0.0248	0.2477	0.9202
	PC12	537	0.0437	0.2486	0.8604
	PC13	537	-0.2120	0.2371	0.3718
	PC14	537	0.1746	0.2459	0.4780
	PC15	537	0.0139	0.2230	0.9505
	PC16	537	-0.3353	0.2465	0.1744
	PC17	537	0.1512	0.2257	0.5033
	PC18	537	0.3362	0.2222	0.1308
	PC19	537	-0.2106	0.2358	0.3722
	PC20	537	-0.2633	0.2220	0.2362
	PC21	537	0.0895	0.2286	0.6956
	PC22	537	-0.1449	0.2232	0.5166
	PC23	537	-0.0537	0.2364	0.8202
	PC24	537	-0.4909	0.3130	0.1174
	PC25	537	0.0655	0.2907	0.8218
	PC26	537	0.0632	0.2782	0.8204
	PC27	537	0.2193	0.2761	0.4272
	PC28	537	0.3742	0.2889	0.1958
	PC29	537	0.4116	0.2427	0.0905
	PC30	537	-0.2915	0.2556	0.2547

Table 4.2b. Significance of principal components estimated using SNPs in association with logRWT after adjusting for age, age^2 sex, height, weight, waist, diabetes, and number of anti-hypertensive medications.

Outcome	Principal Component	N	β Estimate	β St. Error	P-value
logRWT	PC1	537	0.3471	0.1614	0.0320
	PC2	537	-0.1376	0.1477	0.3518
	PC3	537	-0.0511	0.1655	0.7578
	PC4	537	0.0566	0.1499	0.7059
	PC5	537	0.1057	0.1563	0.4991
	PC6	537	0.0196	0.1500	0.8958
	PC7	537	0.1233	0.1546	0.4254
	PC8	537	-0.2278	0.1603	0.1559
	PC9	537	0.0578	0.1538	0.7072
	PC10	537	-0.0069	0.1518	0.9636
	PC11	537	-0.0755	0.1671	0.6517
	PC12	537	-0.0711	0.1678	0.6718
	PC13	537	0.0509	0.1601	0.7507
	PC14	537	0.0726	0.1660	0.6621
	PC15	537	0.1234	0.1504	0.4122
	PC16	537	-0.0608	0.1666	0.7151
	PC17	537	0.1836	0.1522	0.2282
	PC18	537	0.1564	0.1501	0.2980
	PC19	537	-0.0279	0.1592	0.8607
	PC20	537	-0.4949	0.1485	0.0009
	PC21	537	-0.0084	0.1543	0.9566
	PC22	537	0.2920	0.1501	0.0523
	PC23	537	-0.2582	0.1592	0.1053
	PC24	537	-0.3296	0.2112	0.1193
	PC25	537	-0.1259	0.1961	0.5213
	PC26	537	-0.0803	0.1877	0.6690
	PC27	537	-0.1310	0.1863	0.4822
	PC28	537	-0.0618	0.1953	0.7517
	PC29	537	0.3357	0.1636	0.0406
	PC30	537	0.0765	0.1727	0.6579

Figure 4.1. Triangle plot from *Structure* depicting admixture between Rochester (lower right corner, green color) and Yoruba/Mandenka (lower left corner, blue color) for the Jackson, MS cohort of GENOA (across the bottom, red color). Seventy-six microsatellite markers were used to generate this plot with K=2 clusters specified.

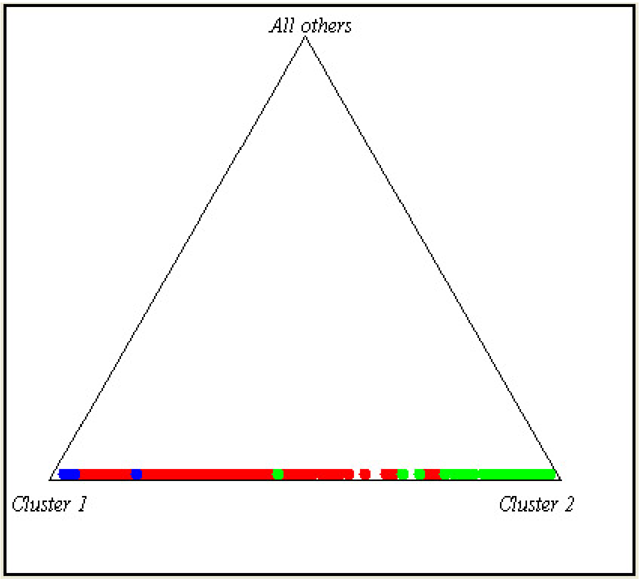


Figure 4.2. Distribution of African Ancestry in the Jackson cohort of GENOA based on the estimates provided by *Structure* when using seventy-six microsatellite markers and Yoruba/Mandenka HGDP Africans and Rochester GENOA participants as the parental populations.

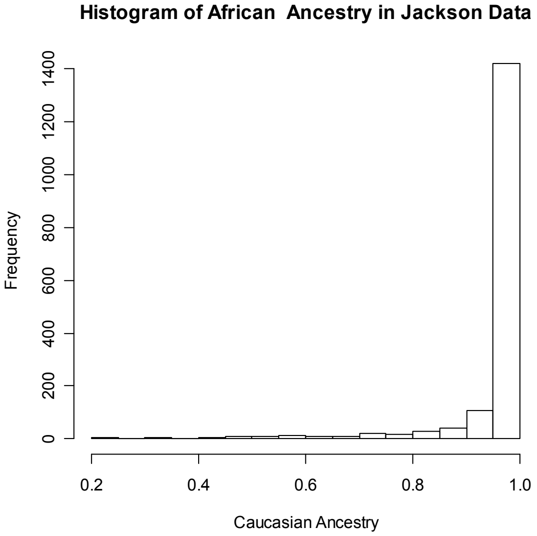


Figure 4.3. Triangle plot from *Structure* with K=2 clusters and 365 African-American ancestry informative SNPs. Cluster one is European (GENOA Rochester in red, n_{roch}=1,236; HapMap CEPH in blue, n_{CEPH}=90). Cluster two is African (GENOA Jackson in green, n_{jack}=680; HapMap Yoruba in yellow, n_{YRI}=90)

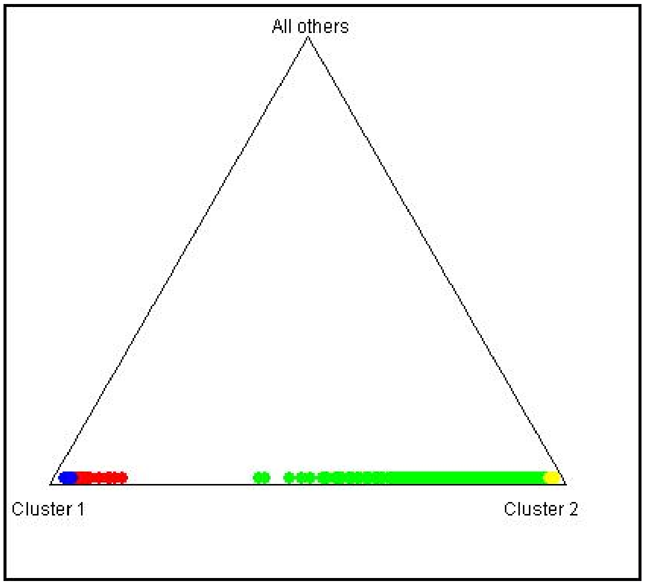


Figure 4.4 Distribution of percent African ancestry in African-American cohort of GENOA (n=680) based on results from *Structure* when K=2 clusters and 365 African-American ancestry informative markers are used.

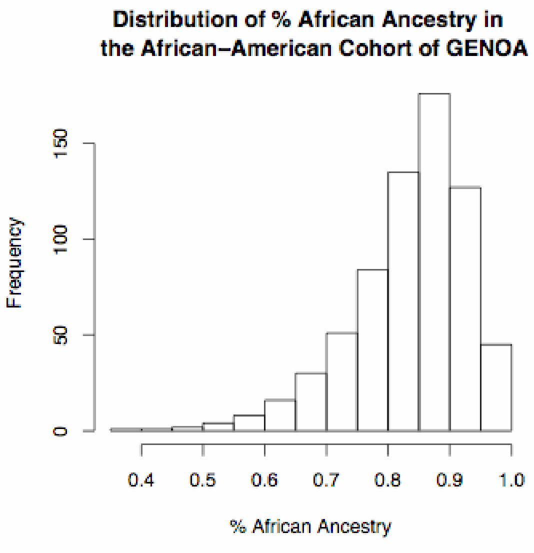


Figure 4.5. Distribution of proportion of genomic variance explained by principal components in the Jackson GENOA individuals with microsatellite genotyping done at University of Texas – Houston, n=706. The area shaded in red represents the first 20 components which cumulatively explain 25.2% of the total variation.

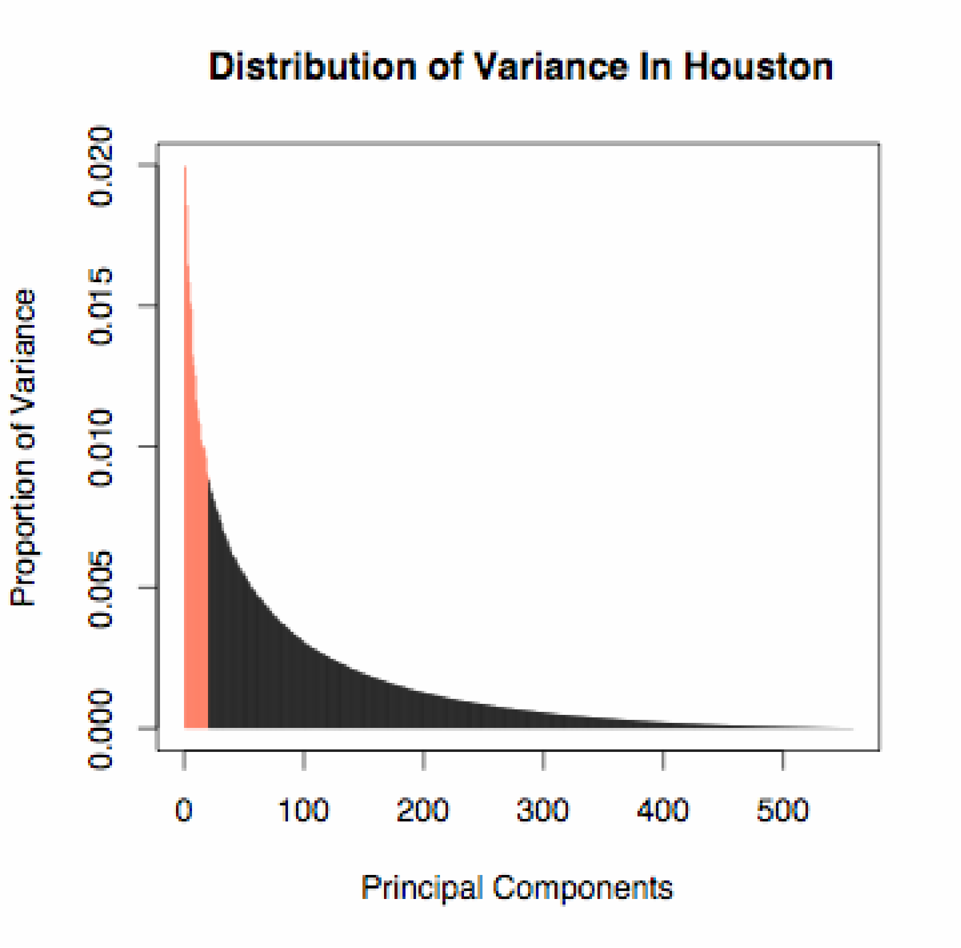


Figure 4.6. Plots of the first five principal components estimated from 279 microsatellites in the subset of GENOA genotyped at UT-Houston, n=706.

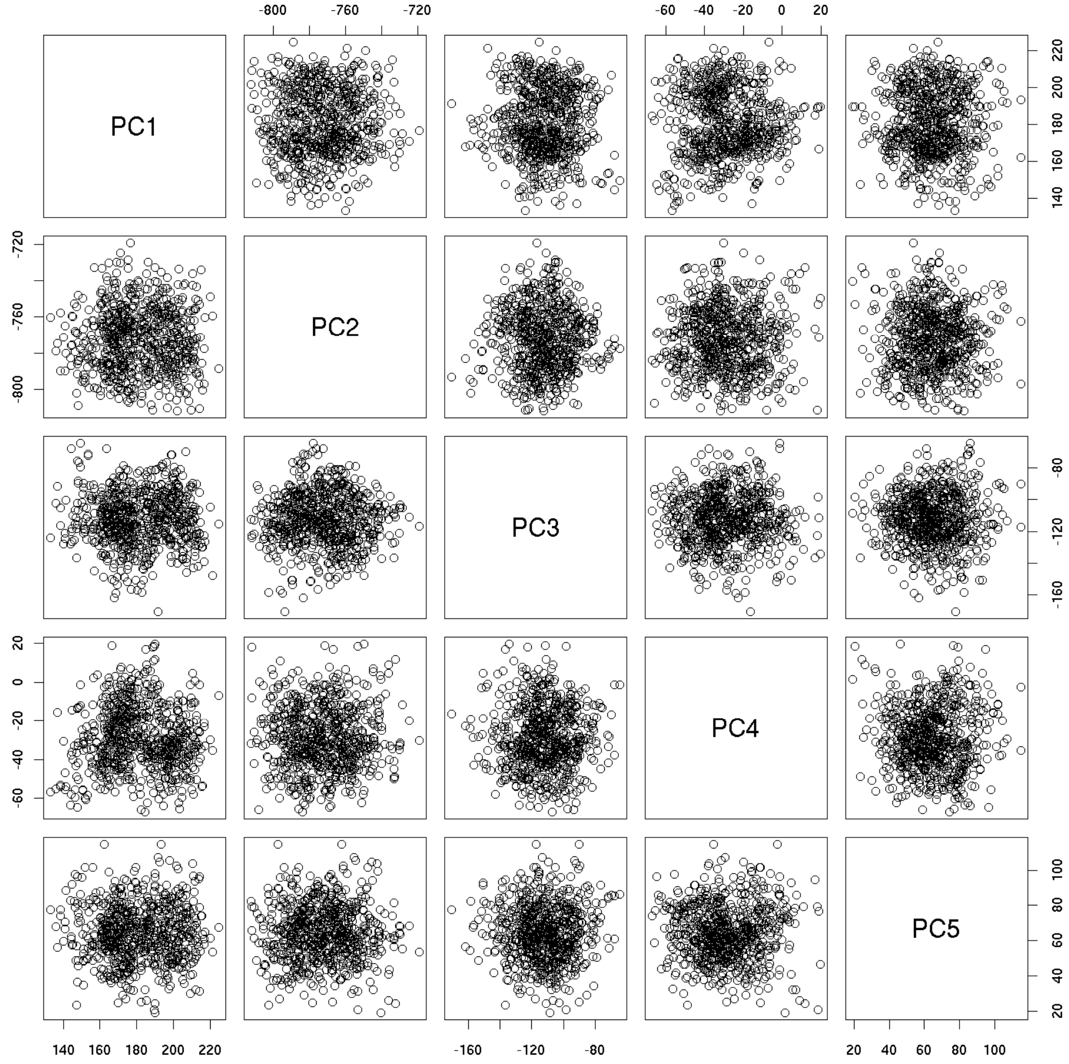


Figure 4.7. Distribution of proportion of genomic variance explained by principal components in the Jackson GENOA individuals with microsatellite genotyping done at the Marshfield clinic, n=993. The area shaded in red represents the first 20 components which cumulatively explain 21.3% of the total variation.

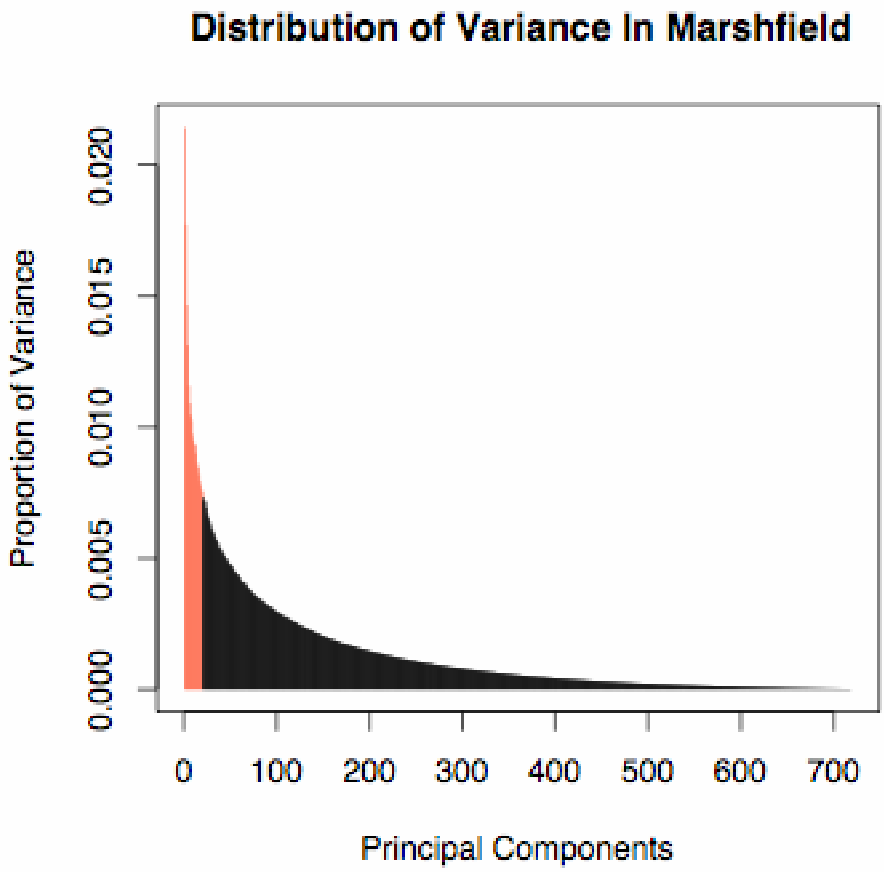


Figure 4.8. Plots of the first five principal components estimated from 360 microsatellites in the subset of GENOA genotyped at Marshfield, n=993.

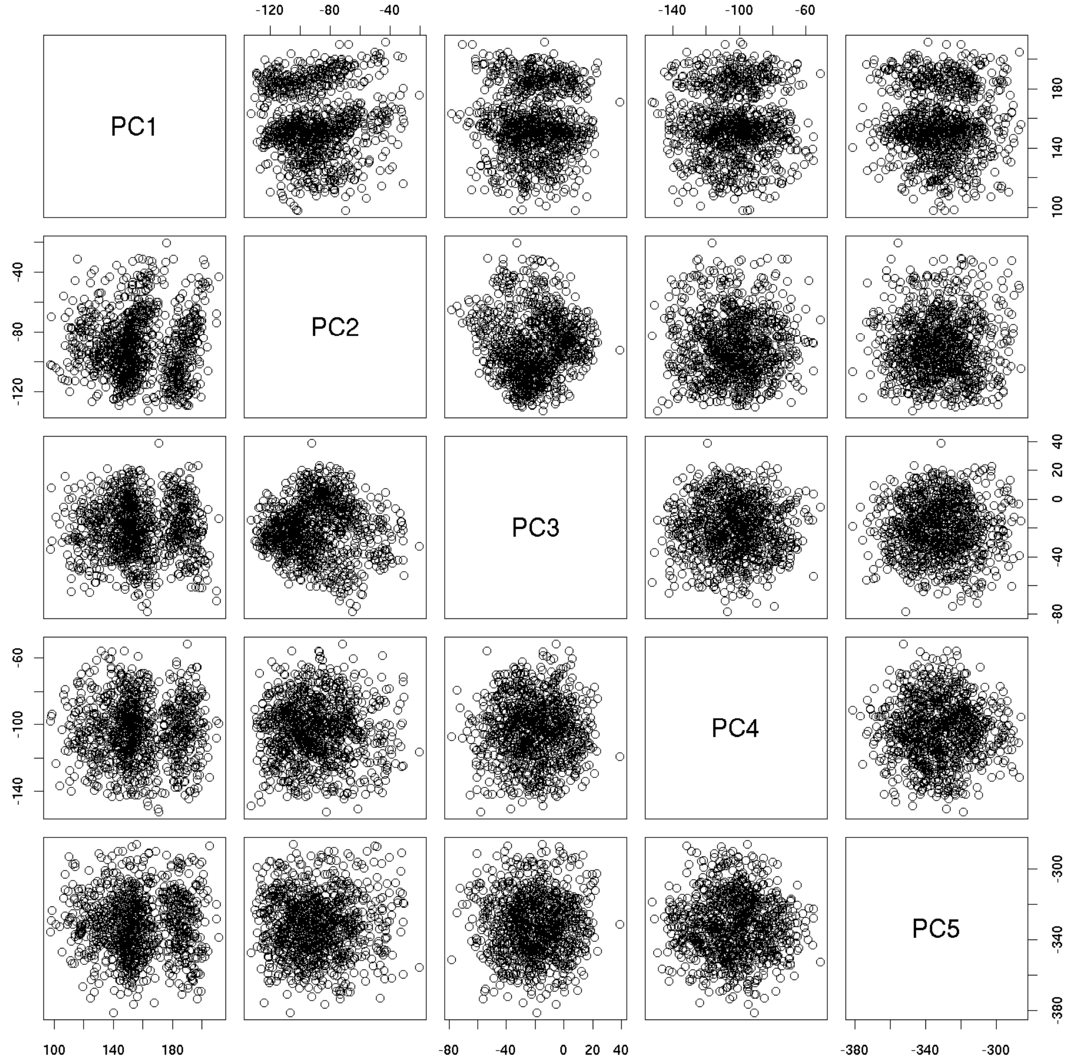
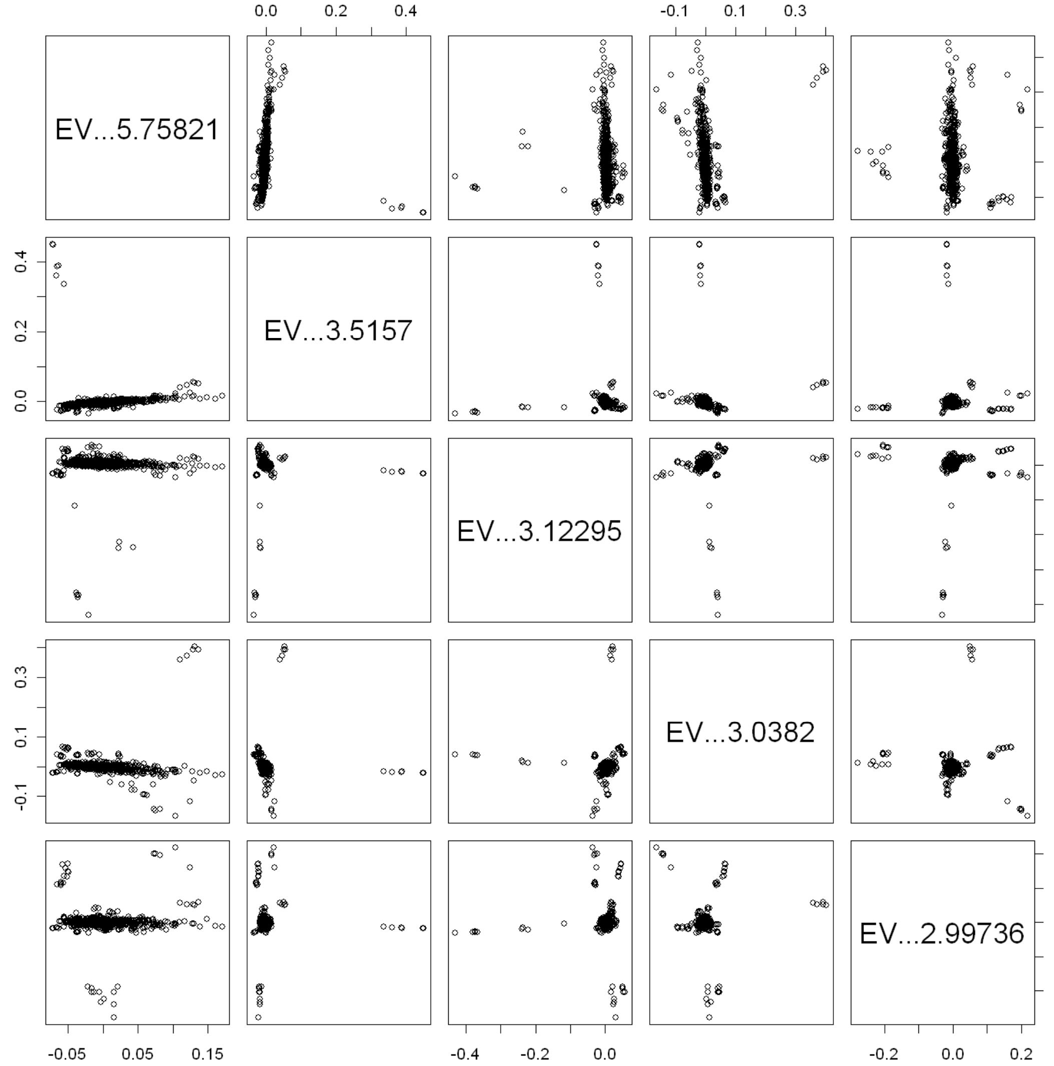


Figure 4.9. Plots of the first five principal components estimated from 738,451 SNPs in the subset of GENOA African-Americans with Affymetrix 6.0 genotype data and echocardiogram outcome data, n=537.



References

1. Freedman ML, Reich D, Penney KL, et al. Assessing the impact of population stratification on genetic association studies. Nat Genet. 2004; 36(4):388-393.

2. Thomas DC, Witte JS. Point: Population stratification: A problem for case-control studies of candidate-gene associations? Cancer Epidemiol Biomarkers Prev. 2002; 11(6):505-512.

3. Hoggart CJ, Parra EJ, Shriver MD, et al. Control of confounding of genetic associations in stratified populations. Am J Hum Genet. 2003; 72(6):1492-1504.

4. Deng HW. Population admixture may appear to mask, change or reverse genetic effects of genes underlying complex traits. Genetics. 2001; 159(3):1319-1323.

5. Devlin B, Roeder K, Wasserman L. Genomic control, a new approach to genetic-based association studies. Theor Popul Biol. 2001; 60(3):155-166.

6. Jackson FL. Illuminating cancer health disparities using ethnogenetic layering (EL) and phenotype segregation network analysis (PSNA). J Cancer Educ. 2006; 21(1 Suppl):S69-79.

7. International HapMap Consortium, Frazer KA, Ballinger DG, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007; 449(7164):851-861.

8. National Center for Health Statistics. Health, 2006: with chartbook on trends in the health of Americans. Hyattsville, MD: 2006.

9. Kramer H, Han C, Post W, et al. Racial/ethnic differences in hypertension and hypertension treatment and control in the Multi-Ethnic Study of Atherosclerosis (MESA). Am J Hypertens. 2004; 17(10):963-970.

10. Hertz RP, Unger AN, Cornell JA, Saunders E. Racial disparities in hypertension prevalence, awareness, and management. Arch Intern Med. 2005; 165(18):2098-2104.

11. Kizer JR, Arnett DK, Bella JN, et al. Differences in left ventricular structure between black and white hypertensive adults: The Hypertension Genetic Epidemiology Network study. Hypertension. 2004; 43(6):1182-1188.

12. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science. 1996; 273(5281):1516-1517.

13. Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999; 55(4):997-1004.

14. Bacanu SA, Devlin B, Roeder K. The power of genomic control. Am J Hum Genet. 2000; 66(6):1933-1944.

15. Bacanu SA, Devlin B, Roeder K. Association studies for quantitative traits in structured populations. Genet Epidemiol. 2002; 22(1):78-93.

16. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38(8):904-909.

17. Zhang F, Wang Y, Deng HW. Comparison of population-based association study methods correcting for population stratification. PLoS ONE. 2008; 3(10):e3392.

18. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000; 155(2):945-959.

19. Choudhry S, Taub M, Mei R, et al. Genome-wide screen for asthma in Puerto Ricans: Evidence for association with 5q23 region. Hum Genet. 2008; 123(5):455-468.

20. Hayes MG, Pluzhnikov A, Miyake K, et al. Identification of type 2 diabetes genes in Mexican Americans through genome-wide association studies. Diabetes. 2007; 56(12):3033-3044.

21. Tian C, Hinds DA, Shigeta R, Kittles R, Ballinger DG, Seldin MF. A genomewide single nucleotide polymorphism panel with high ancestry information for African American admixture mapping. Am J Hum Genet. 2006; 79(4):640-649.

22. Menozzi P, Piazza A, Cavalli-Sforza L. Synthetic maps of human gene frequencies in Europeans. Science. 1978; 201(4358):786-792.

23. Jolliffe IT. Principal Component Analysis. 2nd ed. New York: Springer-Verlag, 2002.

24. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006; 2(12):e190.

25. Stokowski RP, Pant PV, Dadd T, et al. A genomewide association study of skin pigmentation in a south Asian population. Am J Hum Genet. 2007; 81(6):1119-1132.

26. Rosenberg NA, Pritchard JK, Weber JL, et al. Genetic structure of human populations. Science. 2002; 298(5602):2381-2385.

27. Zhu X, Luke A, Cooper RS, et al. Admixture mapping for hypertension loci with genome-scan markers. Nat Genet. 2005; 37(2):177-181.

28. R Core Development Team. R: A Language and Environment for Statistical Computing. 2007; 2.6.0.

Chapter 5

SNP-SNP Interactions Dominate the Genetic Architecture of Candidate Genes Associated with Left Ventricular Mass

Abstract

Left ventricular mass (LVM) is one of the strongest, independent predictors of heart disease incidence and mortality. This research attempts to characterize the complex genetic architecture of LVM in an African-American population by examining the main and interactive effects of individual candidate gene SNPs and conventional risk factors for increased LVM, while also addressing the primary issues of genetic association studies: replication, multiple testing, and population substructure. We used least squares linear regression analysis to investigate the association of 1,878 SNPs in 268 candidate genes with LVM in 1,368 African-Americans from the Genetic Epidemiology Network of Arteriopathy (GENOA) study. We reduced the possibility of false positive results by implementing three analytic strategies: 1) correcting for multiple testing using the false discovery rate, 2) testing for internal replication of results, and 3) using four-fold cross-validation. A multivariable model was built with any SNP main or interactive effects that passed all three criteria. No SNP main effects or SNP-covariate interactions passed all three multiple-testing criteria, 409 SNP-SNP interactions passed all three criteria. A multivariable model including 4 SNP-SNP interactions explained 11.3% of the variation in LVM and increased the predictive ability of LVM in cross-validation procedures. The results of this research underscore that context dependent effects may dominate genetic contributions to complex traits such as LVM.

Introduction

Heart disease, defined as myocardial infarction (MI), hypertensive and ischemic heart disease, and heart failure, is the leading cause of mortality and morbidity in the United Sates.[1] Increased left ventricular mass (LVM) in a well-known, independent risk factor for increased heart disease incidence, mortality, and all-cause mortality.[2-4] LVM can be measured non-invasively via echocardiography and risk factors associated with increases in LVM include high blood pressure, high dietary salt intake, increased age, male gender, diabetes, and increased body mass index.[5, 6, 6-8] African-Americans experience higher mean values of LVM and have almost twice the amount of left ventricular hypertrophy (clinical threshold for high LVM) compared to a non-Hispanic white population.[5] African-Americans also demonstrate higher rates of heart disease incidence and mortality compared to non-Hispanic white populations.[1] Whether this increased heart disease incidence and mortality is partially due to increased LVM is not clear; however understanding the underlying contributions to increased LVM in an African-American population could provide insight into 1) the racial/ethnic disparities of heart disease incidence and mortality and 2) the etiology of increased LVM, which is highly predictive of later heart disease incidence and mortality.

Heritability and twin studies have demonstrated that genetic factors significantly contribute to the inter-individual variation in LVM in numerous racial/ethnic groups, including African-Americans.[9-12] As a follow-up to heritability studies, numerous candidate gene association studies have attempted to test for associations with genetic variants in pathways involved in LVM. While some of the candidate gene results are promising, there are three broad considerations limiting inferences from the results: 1) the lack of replication of the results, 2) potential confounding due to population substructure, and 3) failure to consider the full spectrum of genetic effects involved in complex traits (ie. context dependent effects). LVM is a complex, quantitative trait and by definition is the result of environmental factors, genetic factors, and interactions between. It is entirely plausible that true genetic effects will not replicate in different study populations because they are specific to a given population, in a given environment[13] or because the true architecture involves unaccounted for epistasis (i.e. gene-gene interactions)[14]. While candidate gene studies are a good first step in further informing genetic studies, most candidate gene studies to date have looked only single SNP contributions to LVM. In order to truly understand genomic contributions to complex traits such as LVM, single SNP associations must be considered in the context of, and in conjunction with, environmental factors and other genetic variants.

The goal of this research is to explore the genetic architecture of LVM by identifying robust, replicated single SNP effects, SNP-environment interactions, and SNP-SNP interactions associated with LVM after adjusting for population substructure and relevant risk factors. In achieving this goal, we are implementing a multi-stage approach that focuses on reducing the number of false-positive results and shows replication of effects within the study sample. Harnessing the complexity underlying quantitative traits by testing for main and interactive genetic contributions to the inter-individual variation in LVM will inform and improve the ability to build reliable multivariable models that can predict an individual's LVM.

Methods

Study Population

The National Heart Lung and Blood Institute established the Family Blood Pressure Program (FBPP) in 1996, joining established research networks investigating hypertension and cardiac diseases. One of the four networks in FBPP is the Genetic Epidemiology Network of Arteriopathy (GENOA), which recruited hypertensive African-Americans and non-Hispanic white sibships for linkage and family-based association studies to investigate genetic contributions to hypertension and hypertensive target organ damage. Subjects for this particular GENOA sub-study were African-Americans recruited from Jackson, Mississippi. GENOA recruited sibships containing at least two individuals with clinically diagnosed essential hypertension before age 60. Participants were diagnosed with hypertension if they had a previous clinical diagnosis of hypertension by a physician with current anti-hypertensive treatment, or an average SBP ≥140 mmHg or diastolic blood pressure (DBP) ≥90 mmHg on the second and third clinic visit as stipulated by the Joint National Committee-7 guidelines.[15] After identifying each hypertensive sibship, all members of the sibship were invited to participate regardless of their hypertension status. Exclusion criteria included secondary hypertension, alcoholism or drug abuse, pregnancy, insulin-dependent diabetes mellitus, or active malignancy. Informed consent was obtained from all subjects and approval was granted by participating institutional review boards. Data collection for GENOA was conducted over two phases: phase I (1995-1999) and phase 2 (2000-2004). By the end of Phase II, GENOA had enrolled 1,482 African-American subjects from Jackson, MS.

Phenotype Measurement

Phase I and Phase II data collection consisted of demographic information, medical history, clinical characteristics, lifestyle factors, and blood samples for genotyping and biomarker assays. Study visits were conducted in the morning after an overnight fast of at least eight hours. Blood pressure was measured with random zero sphygmomanometers and cuffs appropriate for arm size. Three readings were taken in the right arm after the participant rested in the sitting position for at least five minutes; the last two readings were averaged for the analysis. Diagnosis of hypertension was defined as described above. Height was measured by stadiometer, weight by electronic balance, and body mass index (BMI) was obtained by the standard calculation of dividing each weight (kg) by the corresponding square of height (m^2). Diabetes was considered present if the subject was being treated with insulin or oral agents or had a fasting glucose level ≥126 mg/dL. Smoking status was defined as self-described smoker within the past year. Use of anti-hypertensive medication was based on self-report during the clinical exam. All phenotype data reported in this paper were collected during the Phase II exam.

The outcome of interest, LVM, was derived during Phase II of GENOA using phased-array echocardiographs with M-mode, two-dimensional and pulsed, continuous wave, and colorflow Doppler capabilities. Standardized methods, along with training and certification, were used by field-center technicians to achieve high-quality recordings. Readings were performed at the New York Presbyterian Hospital-Weill Cornell Medical Center and verified by a single highly experienced investigator. The parasternal acoustic window was used to record at least 10 consecutive beats of two-dimensional and M-mode recordings of the left ventricular internal diameter (LVID) and wall thicknesses at, or just below, the tips of the anterior mitral leaflet in long- and short-axis views. Correct orientation of planes for imaging and Doppler recordings was verified using standardized protocols. Measurements were made using a computerized review station equipped with digitizing tablet and monitor screen overlay for calibration and performance of each measurement. LVID and interventricular septal and posterior wall thicknesses (PWT) were measured at end-diastole and end-systole according to the recommendations of the American Society of Echocardiography in up to three cardiac cycles.[16] Calculations of LVM were made using a necropsy-validated formula.[17] LVM has excellent reliability when measured through echocardiography; the correlation between repeated measures of LVM was 0.93 between paired echocardiograms in hypertensive adults.[18] LVM was measured on a total 1,440 African-American participants of GENOA during Phase II.

SNP Selection and Genotyping

One thousand nine hundred and fifty six SNPs from 268 genes known or hypothesized to be involved in blood pressure regulation, lipoprotein metabolism, inflammation, oxidative stress, vascular wall biology, obesity and diabetes were identified from the genetic association literature and positional candidate gene studies to be genotyped in the entire GENOA population.[19] SNPs were chosen based on a number of different criteria including the published literature, non-synonymous SNPs with a minor allele frequency (MAF) >0.02, and tag SNPs identified using public databases such as dbSNP (http://www.ncbi.nlm.nih.gov/SNP/) and the Seattle SNPs database (http://pga.mbt.washington.edu).

DNA was isolated using the PureGene DNA Isolation Kit from Gentra Systems (Minneapolis MN). Genotyping, based on polymerase chain reaction (PCR) amplification techniques, was conducted at the University of Texas-Health Sciences Center at Houston using the TaqMan assay and ABI Prism® Sequence Detection System (Applied

Biosystems, Foster City CA). Primers and probes are available from the authors upon request. Quality control measures for genotyping assays included robotic liquid handling, separate pre- and post-PCR areas, standard protocols and quality control analyses including 5% duplicates, positive and negative controls, computerized sample tracking, and data validity checks. After removal of SNPs that were monomorphic in the study sample, 1,878 SNPs remained for analysis.

Statistical Analysis Methods

Population Substructure

Population substructure is a term used to describe a single population that is composed of individuals admixed from two or more "parent" populations.[20] The presence of population substructure is a concern for genetic epidemiological studies because this distribution of admixture proportions within a study sample can be a source of confounding, resulting in spurious SNP-disease associations.[21, 22] Population substructure confounds the association of interest when the risk of disease varies with admixture proportions and the frequencies of the SNPs of interest vary based on admixture proportion.[20] This type of confounding is particularly of concern for genetic association studies in African-Americans, a population with high levels of genetic admixture.

The program *Structure* was used to test for the presence of population substructure in our sample of African-Americans from Jackson, MS.[23] We used 76 microsatellite marker genotypes that have been measured in both the GENOA cohort and also in the Human Genome Diversity Project (HGDP).[24] The populations that served as "parents" to the African-American cohort of GENOA were the HGDP African Yoruba and Mandenka populations and the Caucasian GENOA population from Rochester, MN.

Structure indicated that K=2 clusters had the highest posterior probability, meaning that there were no distinct underlying subgroups (ie. clusters) in our dataset, only admixture between the two "parental" datasets.

The underlying genetic structure of admixture within the African-American GENOA sample can be accounted for through principal component analysis (PCA).[25] Of the 1,482 African-American individuals in phase II of GENOA, 1,368 had 453 microsatellite markers previously genotyped in GENOA for linkage analysis. Chapter 4 describes in detail the procedure used to run PCA using the microsatellites. The first 20 principal components were used to adjust the outcome variable, logLVM, using least-squares linear regression. These 20 components described approximately 20% of the underlying genetic variation in the sample.

Descriptive Statistics

High throughput data analyses were conducted using the statistical language R version 2.6.[26] Descriptive statistics for covariates, outcome variables, and SNPs were generated in the full sample and the replication subset samples. LVM was transformed using the natural logarithm. Allele and genotype frequencies were calculated using standard gene counting methods. Hardy-Weinberg equilibrium (HWE) was assessed using a chi-square test or Fisher's exact test if a genotype class had less than 5 individuals.[27] Associations between each SNP and covariate where tested using least-squares linear regression and covariate correlations where assessed in order to identify potential confounders and to understand the underlying correlational structure of the associations tested for with LVM. LVM was adjusted for age, sex, SBP, height, weight, and admixture (via the first 20 principal components from PCA) using least-squares linear regression. The residuals from the adjustment model were used as the dependent variable for all association tests.

Association Testing

We used a multi-stage approach in order to identify both main and interactive genetic effects associated with adjusted logLVM. The final sample size for association analyses is 1,328 due to a limited number of individuals after PCA that were missing outcome data or missing risk factor adjustment data.

In the first stage of association testing, we tested each of the 1,878 SNPs for association with adjusted logLVM using least-squares linear regression methods in the full sample (n=1,328).[27, 28] The SNPs were modeled with two degrees of freedom, therefore assuming no underlying genetic model. The SNP_{ii} genotype for each SNP is the reference genotype and is determined by R to be the homozygous genotype of the first alphabetical allele. There are then two dummy variables in the model where variable SNP_{ij}=1 if heterozygous, and 0 if else and SNP_{jj}=1 if homozygous for the latter alphabetical allele and 0 if else. The statistical model used for this test is written below. The model p-value from the below model was used to determine statistical significance of the main effect of each respective SNP.

$$\text{Adjusted logLVM} = \alpha + \beta_1 SNP_{ij} + \beta_2 SNP_{jj}$$

Based on the 1,878 SNPs and 15 chosen covariates, all possible SNP-covariate interactions where assessed for association with adjusted logLVM using least-squares linear regression for a total of 28,075 tests. The covariates considered in the interactions included age, sex, SBP, DBP, height, weight, diabetes status (0/1), hypertension status (0/1), use of anti-hypertensive medication (0/1), duration of hypertension, smoking status (0/1), myocardial infarction (0/1), total cholesterol, low density lipoprotein cholesterol (LDL), and triglycerides. We hypothesize that the SNPs involved in interactions will have weak or no marginal effects, and will therefore determine significance of the SNP-covariate interaction with a partial F-test comparing a full model (including interaction terms and main effects of the variables in the interaction term) to a reduced model that contains the main effects of the covariate and SNP being tested. Below is an example of the full model:

$$\text{Adjusted logLVM} = \alpha + \beta_1\text{Cov} + \beta_2\text{SNP}_{ij} + \beta_3\text{SNP}_{jj} + \beta_4(\text{Cov*SNP}_{ij}) + \beta_5(\text{Cov*SNP}_{jj})$$

The reduced model is:

$$\text{Adjusted logLVM} = \alpha + \beta_1\text{Cov} + \beta_2\text{SNP}_{ij} + \beta_3\text{SNP}_{jj}$$

All possible SNP-SNP interactions were also tested using a partial F-test. Given the exponential number of tests that are being conducted when assessing SNP-SNP interactions (i.e. epistasis), it has been recommended by some to condition tests for SNP-SNP interactions on marginal main effects of SNPs that are at least weakly significant (ex. p-value<0.10).[29] While this method does reduce the number of tests being conducted, it is known that this approach may miss many epistatic associations because not all possible types of SNP-SNP interactions are expected to demonstrate main SNP effects.[30] In the interest of exploring the full genetic architecture of these candidate gene SNPs, all possible SNP-SNP interactions will be considered, a total of 1,740,614 pairwise interactions.

SNPs were again coded with two dummy variables to allow testing for all possible statistical epistatic effects: additive x additive, additive x dominant, dominant x additive, and dominant x dominant.[31] The statistical significance of the SNP-SNP interaction is based on a partial F-test comparing the full model including all interaction terms to a reduced model with only the main effects of each SNP.[31] When all interaction term beta's are equal to zero, no interaction is present.[31] Below is an example of the full and reduced models for SNP-SNP interactions when both SNPs have all three genotypes present, this would be a four degree of freedom test.

The full model is:

$$\text{Adjusted logLVM} = \alpha + \beta_1\text{SNP1}_{ij} + \beta_2\text{SNP1}_{jj} + \beta_3\text{SNP2}_{ij} + \beta_4\text{SNP2}_{jj} + \beta_5(\text{SNP1}_{ij}*\text{SNP2}_{ij}) + \beta_6(\text{SNP1}_{ij}*\text{SNP2}_{jj}) + \beta_7(\text{SNP1}_{jj}*\text{SNP2}_{ij}) + \beta_8(\text{SNP1}_{jj}*\text{SNP2}_{jj})$$

The reduced model is:

$$\text{Adjusted logLVM} = \alpha + \beta_1\text{SNP1}_{ij} + \beta_2\text{SNP1}_{jj} + \beta_3\text{SNP2ij} + \beta_4\text{SNP2}_{jj}$$

Multiple Testing Adjustments

After testing for main and interactive genetic effects associated with adjusted logLVM, the second stage of analysis was focused on reducing the possibility of false-positive association results and replication of results within our GENOA sample. We did this by implementing three analytic approaches: 1) applying a False Discovery Rate (FDR) q-value threshold,[32] 2) internal replication of results between two subsets of the data, and 3) four-fold cross validation (repeated 10 times)[33]. Only associations that passed the pre-determined thresholds for all three approaches where considered positive associations.

The first step for reducing the probability of false positive results was to calculate the FDR q-value for all association tests.[32] FDR is a method that controls for the proportion of "rejected hypotheses" that are rejected falsely. In other words, the proportion of all "significant" association tests you are willing to consider false positives. For the single SNP associations, the vector of model p-values was used to calculate the q-value. While for the SNP-covariate and SNP-SNP interactions, the vectors of partial F test p-values were used to calculate the q-value. An FDR q-value threshold <0.30 was used to determine significance.

The second step to reduce false positive results was to demonstrate replication of effects within our GENOA sample. The replication subset samples were generated by randomly sampling one sibling from each sibship, without replacement, to create the first sample. From the remaining people, we randomly sampled a second sibling from each sibship to establish the second sample. There were a small number of individuals within the GENOA sample whose siblings did not complete the study enrollment; those "singletons" were equally divided between the two samples. A total of n=491 individuals where in subset 1 and n=496 individuals in subset 2. The slight inequality in sample sizes is because the division of individuals into subsets was done prior to adjustment for admixture. Therefore the 5 less individuals in subset 1 were missing microsatellite data and were not included in the association analyses because no adjustment for admixture was possible in those individuals. If a SNP, SNP-covariate, or SNP-SNP association replicated across these two samples with an α=0.10 and passed the other two multiple testing criteria (FDR and cross-validation), it was tested for homogeneity of direction and magnitude of effect. This was done by creating a dummy variable for sample membership (0 if in subset 1, and 1 if subset 2) and using this variable to test for interaction of effect with a partial F test. Failure to reject the null of "no interaction" in the test for heterogeneity was used to consider the association homogeneous.

The third and final approach for minimizing false positive results and testing for replication was to use cross-validation, a method that reduces false positive results by eliminating associations that lack predictive ability in independent test samples. We performed four-fold cross-validation by dividing the full sample into four equally sized groups. Three of the four groups were combined into a training dataset, and the modeling strategy outlined above was carried out to estimate model coefficients. These coefficients were then applied to the fourth group, the testing dataset, to make predictions about the value of the outcome variable of each individual in the independent test sample. This process was repeated for each of the four testing sets. Predicted values for all individuals in the test set were then subtracted from their observed values, yielding the total residual variability (SSE), $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. The total variability in the outcome (SST) – the difference between each individual's observed value and the mean value for the outcome – was then calculated, $\sum_{i=1}^{n}(\bar{y} - y_i)^2$. In order to estimate the proportion of variation in the outcome predicted in the independent test samples, the cross-validated R^2 (CV R^2) was calculated as follows: $CV\ R^2 = \frac{SST - SSE}{SST}$. This cross-validation method provides a more accurate measure of the predictive ability of the genetic models and will be negative when the model's predictive ability is poor. Because random variations in the sampling of the four mutually exclusive test groups can potentially impact the estimates of CV R^2, this procedure was repeated ten times and the CV R^2 values were averaged.[33] Univariate associations were considered cross-validated if the average percent variation predicted in independent test samples was greater than 0.5% and interactions were considered cross-validated if the difference in average percent variation predicted in independent test samples between the full model containing the interaction term and the reduced model containing only main effect terms was greater than 0.5%.

Based on the association tests that passed all three of the above criteria (FDR q-value <0.30, replication in replication datasets with α=0.10, and CV R^2 >0.005), we built a multivariable linear regression model using forward selection in the full sample of GENOA African-Americans (n=1,328). The increase in percent variation of LVM explained (linear regression R^2 and adjusted R^2) was then calculated, as was the increased predictive ability of the model based on the full model CV R^2 with the addition of each term. Because the full sample of individuals contains siblings, the associations that were included in the final multivariable model were also tested using a linear mixed effects model to account for the familial correlation and ensure that the results were not dependent upon the underlying correlation of data.

Results

Descriptive information regarding the African-American individuals of GENOA used for association analysis can be found in Table 5.1. In general, this is an older (mean age is 63) hypertensive cohort (79% hypertensive) with an average BMI of 31.68. Twenty nine percent of the cohort is diabetic and only 13.9% of the cohort is a self-described smoker within the last year. The average LVM for the entire sample is 160.80g. There were no statistically significant differences identified between subset samples 1 and 2, confirming our process for separating the individuals lends to comparable subsets.

There were a total of 1,878 SNPs that were tested for association with adjusted logLVM in both the full sample and in the replication samples. Of these, 221 had a p-value <0.10 in the full sample, the minimum p-value was $9.24\text{x}10^{-4}$ (SNP: rs12460421, q-value=0.738, CV R^2=0.0033). None of these single SNP associations had a q-value

<0.30 and only one had a CV R^2 >0.005 (SNP: rs2182833). Table 5.2 provides a summary of the number of results for the SNP main effects, the SNP-covariate interactions, and the SNP-SNP interactions passing each of the three pre-determined multiple testing criteria.

There were a total of 28,075 SNP-covariate interactions tested. Ten of those had a q-value <0.30 (p-values ranging from 1.95×10^{-6} to 9.59×10^{-5}), 303 replicated across sample subsets, and 112 had a CV R^2 >0.005. However, none of the SNP-covariate interactions passed all three criteria.

Based on the 1,878 SNPs, all possible SNP-SNP interactions were tested for a total of 1,740,614 associations. A total of 409 of these associations passed all three criteria with an FDR q-value <0.30, replicating in both subsets of the data, and had a CV R^2 >0.005. The interaction with the lowest partial F-test p-value in the full sample was rs17876148*rs12971616 (p-value=4.35×10^{-8}, q-value=0.0139, CV R^2=0.0219).

A multivariable model was then built to determine if a significant additional amount of the variation in LVM could be explained by the joint effect of these SNPs and their interactions. In an effort to avoid over-parameterizing the model, only four SNP-SNP interactions where chosen for the final multivariable model. Beginning with the interaction (and respective main effects) with the most significant model p-value in the full sample (rs35314437*rs7552841), a forward selection process was used with the remaining top nine SNP-SNP interaction models. At each decision point of forward selection, the SNP-SNP interaction, and respective main effects that resulted in the lowest partial F-test p-value was chosen to add to the model. The full model for this F test contained the main and interactive effects of the interaction term fixed in the model and the “new” main and interaction terms of interest. The reduced model removed the interaction and main effect terms for the interaction of interest. Table 5.3 shows the detailed association results for the ten most significant SNP-SNP interaction models that were considered in the forward selection process. Ultimately the following four interactions (and eight SNP main effects) were included in the final multivariable model in the order listed: rs35314437*rs7552841, rs257376*rs5267, rs17876148*rs12971616, and rs6745660*rs12460421. The pairwise genotype specific means for these four interactions are shown in Figures 5.1-5.4. Combined, these main effects and interactions explained 11.3% (adjusted R^2) of the variation in log transformed LVM after adjustment for age, sex, SBP, height, weight, and admixture. Table 5.4 outlines the additional variation of LVM explained with the addition of the main and interactive effects of each SNP-SNP interaction and the p-value for the significance of adding those SNPs to the model. Furthermore, the predictive ability of the model increases steadily with the addition of each interaction term as indicated by the increased in CV R^2. The full model CV R^2 with all four SNP-SNP interactions included was 5.56%.

Discussion

LVM is a complex, quantitative trait highly predictive of incident heart disease. While many studies have investigated candidate gene associations with LVM, to our knowledge, no one has investigated the full spectrum of candidate gene effects for association with LVM including SNP main effects, SNP-covariate interactions, and SNP-SNP interactions. Our motivating hypothesis was that variations within positional and functional candidate genes for hypertension and heart disease may be associated with LVM via interactive effects, in addition to single SNP effects. In examining this hypothesis, we have demonstrated that SNP-SNP interactions predominate the genetic architecture of LVM in the African-American cohort of GENOA, while main SNP effects and SNP-covariate interactions of these candidate genes appear to have little impact on LVM.

One notable aspect of the results presented above is the overwhelming presence of statistically significant epistasis in the absence of marginal SNP effects. Given the increasing availability of genome-wide data and knowledge that interactions (SNP-covariate and SNP-SNP) are important in the variation of complex traits, there has been much debate in the literature about how to best test for these interactions while minimizing computational burden and the possibility of false positives.[29, 30] One popular method is to condition searches for interaction on SNPs with initially significant main effects, even if this threshold for "significance" is weak, for example SNPs with main effect p-value <0.10. There are two major flaws in conditioning searches for interaction on SNP main effects. First, this assumes that the SNPs involved in an interaction will exhibit marginal effects, which is not always expected to be the case.[30] Many previous studies have identified epistasis in the absence of main effects, examples are found in dyslipidemia[34], atrial fibrillation[35] and coronary artery disease[36]. Second, conditioning searches for interaction based on initially significant main effects is likely to be biased by the "winners curse", meaning the subsequent search for interaction will be underpowered since it was predicated on an inflated initial estimate. The "winners curse" is a type of ascertainment bias that describes the phenomena when the first positive report for a genetic variant overestimates the true effect size.[37, 38] While an initial effect estimate of a SNP might be high, subsequent studies of this variant are likely to exhibit "regression to the mean", meaning that follow-up studies will result in reduced effect estimates.[39] The "winners curse" is most often discussed in relation to planning and conducting replication studies for reports of main SNP effects, or in regards to estimating the penetrance parameter for a given single variant. However, this winners curse concept is highly relevant to searches for statistically significant interactions when the SNPs considered were conditionally chosen based on the initial report of a SNP. Even if there was reason to hypothesize marginal effects for a SNP involved in an interaction, effects of SNPs meeting this predetermined "significance" threshold would have been overestimated and therefore the follow-up study underpowered and unlikely to identify the interaction.

Because we hypothesize that marginal effects will not always be present, and in an effort to avoid the "winners curse", we choose not to condition association tests for interaction based on main effects. Our results lend evidence to this approach, no matter how weak the marginal effect threshold might be. Of the eight SNPs involved in the four interactions from the multivariable model we built, the range of main effect SNP p-values was 9.24×10^{-4} (rs12460421) to 0.415 (rs12971616) (Table 5.5). Conditioning searches for interaction on main effects in this study would have precluded investigation of two of the four interactions included in the final multivariable from investigation. In general, we found that regardless of the "significance" level of an interaction, approximately 50% of the interactions will not exhibit any marginal effect in either SNP (Table 5.6). If computational burden is a serious concern, or if you truly do expect to see marginal effects of a SNP involved in interactions, then it would be prudent to first adjust the marginal effect estimates of SNPs for the "winners curse" using one of many methods recently developed to overcome this bias prior to conditioning follow-up searches for interaction.[37, 39, 40]

The concern of type I errors in the face of so many hypothesis tests is substantial and valid. Genetic association studies have suffered from a great lack of replicability in the literature. There are various reasons to attribute this lack of replication. Some of this might be due to truly population specific effects resulting from differing allelic and environmental distributions in various geographical regions. Some of the lack of replication could be because the reported results are false positive or overestimated (the "winners curse"). In addition, it could be due to the limited opportunities to test for replication given that the probability a comparable study cohort has the same exact SNP measured is low. Recognizing that replication in an independent cohort might not be possible because of differences in allelic and environmental distributions; we sought to find genetic associations that replicated within our study sample and were robust across numerous multiple testing adjustment methods. The relative low level of agreement between results filtered through FDR, internal replication, and four-fold cross-validation lends evidence that this is a conservative strategy for determining which results are robust and significant.

In addition to careful consideration of multiple testing issues and reducing the possibility of reporting false positive results, we also took care with the issue of population substructure. While we detected no distinct subpopulations (ie. strata) within our African-American study sample, there was clearly a distribution of admixture, which can also result in confounding by population substructure. Even though we were conservative and adjusted for 20 principal components in the analysis, there is still a possibility that we failed to capture a sufficient amount of admixture. This possibility could arise given that the microsatellites used for PCA were not specifically chosen as ancestry informative markers. Some might be concerned that we "overadjusted" for admixture by using 20 components. However, simulation work done by Price showed that it was more important to sufficiently capture the variation with the principal components and that adjusting for more components than necessary did not alter the results of association tests.[41]

A natural question regarding the results presented above is how might these SNPs interact biologically? Given that these SNPs are in "candidate genes", biological plausibility can be argued for any of the SNPs individually. Table 5.5 outlines positional and functional information for each SNP. "Biological epistasis" (first described in 1908 by William Bateson) and "statistical epistasis" (described by RA Fisher in 1918 as a deviation from additivity in linear regression models) have long been confused, and the interpretations from each are starkly different. Inferences of biological, protein-protein interactions are more difficult to make from this research because statistical tests for SNP-SNP interactions will not necessarily mirror tests for biological interactions.[31] However, we did search within the Michigan Molecular Interactions (MiMI, http://mimi.ncibi.org/MimiWeb/main-page.jsp) database and within PubMed (http://www.ncbi.nlm.nih.gov/pubmed/) for previously reported protein interactions between the four pairwise gene interactions in the multivariable model. No protein interactions were identified for the gene combinations reported in Table 5.5. We do not feel that the lack of previously identified protein interactions negates the results we are reporting. Primarily because it is well understood that making the connection between statistical epistasis and biological epistasis is difficult and arguably not permissible. [31, 42, 43] Furthermore, because association testing relies on the concept of linkage disequilibrium, it is always possible at least one of the "causal" SNPs is in a different, but nearby, gene than the reported gene, and therefore we would not expect to see the biological interaction between reported genes.

The genetic architecture of complex traits such as LVM will consist of a variety of genetic effects (main, SNP-covariate interactions, and epistasis), in candidate genes and in areas of the genome with no currently known function, and a spectrum of allelic frequencies ranging from common to rare variants. In this study we focused on main and interactive genetic effects of SNPs within candidate genes that had a broad range of MAF. Future examinations into the genetic architecture of LVM should include 1) follow-up replication studies of the results reported here in order to fully understand the contribution these interactions have on the variation of LVM and 2) investigation of non-candidate gene regions such as those included in the Affymetrix and Illumina genome-wide chips.

Table 5.1. Descriptive statistics for African-American cohort of GENOA used in the analysis of the candidate gene architecture of LVM.

Variable	N	Full Sample			n	Subset 1			n	Subset 2		
Age, years	1368	62.87	±	9.52	491	62.99	±	9.63	496	63.09	±	9.62
BMI, kg/m^2	1363	31.68	±	6.79	488	31.67	±	7.01	494	31.5	±	6.88
SBP, mmHg	1368	138.40	±	21.12	491	139.3	±	21.49	496	138.5	±	20.77
DBP, mmHg	1368	79.53	±	10.92	491	80.28	±	10.76	496	79.92	±	11.35
Pulse Pressure, mmHg	1368	58.83	±	17.44	491	59.07	±	17.49	496	58.62	±	17.09
Height, inches	1363	168.3	±	8.82	488	169.4	±	9.15	494	169.2	±	9.08
Weight, kilograms	1363	89.62	±	19.36	488	90.66	±	19.83	494	90.01	±	19.5
Duration of hypertension, years	1092	16.58	±	12.74	404	16.79	±	13.24	396	16.17	±	12.49
Duration of anti-hypertension med use, years	980	17.33	±	12.01	365	17.29	±	12.32	350	17.21	±	11.74
Total Cholesterol	1354	202.20	±	42.80	488	201.7	±	45.99	491	201.1	±	41.8
LDL	1354	123.10	±	39.94	488	124.7	±	42.64	491	122.4	±	39.55
Log Triglycerides	1354	2.03	±	0.20	488	2.03	±	0.21	491	2.03	±	0.2
LV Mass, grams	1328	160.80	±	47.07	477	167.4	±	51.66	477	163.5	±	46.42
Sex, male	1368	402 (29.4%)			491	187 (38.1%)			496	175 (35.3%)		
Smoker	1368	190 (13.9%)			491	78 (15.9%)			496	79 (15.9%)		
Diabetic	1368	400 (29.2%)			491	150 (30.5%)			496	144 (29.0%)		
Hypertensive	1368	1,081 (79.0%)			491	400 (81.5%)			496	391 (78.8%)		
Use anti-hypertensive medication	1368	963 (70.0%)			491	357 (72.7%)			496	344 (69.4%)		

Table 5.2. Summary of the number of SNP main effects, SNP-Covariate interactions, and SNP-SNP interactions passing three multiple testing criteria: FDR, internal replication, and 4-fold cross validation.

	SNP Main Effects	**SNP-Covariate Interactions**	**SNP-SNP Interactions**
Total # of Tests	1,878	28,075	1,740,614
P-value <0.10*	221	3,217	192,202
FDR q-value<0.30	0	10	3,083
Cross Validation R^2>0.005	1	112	5,007
Replication (P<0.10 both groups)	14	303	17,593
FDR + CV	0	0	1,031
FDR + Replication	0	2	835
Replication + CV	0	24	1,109
FDR + CV + Replication	0	0	409

* P-values for SNP Main effects are the model-values. The SNP-covariate and SNP-SNP interactions p-value was determined from a partial F test comparing a full model (including all interactions and main effects) to a reduced model only containing main effects of covariates and/or SNPs.

Table 5.3. Detailed multiple testing criteria results (FDR, CV, replication) for the 10 most significant SNP-SNP interaction models of the 409 that passed all three criteria in the African-American cohort of GENOA.

SNP1	SNP2	DF for Interaction Test	Interaction P-value in full sample	Model P-value in full sample	Interaction q-value in full sample	CV R^2 Diff. in full sample	Interaction P-value (Sample1)	Interaction P-value (Sample 2)
rs35314437	rs7552841	2	1.78×10^{-7}	3.88×10^{-8}	0.0142	0.0165	0.0202	4.21×10^{-6}
rs257376	rs5267	3	1.33×10^{-6}	9.11×10^{-8}	0.0218	0.0031	0.0965	0.0442
rs2229169	rs6664855	4	2.45×10^{-7}	1.19×10^{-7}	0.0142	0.0094	0.0004	0.0028
rs10482839	rs7552841	3	2.13×10^{-6}	2.78×10^{-7}	0.0256	0.0143	0.0276	3.96×10^{-5}
rs17876148	rs12971616	4	4.35×10^{-8}	3.14×10^{-7}	0.0139	0.0151	1.85×10^{-7}	0.0389
rs936211	rs521898	2	1.17×10^{-6}	1.07×10^{-6}	0.0211	0.0115	0.0663	0.0023
rs6745660	rs12460421	4	0.0002	1.09×10^{-6}	0.2247	0.0158	0.0856	0.0054
rs945032	rs12028945	4	6.45×10^{-6}	1.11×10^{-6}	0.0385	0.0103	0.0028	0.0011
rs17876144	rs12971616	4	7.29×10^{-8}	1.14×10^{-6}	0.0139	0.0120	2.73×10^{-6}	0.0341
rs35314437	rs4846052	1	1.73×10^{-7}	1.15×10^{-6}	0.0142	0.0177	0.0021	4.14×10^{-5}

Table 5.4. Outline of the additional variation in adjusted log LVM explained (adj R^2) and increased predictive ability of model (CV R^2) as the SNP main and interaction effects were added to the multivariable model based on a forward selection process.

Model	**Interaction Terms in Model**	**Total # of Terms in Model**	**R^2**	**Adj R^2**	**Partial F for Additional Terms**	**Full Model CV R^2**
1	(rs35314437 * rs7552841)	5	0.034	0.030	n/a	0.0165
2	Model 1 + (rs257376 * rs5267)	12	0.073	0.064	2.094x10-8 (df=7)	0.0332
3	Model 2 + (rs17876148 * rs12971616)	20	0.108	0.093	2.208x10-7 (df=8)	0.0460
4	Model 3 + (rs6745660 * rs12460421)	28	0.133	0.113	3.631x10-5 (df=8)	0.0556

Table 5.5. Positional and functional details of SNPs included in the final multivariable model.

SNP	Gene	Chromosome	Position	MAF	Type	Biological Processes*	P-value for Main Effect of SNP
rs35314437	MPO	17q	53704206	0.0153	Synonymous	Response to oxidative stress, anti-apoptosis	0.0440
rs7552841	PCSK9	1p	55291340	0.2370	Intron	Cholesterol homeostasis & metabolic processes	0.0215
rs257376	PRKAR2B	7q	106393948	0.4862	Synonymous	Intra-cellular signaling cascade	0.0512
rs5267	NPPC	2q	232615776	0.1958	Non-synonymous	Regulation of BP & vasoconstriction	0.0067
rs17876148	PON2	7q	94877484	0.0934	Intron	None reported	0.1564
rs12971616	CARM1	19p	10875937	0.1303	Intron	Transcription regulation, histone methylation	0.4149
rs6745660	HSPD1	2q	198057781	0.3505	3' near gene	Protein folding, response to stress	0.0188
rs12460421	CARM1	19p	10842352	0.4382	5' near gene	Transcription regulation, histone methylation	9.24x10^{-4}

* Biological processes of the gene are a subset of that reported in the Michigan Molecular Interactions website: http://mimi.ncibi.org/MimiWeb/main-page.jsp

Table 5.6. The number SNP-SNP interactions where both SNPs had "main effect" p-values that were not significant (p-value>0.2) based on a variety of significance thresholds for the interaction. Total number of SNPxSNP interaction tests conducted was 1,740,614.

Interactions with P-values less than:	**# of Interactions**	**# of interactions with both SNP main effect p-value >0.20 (%)**
<0.10	192,202	117,061 (61%)
<0.01	26,284	14985 (57%)
<0.001	4,808	2,446 (51%)
$<1.0x1^{-4}$	1,096	526 (48%)
$<1.0x10^{-5}$	357	185 (52%)

Figure 5.1. Two-locus genotype specific means of LVM for the first SNP-SNP interaction added to the multivariable model: rs35314437 and rs7552841.

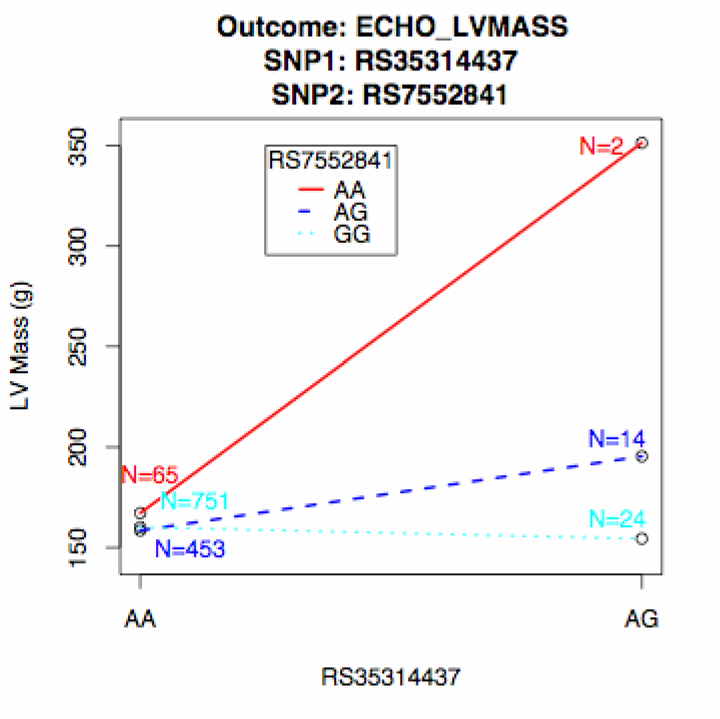


Figure 5.2. Two-locus genotype specific means of LVM for the second SNP-SNP interaction added to the multivariable model: rs257376 and rs5267.

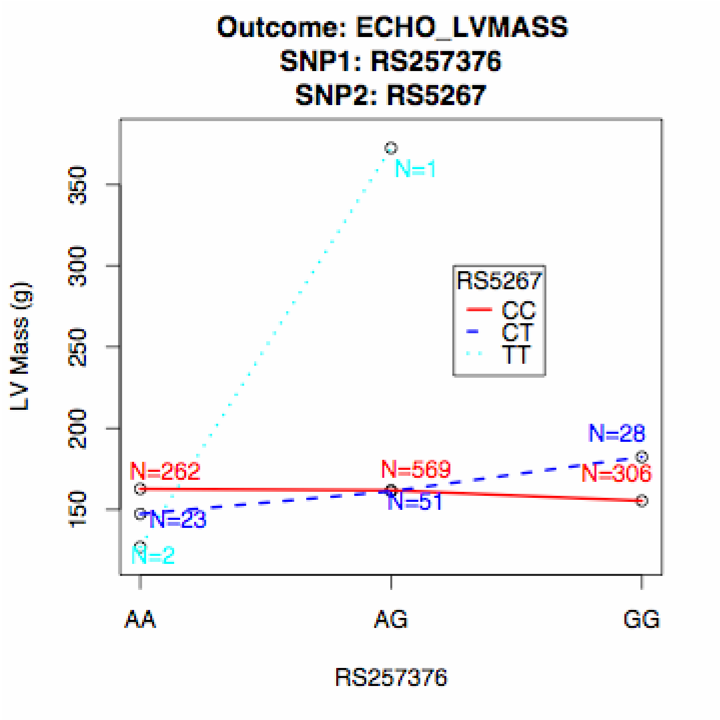


Figure 5.3. Two-locus genotype specific means of LVM for the third SNP-SNP interaction added to the multivariable model: rs17876148 and rs12971616.

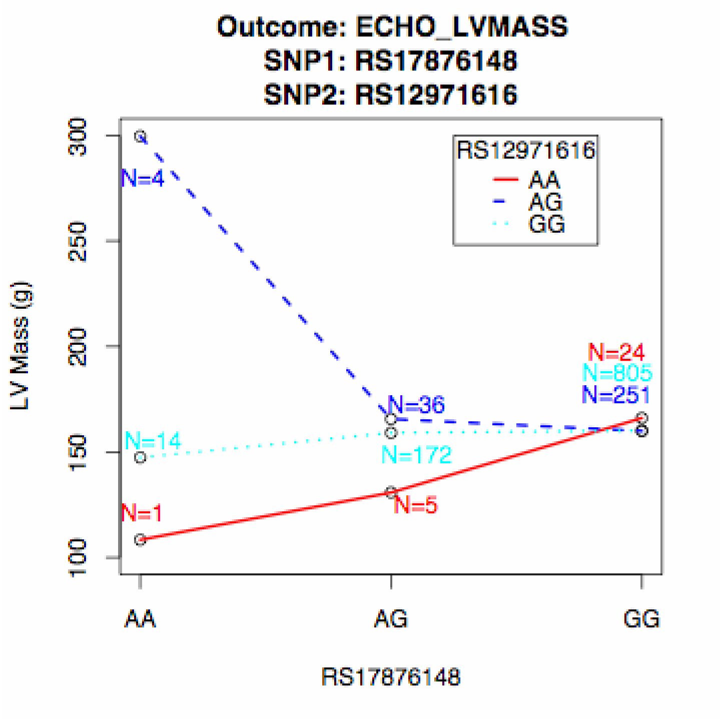
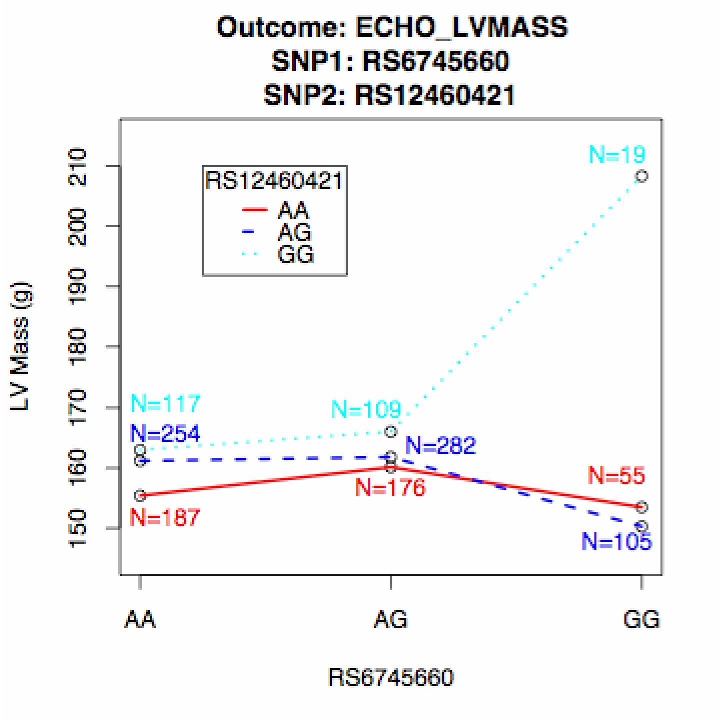


Figure 5.4. Two-locus genotype specific means of LVM for the fourth SNP-SNP interaction added to the multivariable model: rs6745660 and rs12460421.



References

1. National Center for Health Statistics. Health, 2006: with chartbook on trends in the health of Americans. Hyattsville, MD: 2006.

2. Levy D, Garrison RJ, Savage DD, Kannel WB, Castelli WP. Prognostic implications of echocardiographically determined left ventricular mass in the Framingham Heart Study. N Engl J Med. 1990; 322(22):1561-1566.

3. Koren MJ, Devereux RB, Casale PN, Savage DD, Laragh JH. Relation of left ventricular mass and geometry to morbidity and mortality in uncomplicated essential hypertension. Ann Intern Med. 1991; 114(5):345-352.

4. Benjamin EJ, Levy D. Why is left ventricular hypertrophy so predictive of morbidity and mortality? Am J Med Sci. 1999; 317(3):168-175.

5. Kizer JR, Arnett DK, Bella JN, et al. Differences in left ventricular structure between black and white hypertensive adults: The Hypertension Genetic Epidemiology Network study. Hypertension. 2004; 43(6):1182-1188.

6. Galderisi M, Anderson KM, Wilson PW, Levy D. Echocardiographic evidence for the existence of a distinct diabetic cardiomyopathy (the Framingham Heart Study). Am J Cardiol. 1991; 68(1):85-89.

7. Bella JN, Wachtell K, Palmieri V, et al. Relation of left ventricular geometry and function to systemic hemodynamics in hypertension: The LIFE study. Losartan intervention for endpoint reduction in hypertension study. J Hypertens. 2001; 19(1):127-134.

8. du Cailar G, Ribstein J, Mimran A. Dietary sodium and target organ damage in essential hypertension. Am J Hypertens. 2002; 15(3):222-229.

9. Arnett DK, Hong Y, Bella JN, et al. Sibling correlation of left ventricular mass and geometry in hypertensive african americans and whites: The HyperGEN study. Hypertension genetic epidemiology network. Am J Hypertens. 2001; 14(12):1226-1230.

10. Bella JN, MacCluer JW, Roman MJ, et al. Heritability of left ventricular dimensions and mass in American Indians: The Strong Heart Study. J Hypertens. 2004; 22(2):281-286.

11. de Simone G, Tang W, Devereux RB, et al. Assessment of the interaction of heritability of volume load and left ventricular mass: The HyperGEN offspring study. J Hypertens. 2007; 25(7):1397-1402.

12. Sharma P, Middelberg RP, Andrew T, Johnson MR, Christley H, Brown MJ. Heritability of left ventricular mass in a large cohort of twins. J Hypertens. 2006;

24(2):321-324.

13. Sing CF, Stengard JH, Kardia SL. Dynamic relationships between the genome and exposures to environments as causes of common human diseases. World Rev Nutr Diet. 2004; 93:77-91.

14. Wade MJ. Epistasis, complex traits, and mapping genes. Genetica. 2001; 112-113:59-69.

15. Chobanian AV, Bakris GL, Black HR, et al. Seventh report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. Hypertension. 2003; 42(6):1206-1252.

16. Lang RM, Bierig M, Devereux RB, et al. Recommendations for chamber quantification: A report from the American Society of Echocardiography's Guidelines and Standards committee and the Chamber Quantification Writing Group, developed in conjunction with the European Association of Echocardiography, a branch of the European Society of Cardiology. J Am Soc Echocardiogr. 2005; 18(12):1440-1463.

17. Devereux RB, Alonso DR, Lutas EM, et al. Echocardiographic assessment of left ventricular hypertrophy: Comparison to necropsy findings. Am J Cardiol. 1986; 57(6):450-458.

18. Palmieri V, Dahlof B, DeQuattro V, et al. Reliability of echocardiographic assessment of left ventricular structure and function: The PRESERVE study. Prospective randomized study evaluating regression of ventricular enlargement. J Am Coll Cardiol. 1999; 34(5):1625-1632.

19. Barkley RA, Chakravarti A, Cooper RS, et al. Positional identification of hypertension susceptibility genes on chromosome 2. Hypertension. 2004; 43(2):477-482.

20. Hoggart CJ, Parra EJ, Shriver MD, et al. Control of confounding of genetic associations in stratified populations. Am J Hum Genet. 2003; 72(6):1492-1504.

21. Freedman ML, Reich D, Penney KL, et al. Assessing the impact of population substructure on genetic association studies. Nat Genet. 2004; 36(4):388-393.

22. Thomas DC, Witte JS. Point: Population stratification: A problem for case-control studies of candidate-gene associations? Cancer Epidemiol Biomarkers Prev. 2002; 11(6):505-512.

23. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000; 155(2):945-959.

24. Rosenberg NA, Pritchard JK, Weber JL, et al. Genetic structure of human populations. Science. 2002; 298(5602):2381-2385.

25. Menozzi P, Piazza A, Cavalli-Sforza L. Synthetic maps of human gene frequencies in europeans. Science. 1978; 201(4358):786-792.

26. R Core Development Team. R: A Language and Environment for Statistical Computing. 2007; 2.6.0.

27. Weir BS. Genetic Data Analysis II: Methods for Discrete Population Genetic Data. Sudnerland, MA: Sinauer Associates, Inc., 1996.

28. Kleinbaum D, Kupper L, Muller K, Nizam A. Applied Regression Analysis and Other Multivariate Methods. Pacific Grove, CA: Brooks/Cole Publishing Company, 1998.

29. Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet. 2005; 37(4):413-417.

30. Musani SK, Shriner D, Liu N, et al. Detection of gene x gene interactions in genome-wide association studies of human population data. Hum Hered. 2007; 63(2):67-84.

31. Cordell HJ. Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans. Hum Mol Genet. 2002; 11(20):2463-2468.

32. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A. 2003; 100(16):9440-9445.

33. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: A comparison of resampling methods. Bioinformatics. 2005; 21(15):3301-3307.

34. Putt W, Palmen J, Nicaud V, et al. Variation in USF1 shows haplotype effects, gene : Gene and gene : Environment associations with glucose and lipid parameters in the european atherosclerosis research study II. Hum Mol Genet. 2004; 13(15):1587-1597.

35. Tsai CT, Lai LP, Lin JL, et al. Renin-angiotensin system gene polymorphisms and atrial fibrillation. Circulation. 2004; 109(13):1640-1646.

36. Nelson MR, Kardia SL, Ferrell RE, Sing CF. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. Genome Res. 2001; 11(3):458-470.

37. Zollner S, Pritchard JK. Overcoming the winner's curse: Estimating penetrance parameters from case-control data. Am J Hum Genet. 2007; 80(4):605-615.

38. Ioannidis JP. Why most discovered true associations are inflated. Epidemiology. 2008; 19(5):640-648.

39. Zhong H, Prentice RL. Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. Biostatistics. 2008; 9(4):621-634.

40. Xiao R, Boehnke M. Quantifying and correcting for the winner's curse in genetic association studies. Genet Epidemiol. 2009.

41. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38(8):904-909.

42. Moore JH, Williams SM. Traversing the conceptual divide between biological and statistical epistasis: Systems biology and a more modern synthesis. Bioessays. 2005; 27(6):637-646.

43. Kraft P. Curses--winner's and otherwise--in genetic epidemiology. Epidemiology. 2008; 19(5):649-51; discussion 657-8.

Chapter 6

Genome-wide Association Study for Single SNPs Associated with Echocardiography Measures of Left Ventricular Remodeling in African-Americans of the GENOA Study

Abstract

Left ventricular mass (LVM) and relative wall thickness (RWT) are two intermediate traits on the pathway to heart disease. LVM and RWT are strong, independent predictors of heart disease and have a complex etiology involving environmental risk factors, genetics, and their interactions. While many of the environmental risk factors are well understood, the likely numerous genetic variants underlying LVM and RWT are not known. Here we present a genome-wide association study of 738,451 SNPs in 537 African-American subjects of the Genetic Epidemiology Network of Arteriopathy (GENOA) study. After adjustment for known risk factors, population substructure, and family structure, we show an excess of statistically significant results for LVM and RWT. There were 6 SNPs with p-values less than 1.0×10^{-7} in association with LVM. The strongest association was found for adjusted measures of LVM with a p-value=2.59×10^{-8} (rs12102921). The strongest signal for RWT was found at SNP rs1350003 with a p-value=2.18×10^{-7}. Plans for replication of these results with the African-American cohort of HyperGen cohort are in progress.

Introduction

Heart disease, defined as myocardial infarction (MI), hypertensive and ischemic heart disease, and heart failure, is the leading cause of mortality and morbidity in the United Sates.[1] It is well accepted that heart disease is the result of complex interactions between environmental and genetic factors. Unraveling the genetics of heart disease has proven difficult, partly due to the complex nature of the genetics involved, the heterogeneous etiology of heart disease, and because of the heterogeneous 'heart disease' case definitions that are used in research.

The relative lack of success from candidate-gene association studies for heart disease and the identification of a non-coding locus in a gene poor region of chromosome 9 associated with heart disease[2-4] underscore that present knowledge of the genetic and molecular mechanisms leading to heart disease is limited. The agnostic nature of genome-wide association studies (GWAS) provides an ideal tool for identifying new pathophysiologic mechanisms involved in complex traits.[5]

In this study, we tested genome-wide for associations with left ventricular mass (LVM) and relative wall thickness (RWT), in the African-American cohort of GENOA. LVM and RWT are measures used clinically to detect left ventricular remodeling. LVM and RWT are strong, independent predictors of sudden death, ventricular arrhythmias, myocardial ischemia, congestive heart failure, stroke, and angina.[6, 7] Despite the seemingly obvious physiological relationship between increased blood pressure levels and left ventricular remodeling, only a small fraction of the total variability of LVM is explained by blood pressure and upwards of 75% of the variability in LVM may be due to genetic factors.[8] As shown in Chapter 2, in the GENOA African-American cohort, approximately 40% of the variability of LVM after adjusting for risk factors may be due to genetic factors. The goal of this study is to identify novel genetic markers associated with measures of left ventricular remodeling – LVM and RWT – independent of established risk factors such as hypertension.

Methods

Study Population

The Family Blood Pressure Program (FBPP), established by the National Heart Lung and Blood Institute in 1996, joined existing research networks that were investigating hypertension and cardiovascular diseases.[9] One of the four networks in FBPP is the Genetic Epidemiology Network of Arteriopathy (GENOA), which recruited hypertensive, non-Hispanic African-American and Caucasian sibships for linkage and association studies to investigate genetic contributions to hypertension and hypertension-related target organ damage. Sibships containing at least two individuals with clinically-diagnosed essential hypertension before age 60 were recruited from Jackson, Mississippi and Rochester, Minnesota. After identifying each hypertensive sibship, all members of the sibship were invited to participate regardless of their hypertension status. Informed consent was obtained from all subjects and approval was granted by participating institutional review boards.

Participants were diagnosed with hypertension if they had either 1) a previous clinical diagnosis of hypertension by a physician with current anti-hypertensive treatment, or 2) an average systolic blood pressure (SBP) ≥140 mmHg or diastolic blood pressure (DBP) ≥90 mmHg on the second and third clinic visit as stipulated by the Joint National Committee-7 guidelines.[10] Exclusion criteria were secondary hypertension, alcoholism or drug abuse, pregnancy, insulin-dependent diabetes mellitus, or active malignancy. For this study, only African-American subjects from Jackson were included in the analysis of SNP associations with LVM and RWT.

Phenotype measurement

GENOA participants were examined longitudinally in two phases. Phase I (1996-1999) and Phase II (2000-2004) data consisted of demographic information, medical history, clinical characteristics, lifestyle factors, and blood samples for genotyping and biomarker assays. Study visits were conducted in the morning after an overnight fast of at least eight hours. Blood pressure was measured with random zero sphygmomanometers and cuffs appropriate for arm size. Three readings were taken in the right arm after the participant rested in the sitting position for at least five minutes; the last two readings were averaged for the analysis. Diagnosis of hypertension was defined as described above. Height was measured by stadiometer, weight by electronic balance, and body mass index (BMI) was obtained by the standard calculation of dividing each weight (kg) by the corresponding height (m^2). Diabetes was considered present if the subject was being treated with insulin or oral agents or had a fasting glucose level ≥126 mg/dL. All phenotype data reported in this paper were collected during the Phase II exam.

The left ventricular traits of interest, LVM and RWT, were derived using phased-array echocardiographs with M-mode, two-dimensional and pulsed, continuous wave, and colorflow Doppler capabilities during Phase II of the GENOA study. Echocardiograms were only obtained from the African-American cohort of GENOA. Trained and certified field-center technicians used standardized methods in order to achieve high-quality recordings. Echocardiogram readings were performed at the New York Presbyterian Hospital-Weill Cornell Medical Center and verified by a single highly experienced investigator, Dr. Richard Devereux. To measure LVM and RWT, the parasternal acoustic window was used to record at least 10 consecutive beats of two-dimensional and M-mode recordings of the left ventricular internal diameter (LVID) and wall thicknesses at, or just below, the tips of the anterior mitral leaflet in long- and short-axis views. Correct orientation of planes for imaging and Doppler recordings was verified using standardized protocols. Measurements were made using a computerized review station equipped with digitizing tablet and monitor screen overlay for calibration and performance of each measurement. LV internal dimension and interventricular septal and posterior wall thicknesses (PWT) were measured at end-diastole and end-systole according to the recommendations of the American Society of Echocardiography in up to three cardiac cycles.[11] Calculations of LVM were made using a necropsy-validated formula.[12] RWT was calculated as 2*(PWT)/LV internal dimension. LV traits have excellent reliability when measured through echocardiography; e.g. the correlation between repeated measures of LVM was 0.93 between paired echocardiograms in hypertensive adults.[13]

Genotyping

Subjects were genotyped on the Affymetrix Genome-Wide Human SNP Array 6.0 using the Affymetrix protocol.[14] Briefly, 500ng genomic DNA at 50ng/ul in low EDTA-Tris buffer was digested in two separate reaction mixtures using the appropriate restriction enzyme (StyI and NspI, 250ng of DNA for each mixture). This was followed by ligation of an adaptor sequence containing a universal primer sequence. Samples were then subjected to PCR (four PCR reactions per sample for the NspI mixture and three for StyI) with conditions designed to amplify 200-2,000 base pairs. The seven PCR products were combined with Agencourt Ampure beads, passed over an E & K Scientific filter plate, and eluted with buffer. Agarose gel analysis of the PCR products and quantification of the amount of PCR product took place before proceeding further. PCR product concentration was confirmed to be at least 5ug DNA in 1ul EB buffer. Product was then fragmented with DNase I. Agarose gel analysis of the fragmented DNA was used to confirm this step. Following fragmentation, DNA was labeled with Terminal Deoxynucleotidyl Transferase (TdT). Labeled DNA was then hybridized to the appropriate GeneChip and incubated overnight. The chip was stained and washed on the Affymetrix 450 Fluidics station and then scanned on the Affymetrix 3000 GeneChip scanner.

Preliminary SNP genotype calls were generated using the Dynamic Model (DM) algorithm.[15] The final SNP genotype calls were generated by Birdseed, an algorithm designed especially for the Affymetrix Genome-Wide SNP Array 6.0, and based on the robust linear model with Mahalanobis distance classifier algorithm (RLMM).[16] Birdseed utilizes pre-computed models from Affymetrix for each SNP probe set and considers variation across all samples (to help account for experimental variability and population allele probability) to refine the preliminary DM call into genotype calls. Birdseed has been shown to reduce the bias against heterozygous calls and boost call rates to over 99%, while simultaneously increasing concordance rates.[17]

Statistical Analysis

A total of 681 Phase I and Phase II African-American subjects were genotyped for 960,000 Affymetrix 6.0 SNPs. Subjects were excluded if they had an overall SNP call rate <95%. SNPs were excluded if they had a call rate <95%. These quality control filters resulted in 738,451 SNPs available for analysis in 681 African-American GENOA participants.

Descriptive Statistics

Data management and statistical analyses were conducted using version 2.6.0 of R.[18] Descriptive statistics for the covariates and outcome variables are reported in Table 6.1. Histograms and normal quantile plots were created to determine if any variables needed to be transformed. No covariates warranted transformations; however, both outcome variables, LVM and RWT, were transformed using the natural logarithm to reduce skewness and kurtosis. Of the 1,482 African-American individuals in Phase II, 1,440 of them have echocardiogram data. Missing echocardiogram data was a result of either non-participation in that part of the exam or unreadable echocardiograms. Of the 681 African-American subjects with genotype data, 537 of them participated in Phase II of GENOA and thus have LVM and RWT measures available for analysis. Therefore, 537 African-American GENOA subjects from 333 sibships were included in this GWAS. (See Table 6.2 for sibship size distribution)

Population Substructure

Since population substructure is known to be a potential source of confounding in genetic association studies[19-21], we assessed the presence of population substructure in our sample using the program *Structure*.[22] A more detailed description of this analysis can be found in Chapter 4. In short, in order to quantify substructure we identified 365 ancestry informative SNPs that were available in the GENOA Jackson and Rochester samples and in the HapMap Yoruba and CEPH samples. The populations designated as "parents" to the African-Americans in GENOA were the Rochester cohort of GENOA, the HapMap CEPH population, and the African Yoruba from HapMap. Based on these 365 ancestry informative SNPs, no distinct subpopulations were detected in the African-American cohort of GENOA, only admixture between European and African ancestors

(Figure 6.1). The distribution of percent African ancestry within the GENOA African-American sample is plotted in Figure 6.2. Given that population substructure was confirmed to be present in the form of admixture, we used principal component analysis (PCA) to adjust for potential confounding. PCA was implemented in Helix Tree (http://www.goldenhelix.com/SNP_Variation/HelixTree/index.html) using all 738,451 SNPs from the Affymetrix 6.0 array (this includes the 365 AIMs used in *Structure*).[23] The top thirty principal components were retained and used as adjustment variables in association testing.

Association testing

Risk factors for increased LVM and RWT include older age, increased SBP, higher dietary sodium intake, higher BMI, male gender, and the presence of diabetes. Univariate linear regression models for each of these risk factors (excluding dietary sodium intake) were conducted with logLVM and logRWT as the outcomes (Table 6.3). Independent correlates of logLVM included age, gender, BMI, SBP, diabetes status, and number of anti-hypertensive medications (p-value<0.05). Independent correlates of logRWT included age, SBP, diabetes status, and number of anti-hypertensive medications (p-value<0.05).

Replication of results identified through GWAS in an independent cohort is a priority for this research. In order to replicate our findings, we are in the process of establishing collaborations with the HyperGen cohort, another FBPP network. HyperGen recruited hypertensive, African-American sibships from Forsyth County, North Carolina.[24] HyperGen has also performed echocardiography to measure LVM and RWT and conducted genome-wide genotyping from the Affymetrix 6.0 chip. In order to maximize comparability, we mirrored their analysis strategy for our GWAS. LogLVM and logRWT were adjusted for age, age^2, sex, height, weight, waist, diabetes, the number of anti-hypertensive medications a person was on, and 30 principal components from PCA. Based on these adjustment covariates, the multivariable least-squares linear regression model (without principal components) for both outcomes can be found in Table 6.4. Residuals for each of the outcomes of interest were obtained using multivariable linear regression models that included these covariates in addition to the first 30 principal components from PCA. Adjustment using these variables allows us to examine independent (ie. direct) associations between the SNPs and the outcomes.

We then used linear mixed effects (LME) models to test for association between each of the 738,451 SNPs and the multivariable-adjusted, residual phenotypes. LME was used in order to account for the sibship structure among GENOA study participants while retaining a valid type I error rate.[25] No genetic model was assumed, therefore the SNPs were coded using two dummy variables: SNP_{ij}=1 if heterozygous and 0 if otherwise, and SNP_{jj}=1 if homozygous for the latter alphabetical allele and 0 if otherwise. The homozygote for the first alphabetical allele was used as the reference.

Results

Descriptive statistics for the 537 African-American GENOA subjects considered in this analysis are shown in Table 6.1. The study sample is 27.5% male and has a mean age of 62.5 years. The mean SBP is 138 mmHg and they are on average obese with a mean BMI of 32 kg/m^2.

The distribution of minor allele frequencies (MAF) for SNPs analyzed in this study is shown in Figure 6.3. The average MAF was 0.218 (median=0.196). Three hundred eighty four (0.05%) of the SNPs had an MAF less than 0.01, 189,107 (25.6%)

SNPs had an MAF between 0.01 and 0.10, 259,960 (35.2%) SNPs had an MAF between 0.10 and 0.25, and 289,000 (39.1%) SNPs had an MAF between 0.25 and 0.50.

Figures 6.4a-b show the quantile-quantile plots of the expected and observed p-values for the association tests between each SNP and each of the residual phenotypes in the GENOA cohort. Both LVM and RWT exhibited a systematic inflation of p-values across the genome. The inflation factor for both LVM and RWT was approximately 1.9. Figures 6.5a-b shows the distribution of p-values from the LME association tests by chromosome for each of the residual phenotypes. Of the 738,451 SNPs tested, there were 502 SNPs associated with adjusted LVM with a likelihood ratio p-value $<1.0\text{x}10^{-4}$, 70 with a p-value $<1.0\text{x}10^{-5}$, 14 with a p-value $<1.0\text{x}10^{-6}$, and 6 with a p-value $<1.0\text{x}10^{-7}$. For adjusted measures of RWT, there were 301 SNPs with a p-value $<1.0\text{x}10^{-4}$, 25 with a p-value $<1.0\text{x}10^{-5}$, and 4 with a p-value $<1.0\text{x}10^{-6}$. The strongest association for LVM had a p-value=$2.59\text{x}10^{-8}$ (rs12102921), while the strongest association for RWT had a p-value=$2.18\text{x}10^{-7}$. Tables 6.5 and 6.6 list the top 20 SNPs associated with adjusted measures of LVM and RWT, respectively. Tables 6.5 and 6.6 also list the gene the SNP is in, or the distance (bp) to the nearest gene. The strongest association for RWT is SNP rs1350003, which is 492,645 bp away from *G3BP2*. There are two SNPs within the top 20 associations for LVM that also have *G3BP2* as their nearest gene (rs11721411 and rs35570804). The next strongest signal for RWT, rs308717, is 16,236 bp away from *HYAL1*. Again, there is a SNP near this gene (rs6797114) that also was in the top 20 list for association with LVM. There was one more gene, *HADHA*, that had SNPs nearby showing strong associations with both LVM and RWT. While the SNPs are not directly within these genes, the fact that SNPs with the same nearest gene are exhibiting strong statistical associations with both LVM and RWT may provide evidence for pleiotropy.

Discussion

Here we have presented a GWAS attempting to identify new genomic loci associated with LVM or RWT in an African-American cohort. Despite the systematic inflation of p-values across the genome for both traits, there were still an excess of strong p-values observed for both traits compared to chance alone after adjusting for known risk factors, population substructure, and family structure. LVM exhibited stronger p-values compared to RWT, in addition to a larger quantity of significant results. This is consistent with what is expected given that in Chapter 2 we described LVM having a higher estimated heritability compared to RWT in the African-American cohort of GENOA; therefore, we expect to identify more genetic effects associated with LVM. In Chapter 2, we also described the lack of statistical evidence for pleiotropy based on the bivariate genetic correlation for LVM and RWT. However, a cursory look at only the best 20 SNP associations for LVM and RWT indicate that there may indeed be some pleiotropic effects underlying LVM and RWT given that three genes were represented in both lists. If we find this to be the case after replication studies, the reasons the genetic correlation was not statistically significant in Chapter 2 could be due to unrealistic and inappropriate assumptions underlying the test (such as lack of gene-environment or gene-gene interaction) or lack of power to measure the correlation.

One of the strengths of this GWAS is that it was conducted on an intermediate trait along one of the many pathways to heart disease. The molecular and physiological processes leading from variations in genetic code to functional protein changes to clinically detected phenotypes are complex. If we hope to learn more about the etiology of a clinically detectable outcome, investigation of traits preceding clinical disease would be more insightful for understanding etiology.[27] Given that many of the recent GWAS have been conducted with clinical endpoints as the outcome (such as MI and coronary artery disease), much is left to be understood about the genomics of heart disease by studying intermediate traits in the numerous pathophysiological pathways leading to heart disease. LVM and RWT are intermediate traits predictive of heart disease independent of well-established risk factors such as hypertension, cholesterol levels, and smoking. Therefore, investigation of genetic effects underlying intermediate traits such as LVM and RWT may provide new insights into the development of increased LVM or RWT, and subsequent heart disease.

Another notable aspect of this study was the use of an African-American study population. Most of the GWAS to date have been conducted in European samples, regardless of the outcome being studied. Association studies involving admixed populations such as African-Americans are complicated by susceptibility to confounding due to population substructure. However, there are numerous methods available to detect and adjust for population substructure, calming concerns of biased effect estimates. In this study we used principal component analysis to adjust for underlying variation in genomic profiles of the study participants due in part to admixture. Given the varied allele frequencies and environmental distributions across many ethnic groups, future efforts that aim to unravel the complex genetic architecture of various traits must invariably involve understanding how this architecture may vary across racial/ethnic groups.[27] Of the over 300 published GWAS included in the "Catalog of Published

Genome-Wide Association Studies"[34], none were based on an initial search in an African-American population. Therefore, the focus of future research into the genetics of complex traits must give increased consideration to minority groups such as African-Americans.

In addition to the need to replicate these results presented from the GENOA cohort in an independent cohort, some caution must be taken with the interpretation of these results. The quantile-quantile plots in Figure 6.4 exhibit a systematic inflation of observed p-values along the length of the distribution. This is considered a sign of undetected population substructure or other cryptic relatedness.[28] Given that we were conservative and adjusted for 30 principal components in order to correct for potential confounding due to admixture, we are uncertain where this inflation is arising from, it may be due to uncontrolled admixture, families with outlying values of the outcome, or individuals with unique genomic profiles. Regardless of the reason, these results should be interpreted with caution until the origin of the inflation is better understood or replication in an independent cohort has been established.

Future Plans for Replication

Because of the extremely large number of hypothesis tests being conducted, the ability to demonstrate replication of GWAS results in an independent research cohort is important.[29] As previously mentioned, we are in the process of establishing this replication collaboration with the HyperGen cohort. Characteristics of the HyperGen African-American cohort are very similar to GENOA. It is a largely female (69%), obese cohort (mean BMI=32.7) with an average SBP of 134 mmHg.[24] The distribution of LVM and RWT are similar to GENOA, but slightly higher with a mean LVM of 180.6 g and mean RWT of 0.35.[24] In addition, the methods used to measure LVM and RWT are the same in HyperGen and GENOA (echocardiography) and both cohorts used the Affymetrix 6.0 array for genotyping. Thus, the HyperGen cohort is a well-suited replication cohort for our GWAS of LV measures in the African-Americans of GENOA.

In order to strengthen the merits of any replication identified through initial searches, we consulted with HyperGen regarding their statistical modeling strategy for our initial GWAS reported here. Therefore, our modeling strategy matches HyperGen in terms of adjustment covariates, population substructure adjustment methods, and genetic modeling. While a comparison of results for replication will be the first step, it is likely that neither individual study has enough power to detect small effect sizes. Therefore a meta-analysis will also be conducted to increase the power for detecting small effect sizes and to see if statistical evidence was gained for any SNPs found highly significant in either individual study.

Meta-analyses within epidemiology have traditionally used fixed effects models and inverse-variance weighting in order to combine evidence from two or more studies. Fixed effects models assume that the underlying effect for a polymorphism is homogeneous across studies. Tests for heterogeneity of effect (where the null hypothesis is homogeneity) are often cited to defend this assumption. However there are two critical flaws in this argument. First, tests for heterogeneity of effect have low power and therefore may not identify existing heterogeneity. Therefore, failure to reject the null hypothesis of homogeneity does not prove homogeneity.[30] Second, assuming homogeneity of effects across studies undermines the repeated argument that heterogeneity is likely to be the “norm” and a prominent underlying reason genetic associations are difficult to replicate. As discussed by Kavvoura and Ioannidis (2008), the assumption of heterogeneous underlying effect sizes for each study of a meta-analysis is likely to be realistic for genetic association studies.[30] Therefore, using random effects models are preferred over fixed effects models for meta-analyses of genetic associations. The summary estimates of effect from random effects models will be very similar to estimates from fixed effects models, only more conservative yielding wider confidence intervals.[31, 32] For random effects meta-analysis of quantitative outcomes, LME models can be implemented with the estimated regression coefficients as the parameter of interest.[33] The goal of this proposed meta-analysis would be to identify new genomic regions with robust statistical association with LVM or RWT in African-American populations.

Table 6.1. Descriptive statistics for the sub-set of GENOA African-Americans considered in this GWAS**.**

Covariates and Outcome Variables	**N**	**Mean ± SD**
Age, years	537	62.52 ± 9.3
BMI, kg/m^2	537	32.02 ± 6.6
Height, m	537	167.9 ± 8.95
Weight, kg	537	90.14 ± 18.83
Systolic BP, mm Hg	537	137.6 ± 21.29
Diastolic BP, mm Hg	537	79.21 ± 10.81
Left ventricular mass, g	537	162.5 ± 48.31
Log LV Mass	537	5.05 ± 0.28
Relative Wall Thickness	537	0.32 ± 0.05
Log RWT	537	-1.14 ± 0.16
	N	**n (%)**
Male	537	148 (27.5%)
Hypertension	537	427 (79.5%)
Taking anti-hypertensive medications	537	382 (71.1%)
Diabetes	537	168 (31.3%)

Table 6.2. Distribution of sibship size in the sample of 537 African-Americans of GENOA considered in GWAS.

# of Siblings in Sibship	**# of Sibships**	**Total number of individuals**	**% of Total Number of Individuals**
1	187	187	34.8%
2	106	212	39.5%
3	27	81	15.1%
4	9	36	6.7%
5	3	15	2.8%
6	1	6	1.1%
Totals	**333 sibships**	**537 individuals**	**100.0%**

Table 6.3. Univariate associations of covariates with LVM and RWT.

logLVM	**N**	**β**	**P-value**
Age, years	537	0.006346	$5.13\text{x}10^{-7}$
Male gender	537	0.189536	$2.95\text{x}10^{-13}$
BMI, kg/m^2	537	0.010704	$2.06\text{x}10^{-9}$
Height, m	537	0.010369	$9.99\text{x}10^{-16}$
Weight, kg	537	0.006048	$<2.0\text{x}10^{-16}$
Systolic BP, mm Hg	537	0.002924	$1.18\text{x}10^{-7}$
Diastolic BP, mm Hg	537	0.001438	0.191505
Hypertension status	537	0.203637	$1.54\text{x}10^{-12}$
Number of anti-hypertensive Rx	537	0.064873	$5.39\text{x}10^{-11}$
Diabetes status	537	0.118362	$3.16\text{x}10^{-6}$
logRWT	**N**	**β**	**P-value**
Age, years	537	0.004818	$1.73\text{x}10^{-11}$
Male gender	537	-0.00819	0.590862
BMI, kg/m^2	537	0.001961	0.058645
Height, m	537	-0.00132	0.083453
Weight, kg	537	0.000266	0.462207
Systolic BP, mm Hg	537	0.001565	$7.46\text{x}10^{-7}$
Diastolic BP, mm Hg	537	0.000283	0.65372
Hypertension status	537	0.082301	$7.94\text{x}10^{-7}$
Number of anti-hypertensive Rx	537	0.020329	0.000395
Diabetes status	537	0.057703	$7.48\text{x}10^{-5}$

Table 6.4. Multivariable linear regression model with covariates used in the adjustment of LVM and RWT, n=537.

logLVM	**β**	**β St. Error**	**P-value**
(Intercept)	4.0135	0.3978	$< 2 \times 10^{-16}$
Age	-0.0195	0.0096	0.0435
Age^2	0.0002	0.0001	0.0064
Sex	0.102	0.0309	0.001
Height	0.0048	0.0017	0.0058
Weight	0.0049	0.0012	3.52×10^{-5}
Waist	0.0005	0.0014	0.7459
Diabetes	0.0661	0.022	0.0028
Number of anti-hypertensive medications	0.0357	0.0088	6.22×10^{-5}
Adjusted R^2 for model:	**0.3638**		
logRWT	**β**	**β St. Error**	**P-value**
(Intercept)	-1.448	0.2685	1.05×10^{-7}
Age	0.0002	0.0065	0.9769
Age^2	3.16×10^{-5}	5.31×10^{-5}	0.5512
Sex	0.0002	0.0209	0.9906
Height	-0.0002	0.0012	0.8917
Weight	-0.0014	0.0008	0.0734
Waist	0.0029	0.0009	0.0024
Diabetes	0.0361	0.0149	0.0156
Number of anti-hypertensive medications	0.0037	0.006	0.5346
Adjusted R^2 for model:	**0.1144**		

Table 6.5. Top 20 results, based on p-value, in association with logLVM adjusted for risk factors and population substructure, n=537.

SNP rs#	Chromosome	Position	Gene	Distance to Gene (bp)	LR p-value
12102921	16	57020668	*MPG*	23119	2.59E-08
6594004	1	202570244	*PLEKHA6*	0	2.94E-08
1996003	1	202573709	*PLEKHA6*	0	2.94E-08
1956500	14	25904408	*WARS*	80520	4.33E-08
10510440	3	15582005	*HACL1*	0	4.34E-08
1950866	14	25905245	*WARS*	79683	4.41E-08
11721411	4	190586037	*G3BP2*	512930	2.84E-07
7561565	2	136357256	*HADHA*	6775	3.81E-07
7984109	13	33213702	*C13orf31*	76503	5.04E-07
4417762	2	42031828	*HADHA*	96836	5.97E-07
7846615	8	72403490	*EYA1*	0	7.90E-07
11669868	19	58061656	*VASP*	8942	9.65E-07
1902709	10	97994251	*BLNK*	0	9.67E-07
11228152	11	67709704	*SUV420H1*	0	1.04E-06
6797114	3	19831117	*HYAL1*	64852	1.06E-06
11656342	17	64168152	*KRTAP9-8*	59462	1.20E-06
839034	4	57990686	*G3BP2*	319390	1.22E-06
17108317	12	69223360	*PTPRB*	0	1.44E-06
11859393	16	11553229	*LITAF*	0	1.58E-06
35570804	4	22647613	*G3BP2*	217323	1.65E-06

Table 6.6. Top 20 results, based on p-value, in association with logRWT adjusted for risk factors and population substructure, n=537.

SNP rs#	Chromosome	Position	Gene	Distance to Gene (bp)	LR p-value
1350003	4	63113408	*G3BP2*	492645	2.18E-07
308717	3	4361593	*HYAL1*	16236	3.83E-07
7034992	9	124066484	*RBM18*	0	6.50E-07
16929424	8	63770943	*NKAIN3*	0	8.82E-07
17050778	3	36246953	*HYAL1*	150147	1.83E-06
691	15	22915887	*IPW*	0	2.17E-06
6074095	20	10033053	*RPRD1B*	47646	2.26E-06
2064655	20	10038885	*RPRD1B*	53478	2.26E-06
7303392	12	11386358	*LOC100190940*	9665	2.45E-06
7548281	1	231248036	*PCNXL2*	0	2.67E-06
12907340	15	55876467	*CYP1A1*	79422	2.89E-06
3757645	7	111213917	*DOCK4*	0	3.02E-06
3819396	6	15601008	*JARID2*	0	3.18E-06
2016335	20	56633948	*RPRD1B*	25785	3.79E-06
1341058	9	18520187	*ADAMTSL1*	0	3.88E-06
7244187	18	10237424	*MBP*	207200	3.99E-06
207866	2	216830051	*HADHA*	778	4.26E-06
9956842	18	10222898	*MBP*	221726	5.13E-06
16917793	8	53458213	*ST18*	0	6.01E-06
4674710	2	223410632	*HADHA*	23343	6.19E-06

Figure 6.1. Triangle plot from *Structure* with K=2 clusters and 365 African-American ancestry informative SNPs. Cluster one is European (GENOA Rochester in red, n_{roch}=1,236; HapMap CEPH in blue, n_{CEPH}=90). Cluster two is African (GENOA Jackson in green, n_{jack}=680; HapMap Yoruba in yellow, n_{YRI}=90)

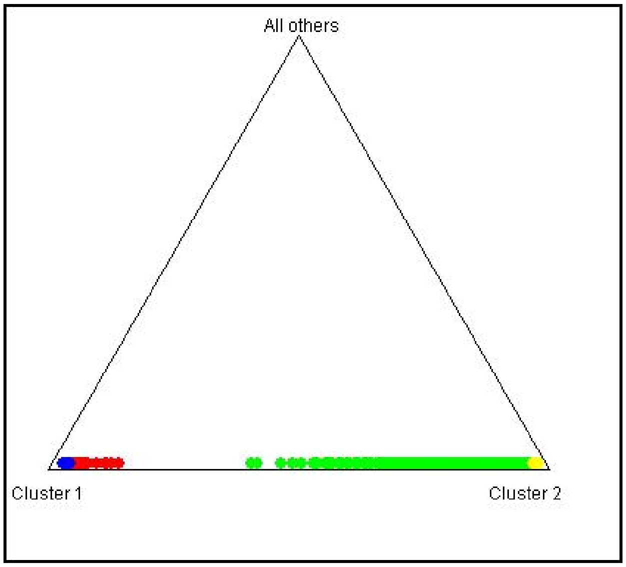


Figure 6.2. Distribution of percent African ancestry in African-American cohort of GENOA (n=680) based on results from *Structure* when K=2 clusters and 365 African-American ancestry informative markers are used.

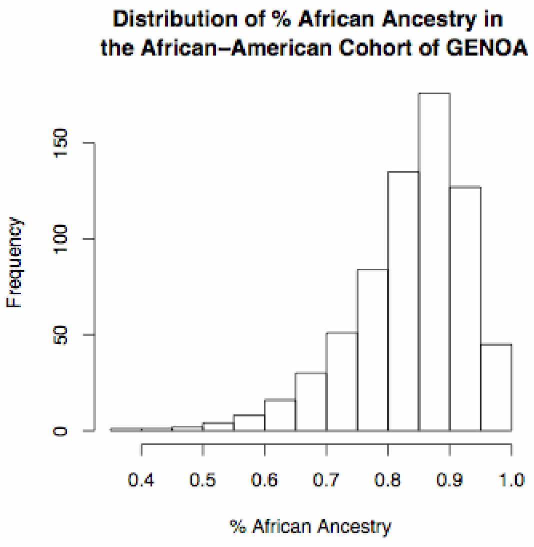


Figure 6.3. Distribution of minor allele frequencies for the 738,451 Affymetrix 6.0 SNPs tested in the African-American cohort of GENOA for association with LVM and RWT.

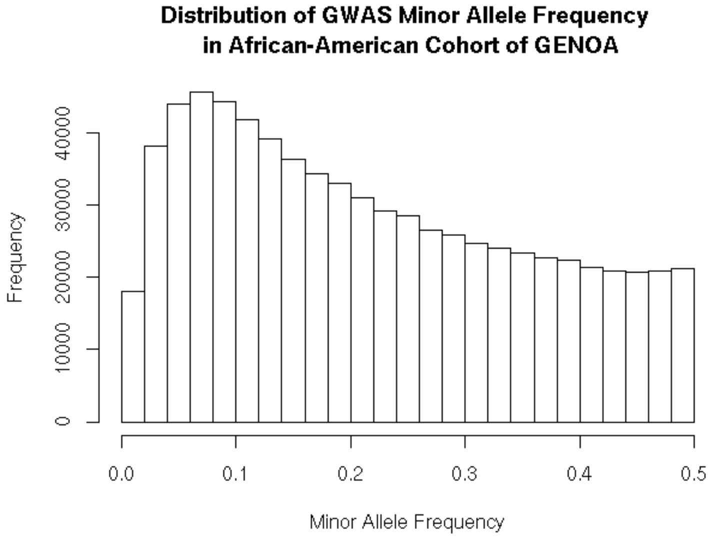


Figure 6.4.a. Quantile-Quantile plot of observed and expected p-values for associations with adjusted logLVM using the LME model.

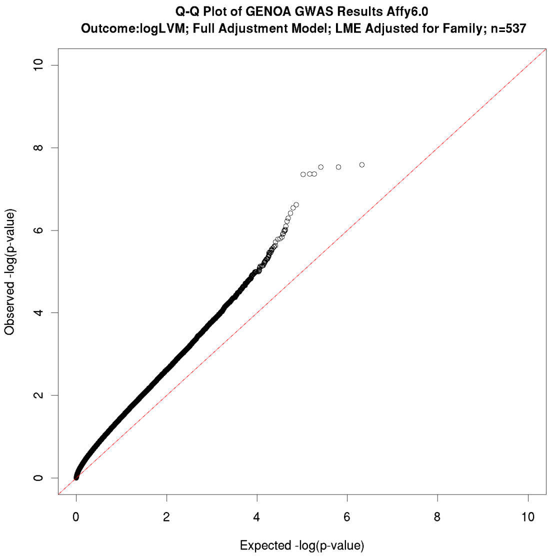


Figure 6.4.b. Quantile-Quantile plot of observed and expected p-values for associations with adjusted logRWT using the LME model.

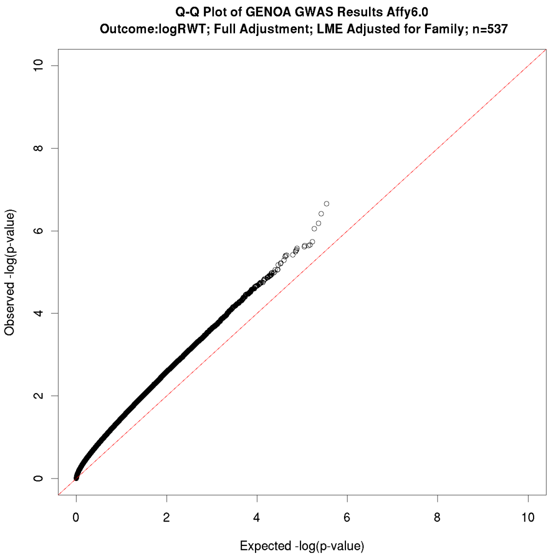


Figure 6.5.a. Distribution of –log p-values from tests of association between Affymetrix 6.0 array SNPs and adjusted LVM in GENOA African-Americans

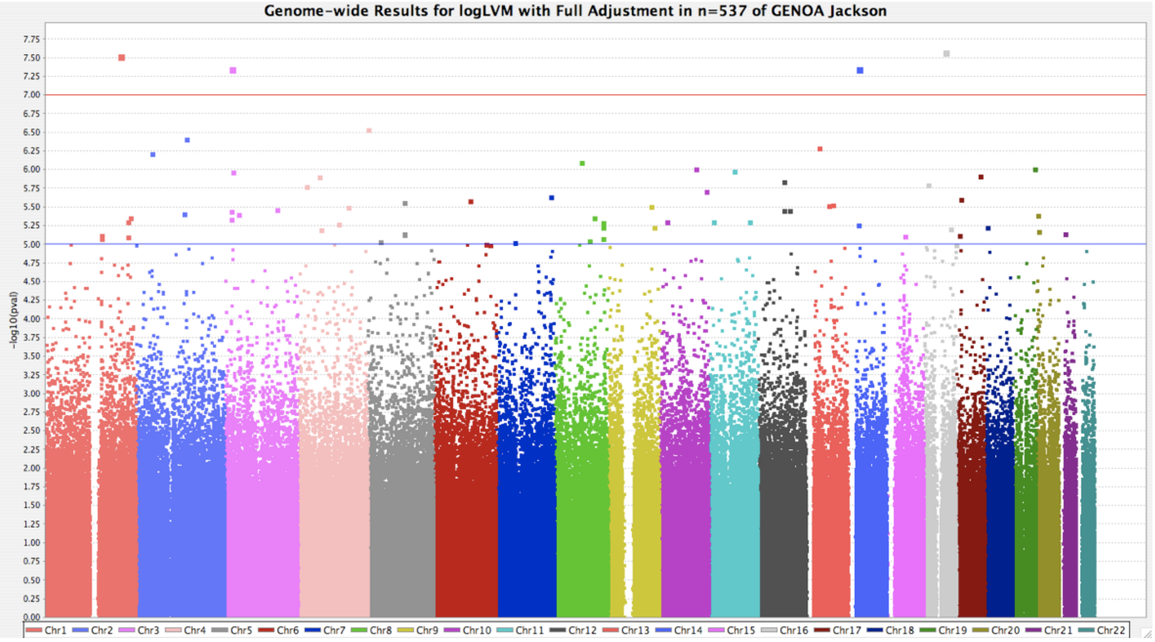
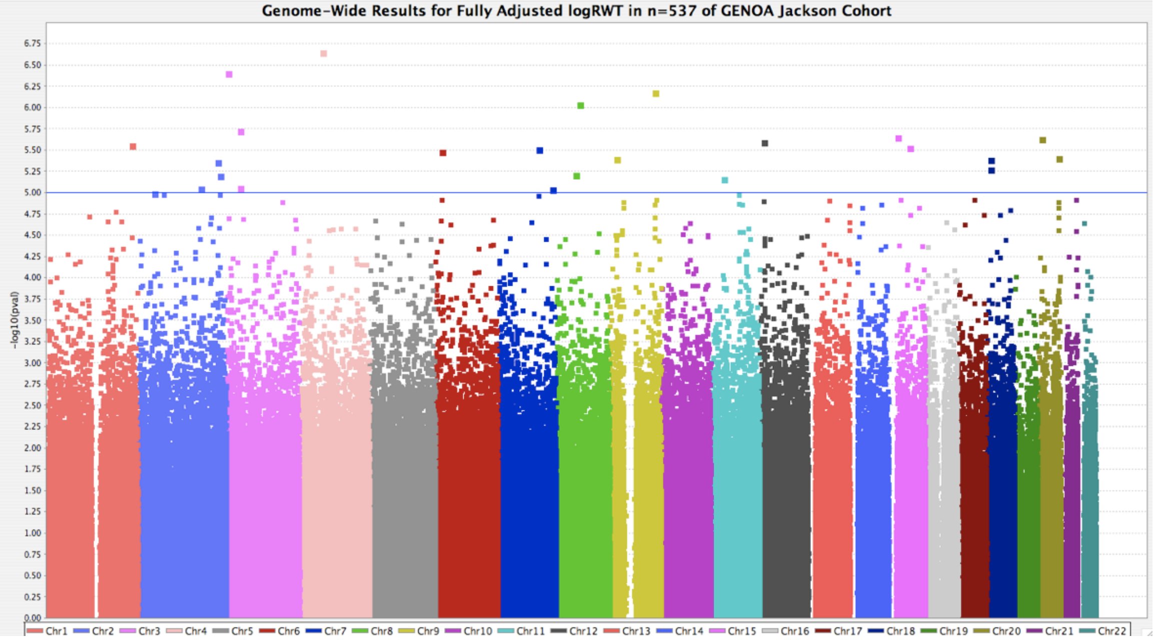


Figure 6.5.b. Distribution of –log p-values from tests of association between Affymetrix 6.0 array SNPs and adjusted RWT in GENOA African-Americans



References

1. National Center for Health Statistics. Health, 2006: with chartbook on trends in the health of Americans. Hyattsville, MD: 2006.

2. Samani NJ, Erdmann J, Hall AS, et al. Genomewide association analysis of coronary artery disease. N Engl J Med. 2007; 357(5):443-453.

3. McPherson R, Pertsemlidis A, Kavaslar N, et al. A common allele on chromosome 9 associated with coronary heart disease. Science. 2007; 316(5830):1488-1491.

4. Helgadottir A, Thorleifsson G, Manolescu A, et al. A common variant on chromosome 9p21 affects the risk of myocardial infarction. Science. 2007; 316(5830):1491-1493.

5. Pearson TA, Manolio TA. How to interpret a genome-wide association study. JAMA. 2008; 299(11):1335-1344.

6. Devereux RB, de Simone G, Ganau A, Roman MJ. Left ventricular hypertrophy and geometric remodeling in hypertension: Stimuli, functional consequences and prognostic implications. J Hypertens Suppl. 1994; 12(10):S117-27.

7. van der Wall EE, van der Laarse A, Pluim BM, Bruschke AVG (eds). Left Ventricular Hypertrophy: Physiology Versus Pathology. Dordrecht, The Netherlands: Kluwer Academic Publishers, 1999.

8. Benjamin EJ, Levy D. Why is left ventricular hypertrophy so predictive of morbidity and mortality? Am J Med Sci. 1999; 317(3):168-175.

9. FBPP Investigators. Multi-center genetic study of hypertension: The family blood pressure program (FBPP). Hypertension. 2002; 39(1):3-9.

10. Chobanian AV, Bakris GL, Black HR, et al. Seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure. Hypertension. 2003; 42(6):1206-1252.

11. Lang RM, Bierig M, Devereux RB, et al. Recommendations for chamber quantification: A report from the american society of echocardiography's guidelines and standards committee and the chamber quantification writing group, developed in conjunction with the european association of echocardiography, a branch of the european society of cardiology. J Am Soc Echocardiogr. 2005; 18(12):1440-1463.

12. Devereux RB, Alonso DR, Lutas EM, et al. Echocardiographic assessment of left ventricular hypertrophy: Comparison to necropsy findings. Am J Cardiol. 1986; 57(6):450-458.

13. Palmieri V, Dahlof B, DeQuattro V, et al. Reliability of echocardiographic assessment of left ventricular structure and function: The PRESERVE study. prospective randomized study evaluating regression of ventricular enlargement. J Am Coll Cardiol. 1999; 34(5):1625-1632.

14. Affymetrix. Affymetrix® Genome-Wide Human SNP Nsp/Sty 6.0 User Guide. 2007; .

15. Cutler DJ, Zwick ME, Carrasquillo MM, et al. High-throughput variation detection and genotyping using microarrays. Genome Res. 2001; 11(11):1913-1925.

16. Rabbee N, Speed TP. A genotype calling algorithm for affymetrix SNP arrays. Bioinformatics. 2006; 22(1):7-12.

17. Affymetrix. Data Sheet: Affymetrix® Genome-Wide Human SNP Array 6.0. 2007; .

18. R Core Development Team. R: A Language and Environment for Statistical Computing. 2007; 2.6.0.

19. Freedman ML, Reich D, Penney KL, et al. Assessing the impact of population stratification on genetic association studies. Nat Genet. 2004; 36(4):388-393.

20. Thomas DC, Witte JS. Point: Population stratification: A problem for case-control studies of candidate-gene associations? Cancer Epidemiol Biomarkers Prev. 2002; 11(6):505-512.

21. Hoggart CJ, Parra EJ, Shriver MD, et al. Control of confounding of genetic associations in stratified populations. Am J Hum Genet. 2003; 72(6):1492-1504.

22. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000; 155(2):945-959.

23. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38(8):904-909.

24. Arnett DK, Hong Y, Bella JN, et al. Sibling correlation of left ventricular mass and geometry in hypertensive african americans and whites: The HyperGEN study. hypertension genetic epidemiology network. Am J Hypertens. 2001; 14(12):1226-1230.

25. Raudenbush SW, Bryk AS. Hierarchical Linear Models: Applications and Data Analysis Methods. 2nd ed. Sage Publications, Inc, 2002.

26. Helgadottir A, Thorleifsson G, Magnusson KP, et al. The same sequence variant on 9p21 associates with myocardial infarction, abdominal aortic aneurysm and

intracranial aneurysm. Nat Genet. 2008; 40(2):217-224.

27. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. Science. 2008; 322(5903):881-888.

28. McCarthy MI, Abecasis GR, Cardon LR, et al. Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. Nat Rev Genet. 2008; 9(5):356-369.

29. NCI-NHGRI Working Group on Replication in Association Studies, Chanock SJ, Manolio T, et al. Replicating genotype-phenotype associations. Nature. 2007; 447(7145):655-660.

30. Kavvoura FK, Ioannidis JP. Methods for meta-analysis in genetic association studies: A review of their potential and pitfalls. Hum Genet. 2008; 123(1):1-14.

31. Chung KC, Burns PB, Kim HM. A practical guide to meta-analysis. J Hand Surg [Am]. 2006; 31(10):1671-1678.

32. Evangelou E, Maraganore DM, Ioannidis JP. Meta-analysis in genome-wide association datasets: Strategies and application in parkinson disease. PLoS ONE. 2007; 2(2):e196.

33. Stram DO. Meta-analysis of published data using a linear mixed-effects model. Biometrics. 1996; 52(2):536-544

34. Hindorff LA, Junkins HA, Mehta JP and Manolio TA. A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/26525384. Accessed May 15, 2009.

Chapter 7

Review of Dissertation

In this dissertation I have systematically investigated the genetic architecture of measures of left ventricular remodeling – LVM and RWT – in an African American cohort from the GENOA study. LVM and RWT are highly predictive of incident heart disease, sudden cardiac death, and all-cause mortality,[1-3] but the reasons these traits are so highly predictive is not well understood.[3] Investigating the genetic contributions to LVM and RWT has the potential to reveal etiologic pathways involved in LV remodeling and heart disease. Furthermore, because of differing allele frequencies and environmental factors in various ethnic populations, discovery of genetic contributions to LV remodeling may also provide insight into disparate incidence of LV remodeling, hypertension, and/or heart disease between African Americans and non-Hispanic white populations.

This dissertation began with a traditional biometric genetic analysis of the heritability of LVM and RWT and an examination of the statistical evidence for genetic correlations between the two traits (Chapter 2). This was followed by a preliminary investigation of the potential single gene effects of 395 SNPs from 80 candidate genes on LVM and RWT variation (Chapter 3). Because of the well-documented concerns for potential confounding of genetic associations by population substructure,[4, 5] Chapter 4 was dedicated to investigating population substructure in the African-American cohort of

GENOA. Subsequent association analyses (Chapters 5 and 6) were adjusted for substructure using principal component analysis. Recognizing that gene-environment and gene-gene interactions are likely to be involved in the genetic architecture of complex traits[6, 7], I then investigated whether 1,878 candidate gene SNPs were associated with interindividual variation in LVM either via main effects, gene-environment interactions, or gene-gene interactions (Chapter 5). Finally, we stepped away from the candidate gene approach and tested approximately 700,000 SNPs from the Affymetrix 6.0 chip for main effects associated with LVM and RWT variation (Chapter 6).

Inferences from Dissertation

Due to the absence of an independent replication cohort for much of this work, inferences about specific variant's association with LVM or RWT are being made with caution. However, this dissertation work has highlighted the complexity of the genetic architecture underlying LVM and RWT in the African-American cohort of GENOA. In Chapter 3 we identified three SNP main effects associated with indexed measures of adjusted LVM and one SNP associated with adjusted RWT. The β_1-adrenergic receptor non-synonymous polymorphism (*ADRB1* Arg389Gly) reported in association with RWT is especially interesting because other studies have shown its association with a gain of contractile function in myocytes and has been implicated in increased risk of heart failure.[8] Furthermore, this variant occurs at higher frequencies in African-American populations and therefore may be partially contributing to increased rates of heart disease in African-American populations.[9] This particular SNP deserves follow-up in replication studies. A lesson learned from this study was the difficulty in replicating genetic effects across populations with different allele frequencies. Of the four SNPs we tried to replicate in the Hispanic cohort of GENOA, two had drastically different allele frequencies, making the comparison inappropriate. While replication is greatly important to genetic epidemiology, the reasons for failed replication deserve careful consideration. Failed replication may not be due to initial false positive reports, but due to other factors such as differing allele frequencies or differences in environmental contexts which influences the effects of a single SNP in a population.

After expanding the number of candidate-gene SNPs investigated and the type of effects by which these SNPs may be associated with LVM, we concluded that gene-gene interactions dominate the genetic architecture of LVM in the African-American cohort of GENOA. This is in line with hypotheses presented by others that epistasis is likely to be ubiquitous in complex traits.[10] Given the difficulties of replicating genetic effects across study populations, this work also highlights the utility of applying numerous multiple-testing criteria to a single study cohort. By considering the intersection of results passing all criteria as "robust", statistical significance is not based solely on the ability to achieve very low p-values but also the ability to predict outcomes in independent samples. Many multiple-testing correction methods have been developed, most with the interest of reducing type I error rates with little concern for type II error. By using multiple criteria that are not prohibitively strict on type I error, the goal was to minimize type II error without significant inflation of type I error.

Recognizing the candidate gene approach is limited by incomplete and imperfect *a priori* knowledge of pathphysiological pathways leading to disease, we also tested approximately 700,000 genome-wide SNPs from the Affymetrix 6.0 array for associations with LVM and RWT. These preliminary investigations will be used in a future meta-analysis with the HyperGen cohort and the analysis will also be redone once the full African-American sample of GENOA has completed genotyping, thereby increasing sample size and power to detect genetic effects. While no firm inferences can be made in this GWAS, we did observe an excess of strong p-values for both traits, even after considering the observed inflation of p-values. Therefore, follow-up studies and meta-analyses are warranted.

Integration

The analyses described above represent an evolution in the field regarding adjustment of LVM, adjustment for population substructure, shifts from candidate gene to genome-wide association studies, and the growing emphasis on reducing type I error in genetic association studies. Tables 7.1 and 7.2 outline the analytical methods used and results reported for the three genetic association Aims of this dissertation. In Aim 1, we conducted a candidate gene association study for main effects associated with either LVM or RWT. LVM was indexed to height raised to a power of 2.7 and was adjusted for age, sex, body mass index (BMI), systolic blood pressure (SBP), diabetes, and stroke volume (another echocardiographically measured trait) to maximally adjust for LVM variation due to concomitants. RWT was adjusted for age, BMI, SBP, and stroke volume. No population substructure adjustment was implemented. In Aim 2, we expanded our candidate gene association study to include interaction effects potentially associated with LVM (we did not investigate RWT). We did not index LVM for this analysis because of a shift in the field's assessment of impact of adjustment and instead adjusted for age, sex, SBP, height and weight. Principal component analysis was performed using microsatellite markers and then the first 20 components were used to adjust analyses for population substructure. Furthermore, Aim 2 applied the most stringent criteria for declaring "significance" of an association because of the vast quantity of tests being conducted, a genetic effect had to pass three pre-determined thresholds from three multiple-testing correction methods. In Aim 3, in order to mirror the analysis strategy of HyperGen, we adjusted LVM for age, age^2, sex, height, weight, waist, and the number of anti-hypertensive medications the individual was taking. Principal component analysis was performed for this Aim using the genome-wide SNPs and the first 30 components were retained for adjustment in association analyses.

These analyses identified a variety of SNPs in numerous genes and the results were not necessarily consistent across Aims. In total, we reported 3 SNPs (from 3 different genes) with main effects associated with LVM, 8 SNPs (from 7 different genes) involved in 4 SNP-SNP interactions associated with LVM, and numerous strong associations from the GWAS for LVM that are currently being validated. Only one SNP was reported with main effects associated with RWT while numerous strong associations from the GWAS are also currently being validated. The evolving analytical methods and seemingly inconsistent results across Aims do not negate any individual results; they only alter the inferences across Aims. One example of this is by considering the covariate adjustment. For Aim 1, we adjusted for stroke volume. Therefore, no genetic effects can be associated with LVM via the pathway of stroke volume. For Aims 2 and 3, no adjustment of stroke volume was made; therefore genetic effects associated with LVM may actually be via the stroke volume pathway. A benefit to adjusting for numerous predictors of LVM is to uncover "direct" genetic effects associated with LVM. However, this approach often limits comparability of results across studies and fails to identify genetic effects with significant contributions towards LVM via risk factor pathways. All adjustment methods used in Aims 1-3 are valid; they only represent different research approaches and changes in the goal of adjustment (e.g. maximally reduce variation in the trait versus create a uniform adjustment for maximum comparability across studies).

An important scientific evolution across Aims was the shift from an emphasis on candidate gene association studies to a more agnostic genome-wide association approach. In addition to the obvious focus of candidate gene studies on SNPs within genes hypothesized to be involved in the disease process, there are numerous reasons to conduct one study versus the other. The candidate gene SNPs chosen for Aims 1 and 2 where identified for genotyping because they were either amino acid substitutions or tagSNPs that could be used to cover both African-American and non-Hispanic White populations. The tagSNPs for the Affymetrix 6.0 genome-wide association chip were chosen to cover non-Hispanic White haplotype block patterns predominately with additional tagSNPs added to better accommodate African-American genomic variation. Furthermore, the candidate gene SNPs were chosen with a lower minor allele frequency (MAF) limit (MAF >0.01) compared to the genome-wide association studies that focus on "common" SNPs with MAF >0.05. Thus the candidate gene approach is poised to identify different aspects of the genetic architecture compared to genome-wide association studies.

One consistency observed across Aims, regardless of genotype selection, analytical methods, or adjustments used is the greater quantity of significant results and the stronger results identified in association with LVM compared to RWT. This is not surprising given that the estimated heritability was higher for LVM compared to RWT, 0.416 versus 0.235. These heritability estimates in the African-American cohort of

GENOA indicate that LVM is likely to have more of its variability explained by genetic variability than RWT. And while genes are likely involved in both traits, our estimates of genetic correlation from Chapter 2 indicated no statistically significant evidence for pleiotropy, which was an unexpected conclusion. However, we feel there still may be shared genetic components between the two traits as evidenced by the GWAS results. For example, in Figure 7.1 all SNPs from the GWAS of 738,451 SNPs in Chapter 6 associated with LVM with a p-value <0.01 (24,721 SNPs) and RWT (23,263 SNPs) are presented. There are 1,523 of the same SNPs associated with both LVM and RWT based on a p-value <0.01. While these SNPs have not yet been replicated, these results exemplify the possibility for pleiotropy.

Future of Genetic Mapping of Complex Traits

In addition to the variety of methods applied in this dissertation, the general field of the genetic mapping of human traits and diseases utilizes a wide variety of sophisticated analytical methods and strategies. Diverse sets of analytic tools have been used for statistical modeling (e.g. logistic and linear regression, data mining methods, variance component analysis), adjustment of population substructure (e.g. *Structure*, genomic control, principal component analysis), and correction for multiple testing (e.g. false discovery rate, false positive report probability, Bonferroni correction). This large tool-kit underscores the complexity of the hypotheses being investigated and the fact that one single analytic method or scientific approach is not uniformly utilized to identify genes underlying complex human traits. There are several reasons for the need to apply a diversity of approaches. Diseases and quantitative traits exhibit a tremendous amount of heterogeneity (by age, sex, body size, geographical region, ethnic background, comorbidities), and it only follows that the underlying genetic architecture would be heterogeneous as well. Therefore, a variety of study designs and analytic methods are needed to study the genetic architecture of disease. Regression techniques may continue to be useful for identifying moderately strong single gene effects and simple gene-environment interactions. However, when testing for gene-gene interactions, the parameter space being estimated using regression methods begins to grow exponentially. In such situations, non-parametric, data mining methods such as multifactor dimensionality reduction (MDR) may be more useful for detecting gene-gene interactions associated with dichotomous outcomes.[12] Methods such as combinatorial partitioning method (CPM)[13] provide an alternative to linear regression and are more conducive for identifying multiple variable genetic models associated with quantitative traits of interest.[12] There are many other methods to be considered as well (e.g. ridge regression, random forests, and other ensemble methods), and even more yet to be developed. The bottom line is that each has unique advantages and disadvantages for detecting different aspects of the genetic architecture of a trait, whether it is multigenic models, interactions, strong effects, weak effects, rare variants, or common variants. Other non-analytical advancements are also needed to enhance the genetic mapping of complex traits including advanced meta-analyses methods, sequencing studies of genes to identify rare variants, improving the methods for measuring environmental factors to improve power for detecting gene-environment interactions, and applying dynamical complex systems approaches that could capture the short term and long term impacts of events over a person's lifecourse.

Meta-analyses of GWAS

A relatively inexpensive, immediate development in genetic mapping of complex diseases involves the unprecedented number of collaborations forming to conduct meta-analyses of GWAS. Single epidemiological studies will often be underpowered to detect small effects, which is problematic because genetic epidemiological studies are searching for small effects. Meta-analysis is a well-developed method for combining evidence from numerous studies. It can be used to increase power to detect effects and provides standards for reporting of results.[14] There has been much discussion of the "winner curse" and inflated effect estimates in genome-wide association studies in this dissertation (see Chapter 5). Meta-analyses improve precision of estimates and increase power to detect associations, reducing the probability of false negatives and minimizing inflation of estimates.[15] As with any meta-analysis, meta-analyses of GWAS should make careful considerations to minimize information bias, selection bias, publication bias, and confounding due to population substructure.[16] A strength of the collaborations being established for GWAS is that the collaborations have been forming early in the discovery process, reducing publication bias concerns. Furthermore, the availability of only a limited number of genotyping platforms lends well to having similar exposure measurement, minimizing information bias. The number of published GWAS meta-analyses is increasing rapidly and numerous review articles have been written that discuss considerations to be made when conducting them.[14-16]

Studies of Rare Genomic Variants

Current genome-wide association studies are using genotyping platforms created with the "common disease common variant" hypothesis in mind (Affymetrix and Illumina platforms), so they are skewed toward representing SNPs with more common frequencies (>5%). It is estimated that only 40% of all SNPs in the genome have a minor allele frequency (MAF) greater than 5%.[17] The majority of variation within the genome is being overlooked by only focusing on "common" SNPs, but the importance of less frequently occurring variants (MAF <5%) is gaining recognition.[17] As described by Altshuler and colleagues, these rare SNPs can be separated into those that are still relatively common (MAF 0.005 – 0.05) and those that are so rare they are essentially unobservable in a study population.[11] While the power is low to detect any single, rare variant in population based studies, power is increased when the large number of these rare SNPs are considered in aggregate.[11, 17] The analytic strategy of considering all rare variants in aggregate not only lends well to achieving high statistical power in population based studies, but also supports the polygenic hypothesis that numerous genes impact a given trait independently. This new emphasis on rare variants is not to say that the "common disease common variant" hypothesis is incorrect, only incomplete. The genes involved in complex traits are likely to exhibit a range of allele frequencies and effect sizes; therefore analysis strategies recognizing this wide range of allele frequencies and sizes of effect will be needed.

Improved Measurement of Exposures and Outcomes

Any epidemiological study depends on the quality and consistency of exposure and outcome measurements. GWAS benefit from standardized, consistent, highly validated exposure measurement from two commonly sourced vendors (Affymetrix and Illumina). The ability and precision with which we measure the genetic exposures has quickly outpaced developments in measuring disease outcomes and environmental exposures. This measurement error in environmental factors is likely impacting the power to detect gene-environment interactions. For example, in this dissertation, Chapter

5 failed to detect any SNP-covariate interactions that passed all three multiple testing criteria, but hundreds of SNP-SNP interactions were identified. Because of the hypothesized importance of gene-environment interactions in the genetic architecture of complex traits, this lack of SNP-covariate interactions was surprising and some of this may be due to imprecise measurement of the covariates. Improved methods for measuring environment and outcomes are needed to keep pace with the precision with which genotypes are measured if we are to truly understand the etiology of these diseases.[11]

Complex Systems Approach

Repeatedly within this dissertation, I have brought attention to the fact that complex, quantitative traits have a complex genetic architecture involving genetic variants (with varying frequencies) acting independently, via interactions (both gene-gene and gene-environment), and with varying magnitudes of effect (weak to strong). It is well accepted that polygenic models, interactions, and heterogeneity are defining features in the genetic architecture of complex traits. Despite this, genetic epidemiological studies continue to apply statistical methods developed to uncover much simpler etiologies.[18] This reductionist approach fails to recognize the complexity of biology. Use of statistical methods that more closely model the complexity of biology may prove more fruitful.

A simple example of how this can be implemented is explained by Kraft and colleagues.[19] Kraft argues that if we know environmental effects may modify genetic effects, than our power to detect any genetic effect (main or interaction) is maximized by considering the joint test of marginal genetic effects and gene-environment interactions.[19] This is a simple implementation that remains within the context of regression methods, therefore it is still dependent on the same analytic tool-box designed to detect simpler genetic effects. More complex methods such as classification and regression trees (CART), random forests (RF), pattern recognition techniques such as neural networks, and Bayesian belief networks are alternative methods that may prove more appropriate for modeling the biological complexity and context dependency by which genes have effects on measurable outcomes.[18, 20] In addition, new bioinformatics capabilities to incorporate pathway information into genome-wide analyses, as well as to incorporate epigenomic, transcriptomic, and metabolomic data into an investigation of the complex dynamical systems underlying genetic architecture of complex diseases.

Future Directions for GENOA

Collaborations with HyperGen

While the individual studies comprising this dissertation have been completed, investigation of the genetic architecture of LVM and RWT in the African-American cohort of GENOA is far from complete. There are many opportunities for future work. The immediate future holds great promise as we develop collaborations with the HyperGen investigators. First, HyperGen will provide a truly independent test cohort to estimate the predictive ability of the gene-gene interactions identified in Chapter 5. I used cross-validation within the GENOA cohort to estimate the predictive ability of four SNP-SNP interactions. Despite the use of internal training and testing sets, using the same cohort for discovery and subsequent predictive ability estimates is subject to inflated estimates and the "winners curse".[21] HyperGen would be an ideal independent study cohort to estimate the increased amount of variation explained in traits by the SNP-SNP interactions because of its similar study design, measured covariates, and similar genotype data. Second, as mentioned at length in Chapter 6, HyperGen and GENOA are planning a meta-analysis for the GWAS of LVM and RWT. This meta-analysis will increase our power to detect numerous, small effect size variations that are hypothesized to dominate quantitative traits such as LVM and RWT. Lastly, we are working with HyperGen to build a larger collaborative involving numerous cohorts of African-American populations with echocardiography data. Within this larger collaborative, the goal is to have identified robust areas of association through genome-wide association study meta-analyses that can be followed up with resequencing studies for rare variants.

Population Substructure

Genetic mapping in African-American populations is a task complicated by increased haplotypic diversity and varying degrees of admixture.[22] Therefore, future directions for this work involve delving deeper into explorations of population substructure by admixture. The adjustment methods reviewed in this dissertation (*Structure*, genomic control, Eigenstrat) are all variations of a global adjustment for substructure.[22] That is to say they adjust for genomic averages of admixture proportions, instead of the actual ancestry estimates immediately adjacent to a SNP of interest. Even after adjusting for genome-wide estimated averages of ancestry, region specific estimates will vary.[23] Numerous methods to estimate regional and local ancestry have been developed, however, they do not scale well to large, genome-wide datasets.[24] Recently a method named LAMP (Local Ancestry in adMixed Populations) was developed that can be implemented in genome-wide datasets with large numbers of subjects.[23, 24] LAMP estimates the ancestry in sliding windows of contiguous SNPs. After the optimal window length is chosen, then for each SNP the ancestry is estimated based on a majority vote from each window containing that SNP.[24] In addition to the computational efficiency of LAMP, other advantages include the fact that parental population genotypes are not required (although they can be used) and the results provide “much more accurate estimates than methods such as *Structure* or Eigenstrat", which should in turn reduce bias resulting from population substructure that has not been taken into account.[24] Given the inflation of results we observed in Chapter 6 even after conducting principal component analysis to adjust for population substructure, conducting association tests after implementing a more precise method such as LAMP is warranted.

Conclusion

Genetic association studies are currently undergoing a revolution both statistical and molecular methodologies used to detect genetic variations with a wide range of effects and allele frequencies. The future of genetic association studies to unravel the complex genetic architecture of LVM and RWT is similarly exciting. Reframing the way we approach analysis, working towards incorporating the known etiological and pathophysiological complexity inherent in LVM and RWT, may prove to be a promising direction for future works in GENOA. In addition, replication in other African-American cohorts will be important to reveal interesting insights about differences in these traits across ethnic groups and could lead to better methods of predicting and intervening on the progression of heart disease. Genetic mapping of complex traits in African-American populations is surprisingly still in its infancy. The dissertation work presented here, and the future opportunities within the GENOA dataset, will continue to contribute to the growing field of genetic epidemiology and offer a better understanding of sub-clinical phenotypes for heart disease in African-American populations.

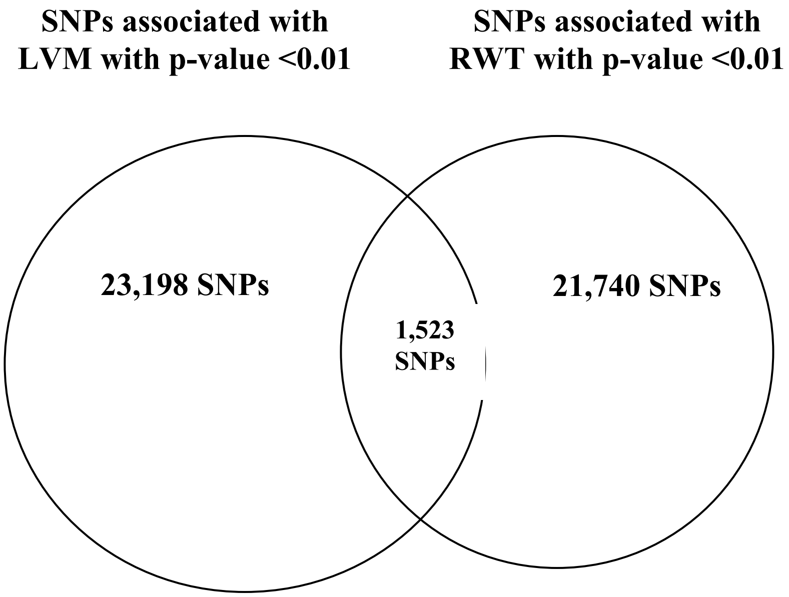
Table 7.1. Summary of analytic methods used for testing genetic associations with LVM for Aims 1, 2, and 3.

Aim	General Approach	Outcome	Sample Size	Covariate Adjustment	Population Substructure Adjustment	Multiple Testing Adjustment	Reported Results
1	Main effects of 385 candidate-gene SNPs	logLVMI ($g/m^{2.7}$)	n1=439, n2=439	age, sex, BMI, SBP, diabetes, and stroke volume	None	1) internal replication, 2) replication in Hispanics of GENOA	rs449647 (*APOE*) cv356952 (*SCN7A*) cv9546580 (*SLC20A1*)
2	Main, GxE, or GxG effects of 1,878 candidate-gene SNPs	logLVM (g)	N=1,328 (n1=477, n2=477)	age, sex, SBP, height, and weight	20 principal components based on microsatellites	1) FDR<0.3, 2) CV R^2 >0.005 3) internal replication	**1)** rs35314437(MPO)*rs7552841 (*PCSK9*), **2)** rs257376(*PRKAR2B*) *rs5267 (*NPPC*), **3)** rs17876148(*PON2*)* rs12971616 (*CARM1*), **4)** rs6745660(*HSPD1*)* rs12460421 (*CARM1*)
3	GWAS for main effects (Affy 6.0 SNPs)	logLVM (g)	n=537	age, age^2, sex, height, weight, waist, diabetes, and # of anti-hypertensive meds	30 principal components based on Affy SNPs	None (future external replication planned)	**Top 5 SNPs:** rs12102921, rs6594004 (*PLEKHA6*), rs1996003 (*PLEKHA6*), rs1956500, rs10510440 (*HACL1*)

Table 7.2. Summary of analytic methods used for testing genetic associations with RWT for Aims 1 and 3.

Aim	General Approach	Outcome	Sample Size	Covariate Adjustment	Population Substructure Adjustment	Multiple Testing Adjustment	Results
1	Main effects of 385 Candidate gene SNPs	logRWT	n1=439, n2=439	age, BMI, SBP, stroke volume	None	1) internal replication, then, 2) replication in Hispanics of GENOA	Arg389Gly (*ADRB1*)
2	n/a	n/a	n/a	n/a	n/a	n/a	
3	GWAS for main effects (Affy 6.0 SNPs)	logRWT	n=537	age, age^2, sex, height, weight, waist, diabetes, # of anti-hypertensive meds	30 principal components based on Affy SNPs	None (future external replication planned)	**Top 5 SNPs:** rs1350003, rs308717, rs7034992 (*RBM18*), rs16929424 (*NKAIN3*), rs17050778

Figure 7.1. Venn diagram depicting the number of SNPs from the genome-wide association study associated with adjusted LVM with a p-value <0.01 (24,721), the number of SNPs associated with adjusted RWT with a p-value <0.01 (23,263), and the number associated with both LVM and RWT (1,523).


SNPs associated with LVM with p-value <0.01
SNPs associated with RWT with p-value <0.01
23,198 SNPs
1,523 SNPs
21,740 SNPs

References

1. Levy D, Garrison RJ, Savage DD, Kannel WB, Castelli WP. Prognostic implications of echocardiographically determined left ventricular mass in the framingham heart study. N Engl J Med. 1990; 322(22):1561-1566.

2. Koren MJ, Devereux RB, Casale PN, Savage DD, Laragh JH. Relation of left ventricular mass and geometry to morbidity and mortality in uncomplicated essential hypertension. Ann Intern Med. 1991; 114(5):345-352.

3. Benjamin EJ, Levy D. Why is left ventricular hypertrophy so predictive of morbidity and mortality? Am J Med Sci. 1999; 317(3):168-175.

4. Hoggart CJ, Parra EJ, Shriver MD, et al. Control of confounding of genetic associations in stratified populations. Am J Hum Genet. 2003; 72(6):1492-1504.

5. Thomas DC, Witte JS. Point: Population stratification: A problem for case-control studies of candidate-gene associations? Cancer Epidemiol Biomarkers Prev. 2002; 11(6):505-512.

6. Kardia SL. Context-dependent genetic effects in hypertension. Curr Hypertens Rep. 2000; 2(1):32-38.

7. Sing CF, Stengard JH, Kardia SL. Genes, environment, and cardiovascular disease. Arterioscler Thromb Vasc Biol. 2003; 23(7):1190-1196.

8. Small KM, Wagoner LE, Levin AM, Kardia SL, Liggett SB. Synergistic polymorphisms of beta1- and alpha2C-adrenergic receptors and the risk of congestive heart failure. N Engl J Med. 2002; 347(15):1135-1142.

9. Moore JD, Mason DA, Green SA, Hsu J, Liggett SB. Racial differences in the frequencies of cardiac beta(1)-adrenergic receptor polymorphisms: Analysis of c145A>G and c1165G>C. Hum Mutat. 1999; 14(3):271.

10. Moore JH. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Hum Hered. 2003; 56(1-3):73-82.

11. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. Science. 2008; 322(5903):881-888.

12. Moore JH. Analysis of gene-gene interactions. Curr Protoc Hum Genet. 2008; Chapter 1:Unit 1.14.

13. Nelson MR, Kardia SL, Ferrell RE, Sing CF. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation.

Genome Res. 2001; 11(3):458-470.

14. Zintzaras E, Lau J. Trends in meta-analysis of genetic association studies. J Hum Genet. 2008; 53(1):1-9.

15. Evangelou E, Maraganore DM, Ioannidis JP. Meta-analysis in genome-wide association datasets: Strategies and application in parkinson disease. PLoS ONE. 2007; 2(2):e196.

16. Kavvoura FK, Ioannidis JP. Methods for meta-analysis in genetic association studies: A review of their potential and pitfalls. Hum Genet. 2008; 123(1):1-14.

17. Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI. Shifting paradigm of association studies: Value of rare single-nucleotide polymorphisms. Am J Hum Genet. 2008; 82(1):100-112.

18. Thornton-Wells TA, Moore JH, Haines JL. Genetics, statistics and human disease: Analytical retooling for complexity. Trends Genet. 2004; 20(12):640-647.

19. Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ. Exploiting gene-environment interaction to detect genetic associations. Hum Hered. 2007; 63(2):111-119.

20. Motsinger AA, Ritchie MD, Reif DM. Novel methods for detecting epistasis in pharmacogenomics studies. Pharmacogenomics. 2007; 8(9):1229-1241.

21. Kraft P. Curses--winner's and otherwise--in genetic epidemiology. Epidemiology. 2008; 19(5):649-51; discussion 657-8.

22. McCarthy MI, Abecasis GR, Cardon LR, et al. Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. Nat Rev Genet. 2008; 9(5):356-369.

23. Redden DT, Divers J, Vaughan LK, et al. Regional admixture mapping and structured association testing: Conceptual unification and an extensible general linear model. PLoS Genet. 2006; 2(8):e137.

24. Sankararaman S, Sridhar S, Kimmel G, Halperin E. Estimating local ancestry in admixed populations. Am J Hum Genet. 2008; 82(2):290-303