Characterization of Gene-by-Age Interaction and Gene-by-Gene Interaction In Coronary Artery Disease

by

Yi Zhao

Department of Computational Biology and Bioinformatics
Duke University

Date: July 5, 2012
Approved:

_____
Elizabeth Hauser, Supervisor

_____
William E. Kraus

_____
Sayan Mukherjee

_____
Simon Gregory

_____

Thesis submitted in partial fulfillment of
the requirements for the degree of Master of Science in the Department of
Computational Biology and Bioinformatics in the Graduate School
of Duke University

2012

ABSTRACT

Characterization of Gene-by-Age Interaction and Gene-by-Gene Interaction In Coronary Artery Disease

by

Yi Zhao

Department of Computational Biology and Bioinformatics
Duke University

Date: July 5, 2012
Approved:

_____
Elizabeth Hauser, Supervisor

_____
William E. Kraus

_____
Sayan Mukherjee

_____
Simon Gregory

_____

An abstract of a thesis submitted in partial
fulfillment of the requirements for the degree
of Master of Science in the Department of
Computational Biology and Bioinformatics in the Graduate School
of Duke University

2012

Abstract

The success of genome-wide association studies (GWAS) has been limited by missing heritability and lack of biological relevance of identified variants. We sought to address these issues by characterizing interaction among genotypes and environment using case-control samples enrolled at Duke University Medical Center. First, we studied the impact of age on coronary artery disease (CAD). Gene-by-age (GxAGE) interactions were tested at genome-wide scale, along with genes' marginal effects in age-stratified groups. Based on the interaction model, age plays the role as a modifier of the age-CAD relationship. SNPs associated with CAD in both young and old demonstrate consistency in effect sizes and directions. In spite of these SNPs, vastly different CAD associated genes were discovered across age and race groups, suggesting age-dependent mechanisms of CAD onset. Second, we explored gene-by-gene interaction (GxG) using a statistical model and compared results to biological evidence. Specifically, we investigated GATA2 as a candidate gene transcription factor, and modeled the interaction with genome-wide SNPs. The genetic effects at interacting loci were modified by GATA2 genotype. Without taking GATA2 variants into account , no marginal main effects were detected. Open access ChIP-seq data was available for comparison with the statistical model, and to relate GWAS findings with biological mechanisms. The agreement between the statistical and biological models was very

limited.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to start by thanking my program Computational Biology and Bioinformatics for recruiting me and giving me the opportunity to start this exciting and unforgettable journey. Especially I want to thank my advisor Elizabeth Hauser, not only for her excellent being an instructor, but also her kindness and support of my life. I would also like to my thankfulness to Jeanette McCarthy, Cavin Ward-Caviness, Abanish Singh, Rong Jiang, Beverly Brummet, students, faculty and staffs at CHG. Finally, I want to express my sincere gratitude to my parents, my boyfriend, my friends Su Wang, Eric Jiang, Zhu Wang, Xiaoyuan Gu, friends from Duke and ZJU.

1. Genome-wide scan for age-related variants in Coronary Artery Disease

1.1 Introduction

Coronary artery disease (CAD) is one of the leading causes of morbidity and mortality in western societies, and the prevalence is rapidly rising in developing countries (J Mackay 2004). It is not only a disease of old men, but also affects women, young adults and even children.

The development of CAD is a complex interplay between genetics and environmental risk factors. Among the over 300 risk factors associated with CAD, aging is the most powerful non-modifiable risk factor(J Mackay 2004).An increasing trend of CAD prevalence associated with age has been observed (Roger et al. 2012). Advancing age correlates with an increase in conventional risk factors such as blood pressure and physical inactivity, as well as impairment of coronary artery systems. Younger adults may predispose less to these conventional risk factors but may be affected by greater heritability (Fornage et al. 2004; Murabito et al. 2005; Scheuner et al. 2006; Hauser et al. 2004). Thus, despite a relatively small fraction of cases, early-onset subjects form an interesting group to study the underlying disease genetics. The impact of age on CAD development can be studied in patients stratified by age of onset, i.e. the age of 55 years old has been widely used to stratify young affected and old affected (Zhang et al. 2010). Comparison of genetic variants contributing to younger age at onset with genetic

variants contributing to older age of onset might elucidate age-specific disease mechanisms. In addition, the heterogeneity of genetic effects on different age groups might be explained by GxAGE interactions. This is within the spectrum of gene by environment interaction. Genetic effects might be masked if uncontrolled for environmental exposures that modify the gene-disease relationship. We might have incorrect estimates of disease variance contributed by genes and environments if we only estimate the separate effects of genes and environmental factors, and overlook their interactions. Therefore, the search for gene by environment interaction has been shown to produce a meaningful increase in the combined effect of the genetic determinants (Lanktree and Hegele 2009; Hunter 2005; Manolio et al. 2006).

Linkage studies and genome-wide association studies have identified genetic variations associated with age-related CAD and traditional risk factors. A genome-wide scan in 493 affected sibling pairs has discovered multiple regions linked to early-onset CAD(Hauser et al. 2004). A genome-wide linkage analysis of 4,175 affected subjects from 1,933 families has mapped both early-onset CAD and MI to chromosome 2p12-2p23.3(Samani et al. 2005). In genome-wide association studies of CAD associated lipid profiles, the proportion of lipid variance explained by associated genes seems to drop in the older cohorts(Aulchenko et al. 2009). Within the same subjects, a longitudinal study has demonstrated that genetic effects on intraindividual variation in CAD risk factors

change over time, which is likely a reflection of gene by environment interaction(Friedlander et al. 1997).

These data strongly indicate the importance of GxAGE interactions that predispose to CAD. To characterize age's effects on CAD incidence, we performed a genome-wide association study in age stratified cases and controls. We also studied GxAGE interaction in a combined sample. These two analyses might jointly and compensatorily identify genetic factors and pathogenic pathways. Assuming patients with collective genetic variants began to have symptoms at younger age, the identification of these variants has the potential to genetically and pathologically distinguish early-onset CAD patients from others.

1.2 Methods

1.2.1 CATHGEN sample

CATHGEN is a cohort study which sequentially enrolled subjects undergoing cardiac catheterization at Duke University Medical Center from 2001 to 2011(Shah et al. 2010). Cases and controls were defined based on the CAD index (CADi) and disease history. With the knowledge that CAD risk increases with age, we set different CADi threshold to determine affected status in young and old samples (Connelly et al. 2006; Hauser et al. 2004). For participants younger than 55 years of age, CADi > 23 was used to define young affected (YA). For older participants, a higher cutoff CADi > 67 was used

to define old affected (OA), as older people have a higher CADi baseline. Both younger affected and older affected used a shared control group: participants older than 61 years of age, whose CADi ≤ 23 and without history of cerebrovascular disease, peripheral vascular disease, MI, ICC, or CABG. The use of older normal (ON) as controls reduced the potential that unaffected young participants develop CAD at an older age.

To study GxAGE interaction, cases and controls were also selected regardless of age by a CADi threshold of 23. From which cases were defined as all affected (AA), and controls as all normal (AN). AN were also restricted by the same criteria as ON. The characteristics of cases and controls are summarized in Table 1. Case-control groups were constructed based on affection status and age. Three comparison groups were defined: YA as cases and ON as controls, OA as cases and ON as controls, AA as cases and AN as controls to avoid biases in the genetic analysis. Due to the potential for population stratification, we performed analysis stratified by white and black for a total of 6 comparisons.

1.2.2 SNP genotyping and quality control

We genotyped 1,140,419 SNPs using the Illumina HumanOmin1-Quad array. After filtering for SNPs with 0 intensity, 1,016,423 SNPs were left. Next, we checked X, Y and MT chromosomes and excluded SNPs based on call frequencies. 1,283 snps with genotype 0 in all samples were removed on the Y chromosome. Three SNPs with call

frequency < 0.98 and one bad SNP were removed from the mitochondrial genome. Sequential clustering of females and males on the X chromosome has been done before SNP statistics were calculated. We then removed 1058 SNPs with low call rate < 0.98, 21,623 SNPs with AB frequency > 0 in males, and we zeroed 20 SNPs with > 2% male Y chromosome heterozygosity. Next, we checked all SNPs on the autosomal chromosomes and excluded 5113 SNPs with call frequency < 0.98 and call frequency > 0. After that, we filtered and zeroed SNPs based on cluster separation and heterozygosity excess, and 1,005,840 SNPs were kept for statistical analysis. We then applied HWE p-value < 1.0E-06 and 928735 SNPs remained on chromosome 1 to chromosome 23.

We also performed sample level quality control for the 3427 subjects with GWAS data. Based on low call rate (< 0.975), 25 samples were removed. One sample with heart transplant PT was excluded. Eight samples were removed because of bad chips. We also removed 50 samples because their reported sex disagreed with sex determined by X chromosome data. Next, we calculated cryptic relatedness. A set of 126,661 LD-pruned SNPs was selected based on SNP pairs in a 500 SNP window with a step size of 5 SNPs and r^2 threshold of 0.2. We calculated pair wise identity-by-decent measures and removed one sample from a pair if IBD sharing > 0.2. Of the 94 pairs of samples with relationships, 9 individuals have relationships with two individuals. Thus 85 samples were excluded at this step. Ethnicity outliers were checked by EIGENSTRAT (Price et al.

2006) in the remaining 3258 individuals, and discovered: 104 outliers out of 2305 in whites, 11 outliers out of 678 in blacks, 3 outliers 158 in Native Americans, 28 outliers of 117 in others.

The subsequent QC filters were applied to each case-control group. Samples were excluded if they have a low call rate (< 97%). Marker exclusion were made by genotyping rates less than 0.97 and minor allele frequencies less than 0.05. In dominant models, markers were removed if any of the cells has a sample count less than 2. All QC steps were done in PLINK(Purcell et al. 2007)

1.2.3 Statistical analysis

We tested and controlled for population stratification by using EIGENSTRAT in whites and blacks. We used PLINK(Purcell et al. 2007) to retrieve a list of LD-pruned SNPs by separately calculating LD between SNPs pairs in 50-SNP window size, with a step size of 20 SNPs and an r^2 threshold of 0.10. SNPs with reversed minor alleles were also removed. EIGENSTRAT removed 47 individuals as outliers. Next, race-stratified Eigenstrat analysis was performed. In whites 104 outliers were removed out of 2305; and in blacks 11 outliers were removed of 678. Based on eigenvalues, we picked the first four principal components in whites and the first two principal components in blacks.

Our analysis plan was aimed at detecting SNPs whose effects were modified by age. Association analysis was performed using two separate logistic regression models.

The first model tested the main effect for each SNP in the age-defined two case-control groups (YA+ON, OA+ON) in whites and blacks separately. Age, sex, cardiovascular risk factors (BMI, history of hypertension, diabetes, hyperlipidemia and smoking) and population stratification were included as covariates. Both additive and dominant genetic models were applied. The second model tested GxAGE interaction in a single case-control group (AA+AN) in whites and blacks separately by adding a genotype-by-age cross product to the first model. Because we were specifically interested in GxAGE interaction, a one degree of freedom test was conducted for the genotype-by-age cross product. An additive genetic model was used for tests.

SNPs were selected for secondary analysis based on significance level. An intersection of SNPs showing nominal significance ($p < 0.01$) of main effect in both YA+ON and OA+ON were selected for comparison of odds ratios in these two groups. SNPs with main effect $p < 1.0E-03$ in both YA+ON and OA+ON were annotated and assessed for functional relatedness. SNPs significantly associated in one group ($p < 1.0E-04$) but not in the other ($p > 0.05$) were analyzed for heterogeneity of effects under different ages. In the interaction model, we followed up the most significantly associated SNPs ($p < 1.0E-04$) and compared with results from the age-stratified tests.

1.3 Results

1.3.1 Clinical characteristics of CATHGEN cases and controls

The CATHGEN dataset is an appealing resource to study the influence of age on CAD development in a case-control design. It consists of subjects aged 18-91, from which we can define case groups as young affected (YA), old affected (OA) and all affected (AA). Table 1 presents the comparison between cases and controls, comparison among case groups, and comparison between whites and blacks. Cases demonstrate a disease-prone risk factor profile. Compared with controls, cases have higher male sex, increased rates of history of hypertension, diabetes, hyperlipidemia, smoking and an increased family history of coronary disease. In addition, cases exhibit a more unfavorable lipid profile: lower HDL level and higher triglycerides level. However, the LDL and total cholesterol levels in controls are higher than one or two case groups, which may be explained by a higher use of cholesterol lowering medication of cases. Older affected exhibit the lowest total cholesterol level among all groups. Younger affected manifest higher prevalence of clinical history risk factors except for history of CABG. It is consistent with the clinical evidence that younger patients are more likely to receive alternative treatments rather than surgery. The rate of family history of coronary disease is higher in younger patients, indicating potentially higher heritability. Compared with older affected, younger affected also demonstrate a high risk lipid

profile including higher levels of total cholesterol, LDL, triglycerides and lower HDL level. These results support the assumption that stronger genetic factors maybe observed in younger affected (Fornage et al. 2004; Murabito et al. 2005; Scheuner et al. 2006; Hauser et al. 2004). The between-race comparison shows a generally higher BMI level in blacks. Moreover, blacks also had higher blood pressure, higher lipoprotein levels, but lower triglycerides.

Table 1: Clinical characteristics of CATHGEN cases and controls, stratified by race and age.

	White					Black				
	Young Affected n=506	Old Affected n=114	Old Normal n=312	All Affected n=1181	All Normal n=781	Young Affected n=152	Old Affected n=14	Old Normal n=69	All Affected n=278	All Normal n=293
Age (SD)	55.2 (9.7)	69.1 (8.2)	69.4 (6.4)	63.2 (11.3)	57.7 (11.7)	50.8 (7.5)	70.7 (8.1)	68.2 (6.4)	58.4 (11.1)	53.6 (10.9)
Age of onset (SD)	47.1 (6.7)	64.2(7.4)	NA	57.7 (11.7)	NA	46.7 (6.0)	67.2 (8.2)	NA	54.9 (11.1)	NA
%Male	71.3%	68.4%	41.0%	65.5%	52.8%	56.6%	35.7%	37.7%	52.9%	43.3%
CAD index	52.6 (19.0)	80.8 (11.1)	6.1 (9.4)	52.1 (18.7)	3.7 (7.9)	50.7 (19.1)	73.3 (12.6)	6.0 (9.5)	50.9 (17.7)	323.9%
History of Hypertension	69.0%	64.0%	61.9%	70.2%	55.3%	83.6%	92.9%	76.8%	84.2%	75.1%
History of Diabetes	36.4%	27.2%	13.1%	32.2%	15.0%	53.3%	50.0%	42.0%	53.6%	35.2%
BMI	30.4 (7.1)	29.1 (6.9)	28.6 (6.9)	29.7 (6.6)	29.9 (7.8)	32.0 (7.0)	31.1 (8.4)	32.8 (8.2)	31.3 (7.0)	32.9 (8.9)
History of Hyperlipidemia	74.5%	69.3%	42.6%	70.4%	41.5%	67.8%	78.6%	47.8%	67.3%	39.6%
History of Smoking	65.4%	51.8%	34.9%	55.0%	40.3%	57.9%	50.0%	42.0%	52.5%	36.9%
Family History of Coronary Disease	53.2%	40.4%	24.4%	44.8%	31.0%	48.0%	28.6%	14.5%	38.1%	28.3%
History of Cerebrovascular Disease	8.7%	7.0%	0.0%	9.7%	0.0%	3.9%	7.1%	0.0%	8.3%	0.0%
History of Peripheral Vascular dz	9.9%	9.6%	0.0%	9.5%	0.0%	10.5%	0.0%	0.0%	11.5%	0.0%
History of MI	48.2%	30.7%	0.0%	38.1%	0.0%	50.7%	21.4%	0.0%	43.2%	0.0%
History of ICC	37.4%	16.7%	0.0%	29.0%	0.0%	38.2%	21.4%	0.0%	29.9%	0.0%
History of CABG	45.5%	50.0%	0.0%	36.7%	0.0%	28.3%	35.7%	0.0%	26.3%	0.0%
Mean Systolic Blood Pressure, mm HG (SD)	141.0 (22.8)	145.4 (24.7)	148.8 (21.2)	145.0 (24.7)	141.4 (21.7)	150.8 (28.5)	164.3 (22.5)	156.4 (26.9)	154.2 (27.7)	147.6 (25.7)
Mean Diastolic Blood Pressure, mm HG (SD)	79.5 (14.1)	78.4 (14.5)	78.3 (13.6)	79.1 (14.5)	78.4 (13.6)	85.8 (16.5)	85.1 (13.0)	79.8 (15.1)	85.1 (16.3)	84.4 (15.7)
Total Cholesterol, mg/dL (SD)	190.2 (51)	175.8 (47.4)	187.2 (43.1)	182.4 (50.4)	191.2 (40.9)	199.0 (83.1)	187.5 (25.6)	190.0 (33.6)	189.6 (69.0)	201.2 (53.7)
LDL, mg/dL (SD)	101.2 (40.5)	95.5 (32.8)	107.2 (34.9)	98.8 (41.7)	109.6 (33.6)	112.6 (51.7)	98.0 (35.8)	108.1 (39.6)	107.9 (47.1)	113.5 (40.7)
HDL, mg/dL (SD)	42.5 (12.7)	44.9 (13.5)	50.7 (19.4)	43.9 (14.0)	48.4 (16.4)	45.9 (13.7)	62.5 (7.1)	55.6 (15.6)	48.0 (14.6)	53.5 (18.0)
Triglycerides, mg/dL (SD)	248.0 (263.9)	180.2 (117.2)	144.9 (81.4)	205.9 (198.3)	168.0 (123.1)	172.6 (198.3)	112.3 (78.2)	123.3 (79.4)	156.6 (162.8)	149.6 (121.5)

1.3.2 Comparison of SNPs associated with disease status in young and old

Manhattan plots and quantile-quantile (Q-Q) plots for comparisons of all case-control groups are shown in Figure 1 and Figure 2. Each case-control group was analyzed for main effects using both additive and dominant genetic models. Population stratification was controlled in whites and blacks, and no inflation has been observed in

all groups. However, the Q-Q plots in blacks stray below the theoretical line, indicating that small sample sizes result in loss of power. The black OA+ON group has only 14 cases, thus is remarkably underpowered to detect even larger effects. A genome-wide significance p-value cutoff 5.38E-08 is defined by a conservative Bonferroni correction which corrects for 928735 comparisons. No genome-wide significant association was found. To characterize the results and identify SNPs for follow-up research, we evaluated SNPs which reached a liberal threshold of $p < 1.0E-04$.

Based on Manhattan plots, additive and dominant models generate vastly different association results in genome-wide SNPs. A comparison of top SNPs ($p < 1.0E-04$) between additive and dominant models shows varying levels of overlap in different case-control groups. In whites, 30 SNPs were found in 45 (66.7%) additive YA+ON SNPs intersecting 66 (45.5%) dominant counterpart; 56 SNPs were found in 76 (73.7%) additive OA+ON SNPs intersecting 88 (63.6%) dominant SNPs. In blacks, 12 SNPs present in 37 (32.4%) additive YA+ON SNPs and 44 (27.2%) dominant SNPs. The use of dominant model may be able to compensate for the small sample sizes in some groups. However, the underlying disease model is still unclear. In spite of the debates over the selection of disease models, we mainly focused on the additive model as chosen by the majority of published GWAS.

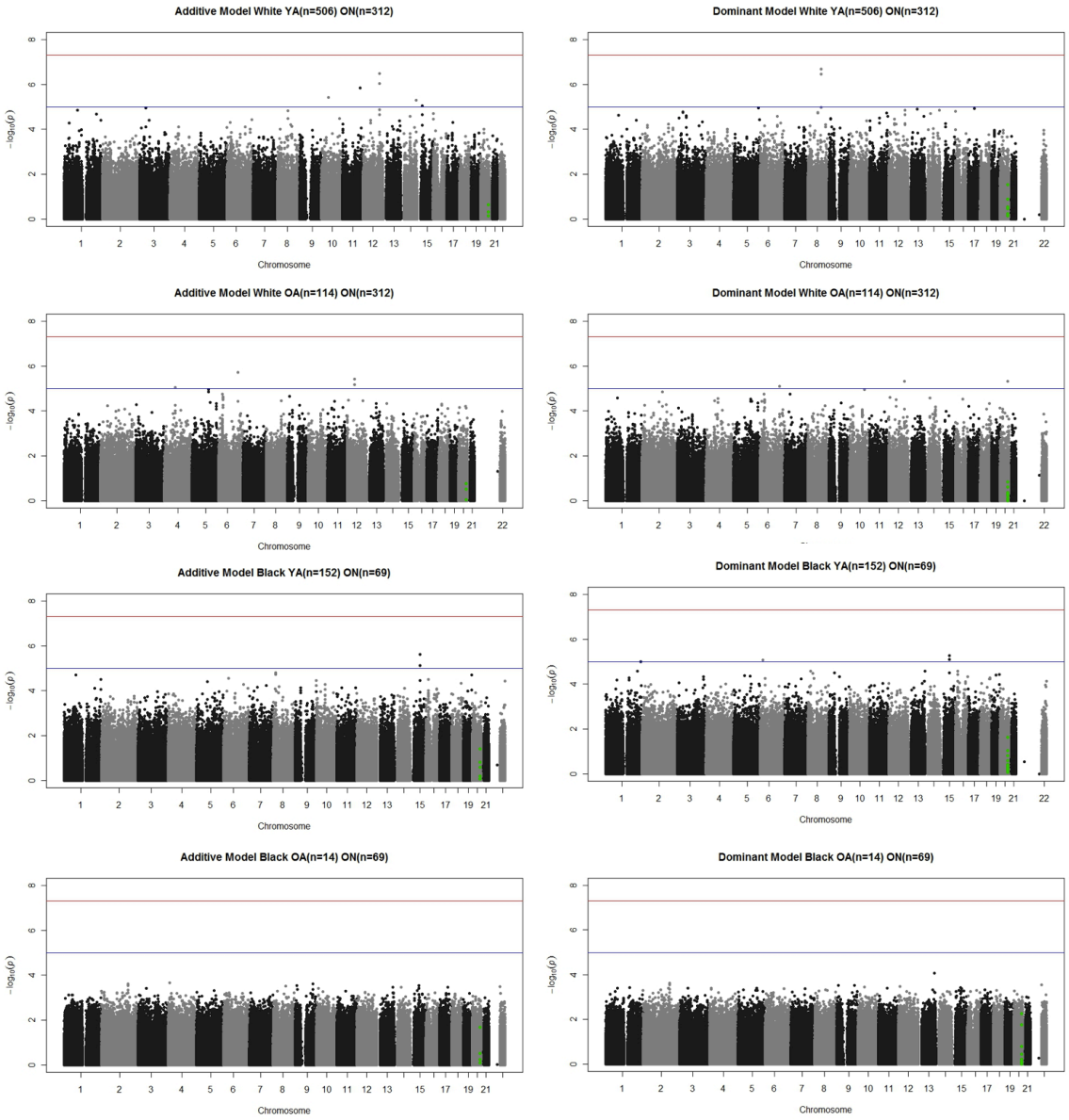

Figure 1: Manhattan plots of CATHGEN GWAS p-values.

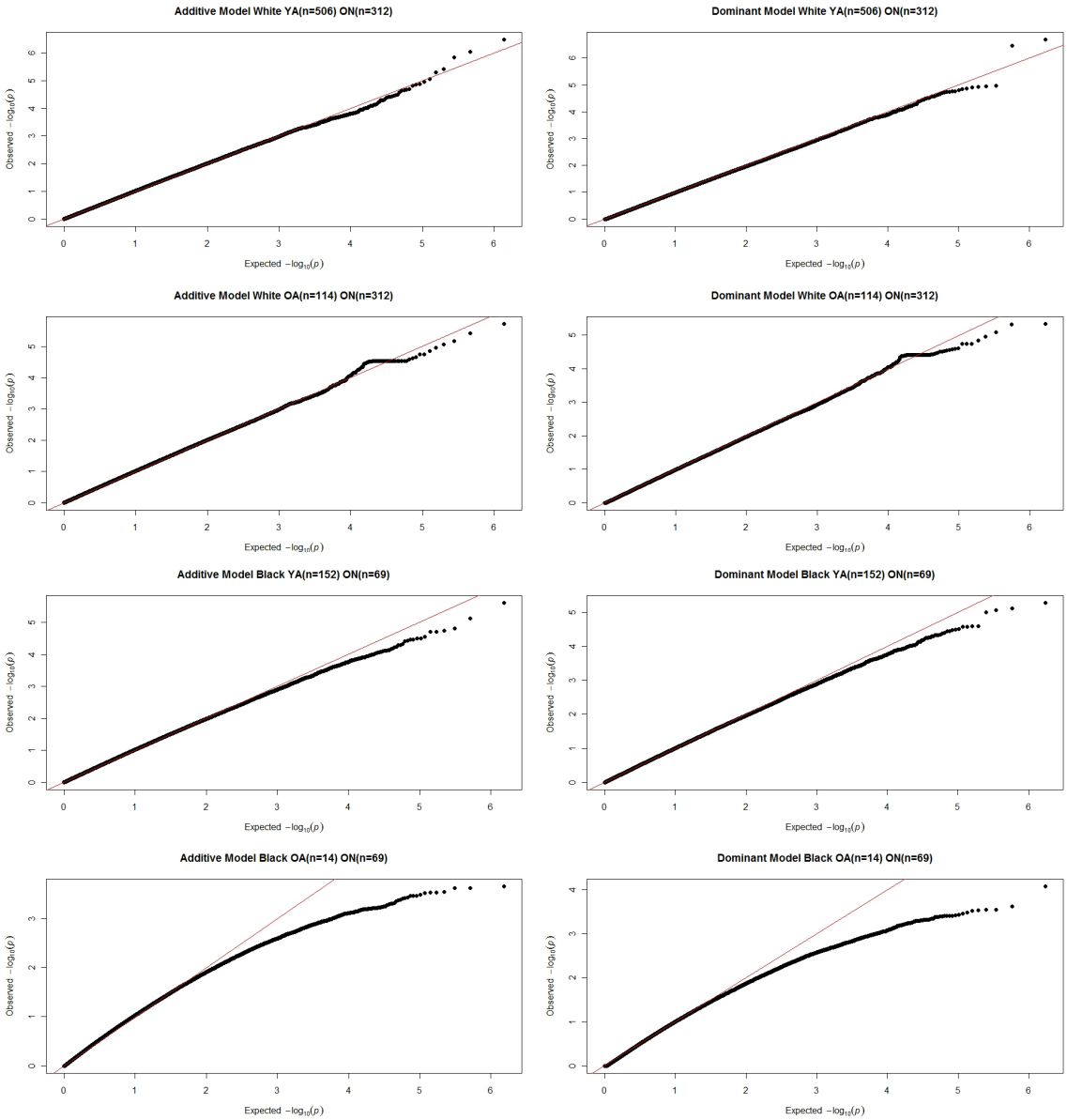

Figure 2: Quantile-Quantile plots of CATHGEN GWAS p-values.

Genetic factors underlying CAD can manifest either as age-dependent or age-independent effects. SNPs that significantly associate in both YA+ON and OA+ON are

candidates in analysis of age-independent genetic effects. The number of independent loci discovered in theadditive model ranges from 0 to 38: 38 in white YA+ON, 33 in white OA+ON, 26 in black YA+ON, 0 in black OA+ON. Only one locus presents in both white YA+ON and white OA+ON, and it consists of two SNPs (rs4762060 and rs9669529) in linkage disequilibrium (LD). This locus is in the intron of KRT80 on chromosome 12q13.13. Its minor allele exhibits protective effects (OR < 1) in both groups. KRT80 is not identified as an obvious CAD susceptibility gene. Further investigation is needed to discover its role in CAD development.

SNPs significantly associated in both young and old are unlikely to be involved in GxAGE interaction. On the other hand, most SNPs discovered by p-value cutoff of 1.0E-04 are not significantly associated SNPs in the opposite age group, thus, define an age-specific SNP as showing significant ($p < 1.0E-04$) CAD association in one group, but not the other ($p > 0.05$). These SNPs are candidates for the study of GxAGE interaction. Twenty-three out of 38 (60.5%) independent loci in white YA+ON are discovered as age-specific, as are 20 out of 33 (60.6%) independent loci in white OA+ON, and 19 out of 26 (73.1%) in black YA+ON. The lack of overlapping loci in young and old groups, together with the large proportion of age-specific loci, highlights the potential for heterogeneity of genetic effects in different age groups.

The use of a less stringent cutoff of $p < 1.0E-03$ yields more SNPs. White YA+ON and white OA+ON shared 59 SNPs, representing 9 independent loci. One of these loci on chromosome 6p22.1 consists of 37 SNPs in LD, and in proximity to LOC100133214, HCG9 and RNF39. Comparison across racial groups shows lack of consistency. Only one SNP rs2235641 on chromosome 16p13.3 presents in YA+ON group in white and black. It is located in the intron of IFT140.

To further characterize the relationship between SNPs associated with early-onset CAD and old-onset CAD, we relaxed the p-value threshold to $p < 0.01$ and plotted odds ratio (OR) for SNPs present in both YA+ON and OA+ON (Figure 3). A vertical line and a horizontal line of OR = 1 were added to differentiate between risk and protective effects. Next a 2 x 2 contingency tables of SNP counts in each OR classification was used to examine association between YA+ON and OA+ON. The p-values from the Fisher's exact test are all zeros, suggesting the dependence of comparison groups. Indeed, linear trends were found. The odds ratios are highly correlated: $r = 0.978$ in additive white YA+ON and white OA+ON, and $r = 0.967$ in the dominant counterpart; $r = 0.895$ in additive black YA+ON and white OA+ON, and $r = 0.874$ in the dominant counterpart.

A majority of SNPs concentrate on the main diagonal and exhibit consistent direction of effects in YA+ON and OA+ON, except for one SNP rs11764226 on 7p12.2, showing opposite effects in dominant YA+ON (OR < 1) and dominant OA+ON (OR > 1).

SNPs with same direction of effects can be divided into risk SNPs (OR > 1), and protective SNPs (OR < 1). In general, the number of risk SNPs exceeds the number of protective SNPs: 274 risk SNPs and 269 protective SNPs in white additive model, 179 risk SNPs and 117 protective SNPs in black additive model, 333 risk SNPs and 291 protective SNPs in white dominant model, 147 additive SNPs and 97 protective SNPs in black dominant model.

However, such strong linear correlation of OR is not observed in the between-race comparison. SNPs with $p < 0.01$ in both whites and blacks were selected to compare OR in two groups (Figure 4). The numbers of SNPs in each plot range from 49 to 79. Compared with plots in Figure3, in which the numbers of SNPs range from 244 to 625, significantly fewer SNPs were shared across races. SNPs in Figure 4 were divided into four sections by a vertical line and a horizontal line of OR = 1. SNPs counts in four sections were analyzed in 2x2 tables using Fisher's exact test, and result shows independence between whites and blacks. Fisher's exact test p-values vary in four comparison groups: $p = 1$ in additive white YA+ON and black YA+ON, $p = 0.018$ in additive white OA+ON and black OA+ON, $p = 1$ in dominant white YA +ON and black YA+ON, $p = 0.568$ in dominant white OA+ON and black OA+ON. In addition, correlation of OR in whites and blacks is not found, given the absolute values of coefficient r from 0.009 to 0.261.

The vast difference between cross-age comparison and cross-race comparison suggests greater discordance between races than between young and old. Based on the linear relationship in SNPs' effects on early-onset and old-onset CAD, this set of CAD-associated SNPs is unlikely to contribute to the age difference of CAD development. Alternatively, SNPs with suggestive association in only one group might account for the age-dependent CAD pathogenesis.

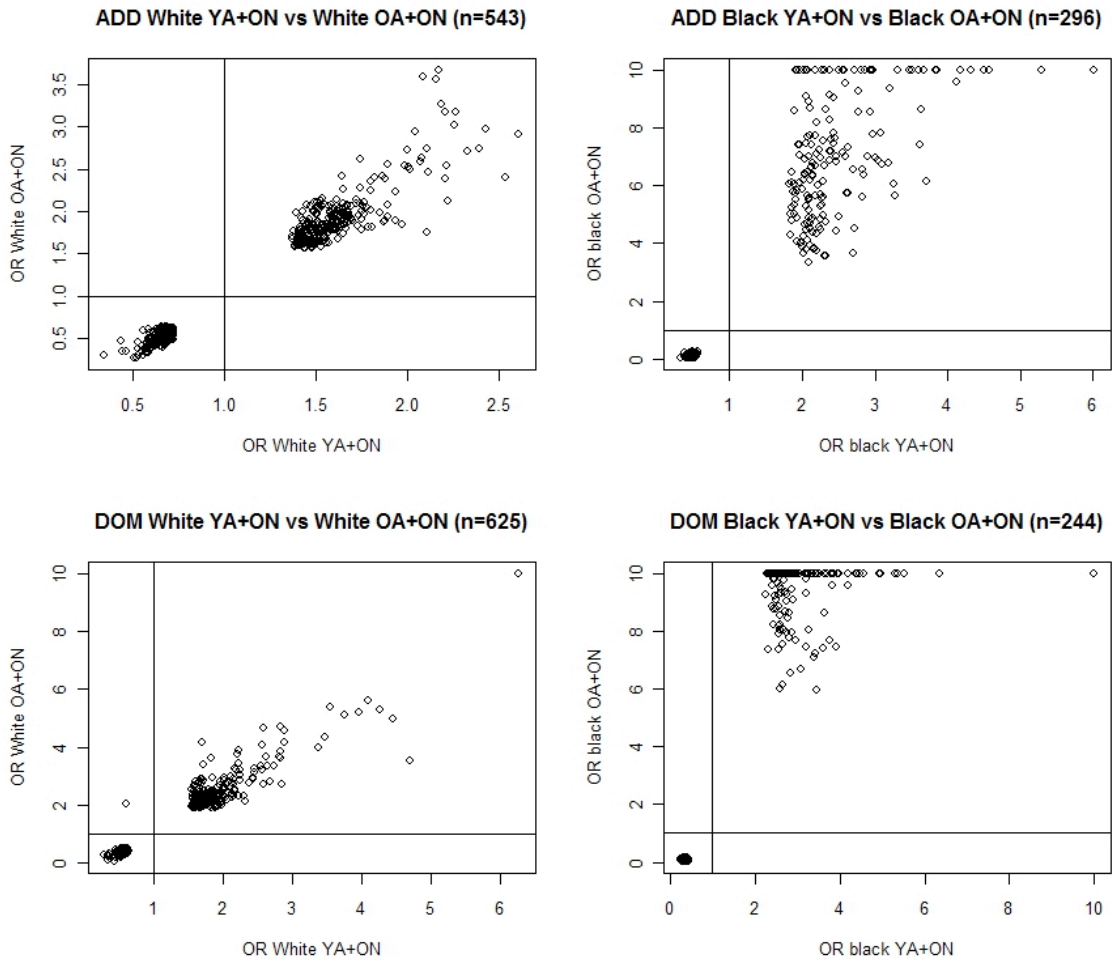

Figure 3: Comparison of odds ratios of YA+ON and OA+ON. For each SNP with suggestive ($p < 0.01$) association with CAD in YA+ON and OA+ON, OR in two groups were plotted. ORs above 10 were set to 10. Comparisons were made in four scenarios: additive and dominant disease models, in whites and blacks. A vertical line and a horizontal line of OR = 1 were added to each plot. SNPs locates in lower left and upper right sections have consistent protective and risk effects in two comparison groups. All except for one SNP (rs11764226) cluster in these two sections.

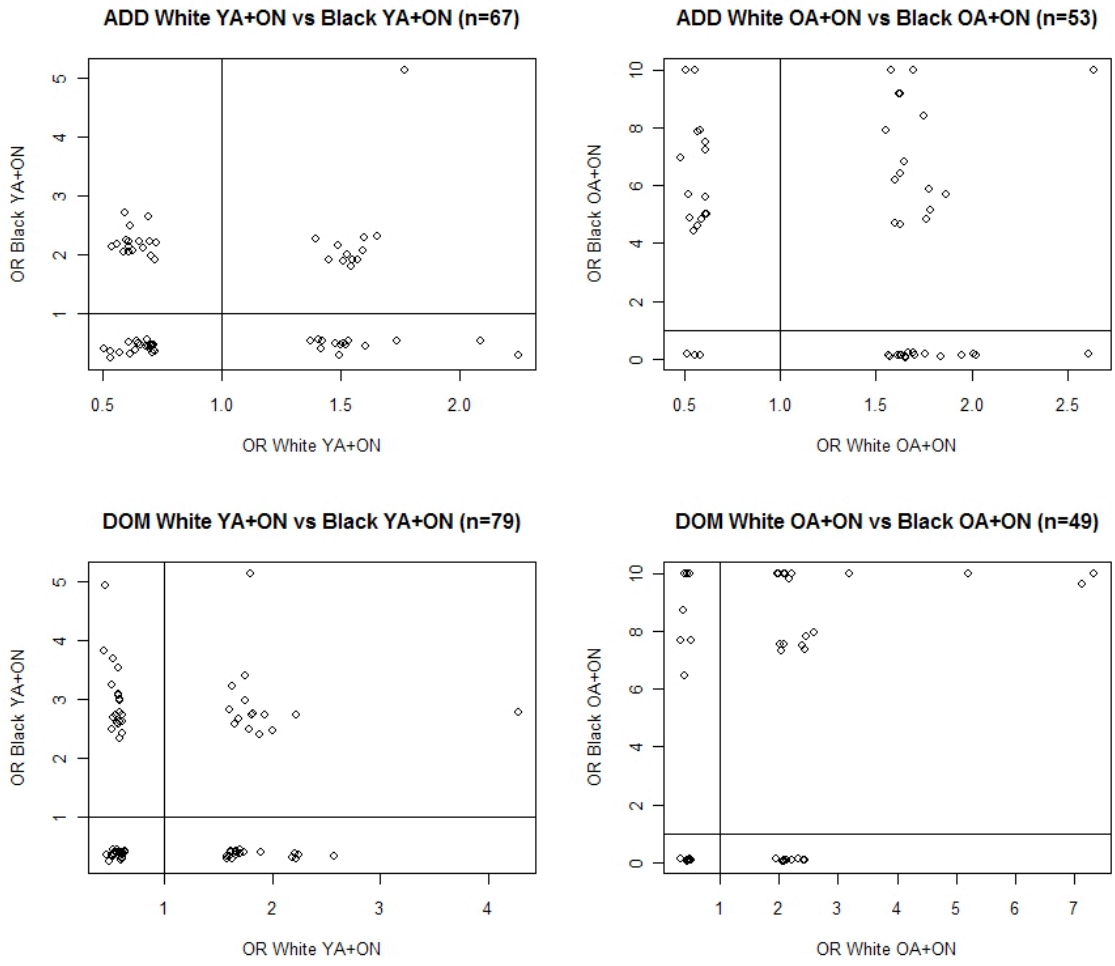

Figure 4: Comparison of odds ratios of whites and blacks. For each SNP with suggestive ($p < 0.01$) association with CAD in whites and blacks, OR in two groups were plotted. ORs above 10 were set to 10. Comparisons were made in four scenarios: additive and dominant genetic models, in YA+ON and OA+ON. A vertical line and a horizontal line of OR = 1 were added to each plot. SNPs counts in each section were used to test association in whites and blacks, and no associations were found.

1.3.3 Comparison of GxAGE interactions and main effect in age-stratified samples

Gene-by-age interaction was calculated for genome-wide SNPs in case-control samples without age stratification. Age at catheterization was used in the logistic regression model instead of using age of onset. Despite the fact that age at catheterization is not a precise measure of CAD onset, it matches the age at CADi measurement and is available for all participants. Manhattan plots and Q-Q plots for GxAGE interaction are shown in Figure 5. From Q-Q plots, p-values deviate from expected values at the lower p-value side. This deviation suggests the insufficiency of power in genome-wide analysis of GxE interaction. No SNPs reached a genome-wide significance cutoff. The most significant interactions are 1.18E-06 in whites and 4.02E-06 in blacks. If we applied a less stringent cutoff of $p < 1.0E-04$, and we discovered 39 independent loci in whites and 48 independent loci in blacks.

To characterize and classify these GxAGE interactions, we analyzed the age-related effects of interaction SNPs with $p < 1.0E-04$. Odds ratios and p-values were extracted from the YA+ON and OA+ON stratified analysis, and were plotted in Figure 6. The use of OR = 1 in two groups divides SNPs into four categories as in Figure 3 and Figure 4. SNP counts were tested for association between YA+ON and OA+ON. Fisher's exact tests show significant association in whites ($p = 0.0026$) and no association in blacks ($p = 0.761$), but the plotted ORs are not convincing without significant main

effects in the stratified analysis. Consequently, we considered not only the direction of OR, but also the significance level of the stratified analysis into classification. In Figure 6, green dots represent significant association ($p < 0.05$) in both YA+ON and OA+ON; blue dots represent absence of association ($p > 0.05$) in both groups; and red dots represent significant association ($p < 0.05$) for only one of the comparisons. The majority of interacting SNPs don't present association in either age group: 71.8% in whites and 79.2% in blacks. In this case, modeling GxAGE interaction is required to detect such an effect. Red dots demonstrate a quantitative difference in effect sizes in two comparison groups. In whites, one out of 7 SNPs discovered this way exhibit YA+ON biased association, whereas three out of nine SNPs in blacks associate in YA+ON only. Green dots indicate that even though association is present in both groups, the relationship between genotype and CAD might change at different age.

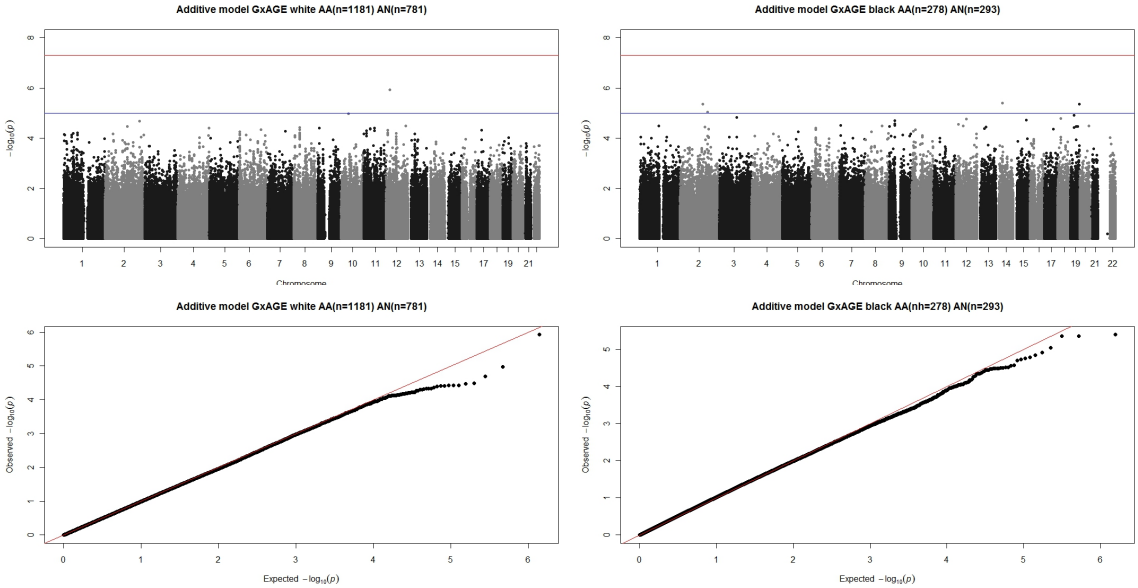

Figure 5: Manhattan plots and Quantile-Quantile plots of GxAGE interaction

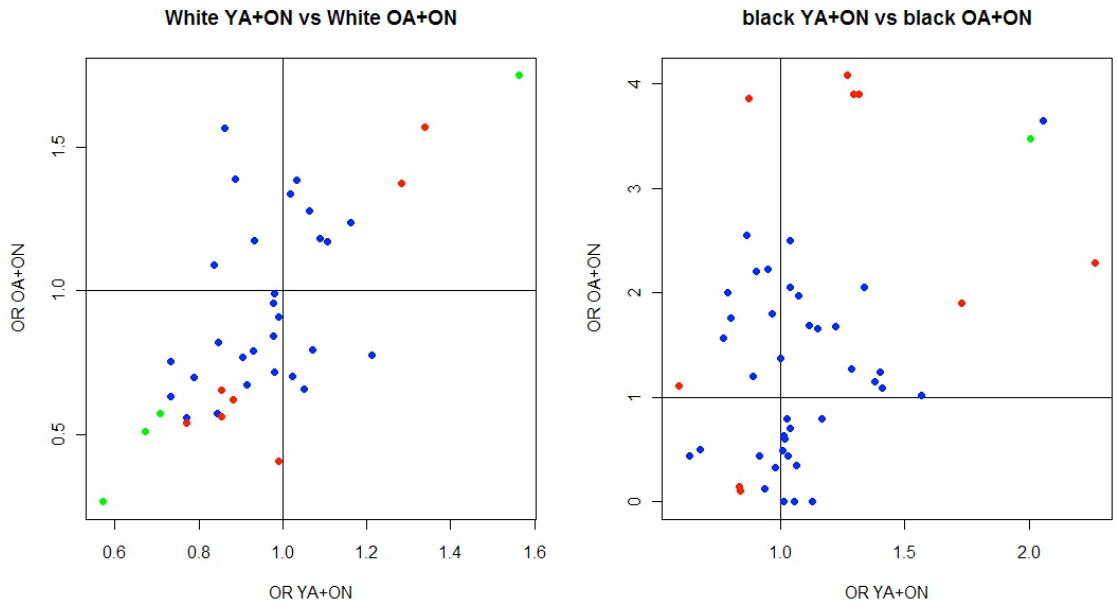

Figure 6: Comparison of odds ratios of YA+ON and OA+ON, with significant GxAGE interaction.

1.3.4 Types of genotype-modified relationship between age and CAD, and disease-modified relationship between age and genotype

Age is marginally associated with CAD in most studies. An increasing trend of CAD risk in relation with age has often been observed. With the presence of genotype, however, the relationship between age and CAD might be modified. In addition, the trends of age-genotype relationship might disagree in cases and controls. SNPs showing significant GxAGE interaction form an appealing set to be followed up in analysis of these relationships. Despite the fact that these SNPs show significant GxAGE interaction, it is anticipated that variation exists. The three categories of significant interaction

colored in Figure 6 have the potential to exhibit more divergence if we studied those relationships. Therefore we extracted the first SNP based on chromosomal coordinate order from each category in whites, stratified by genotype, and plot the relationship between age and the probability of CAD from the fitted logistic regression model (Figure 7). For blue dots, we selected the second SNP rs11211611 instead of the first SNP rs4648808 because rs4648808 shows suggestive association in OA+ON ($p = 0.076$) which might obscure the age-CAD relationship for this particular type of interaction. Association test results of these three SNPs are shown in Table 2. In Figure 7, GxAGE interactions were depicted by the cross of three age-CAD lines. Three types of interactions were differentiated if we concentrated at the lower and upper end of the age range, i.e. < 45 years of age, and > 70 years of age. rs2753244's genotypes separate the age-CAD relationship in both young and old, and consequently we observed significant association in both YA+ON and OA+ON. In contrast, rs11211611 failed to separate the age-CAD relationship, and rs9439550 only separated in the old. rs11211611 represents the blue dots which account for the majority of interactions detect by modeling GxAGE cross product. In all three cases, the CAD risk lines cross around the age of 60. Next we compared the age-genotype relationship between cases and controls (Figure 8). In general, cases are older than controls. The trends of age-genotype relationship are opposite in cases and controls.

Table 2: Three SNPs showing significant gene-by-age interaction.

			GxAGE					YA+ON					OA+ON				
CHR	SNP	BP	OR	P	Major/ Minor	MAF	HWE	OR	P	Major/ Minor	MAF	HWE	OR	P	Major/ Minor	MAF	HWE
6	RS2753244	5257610	0.9681	6.66E-05	A/G	0.1764	0.3511	0.6721	0.008425	A/G	0.1877	0.3582	0.5101	0.004917	A/G	0.2054	0.4582
1	RS11211611	48288016	1.0287	7.36E-05	C/A	0.3713	0.7351	1.163	0.2336	C/A	0.3612	0.9397	1.238	0.2336	C/A	0.3592	0.8335
1	RS9439550	5357171	0.9733	7.39E-05	A/C	0.3419	0.3158	0.8819	0.3214	A/C	0.3637	0.651	0.6204	0.00893	A/C	0.3603	0.1147

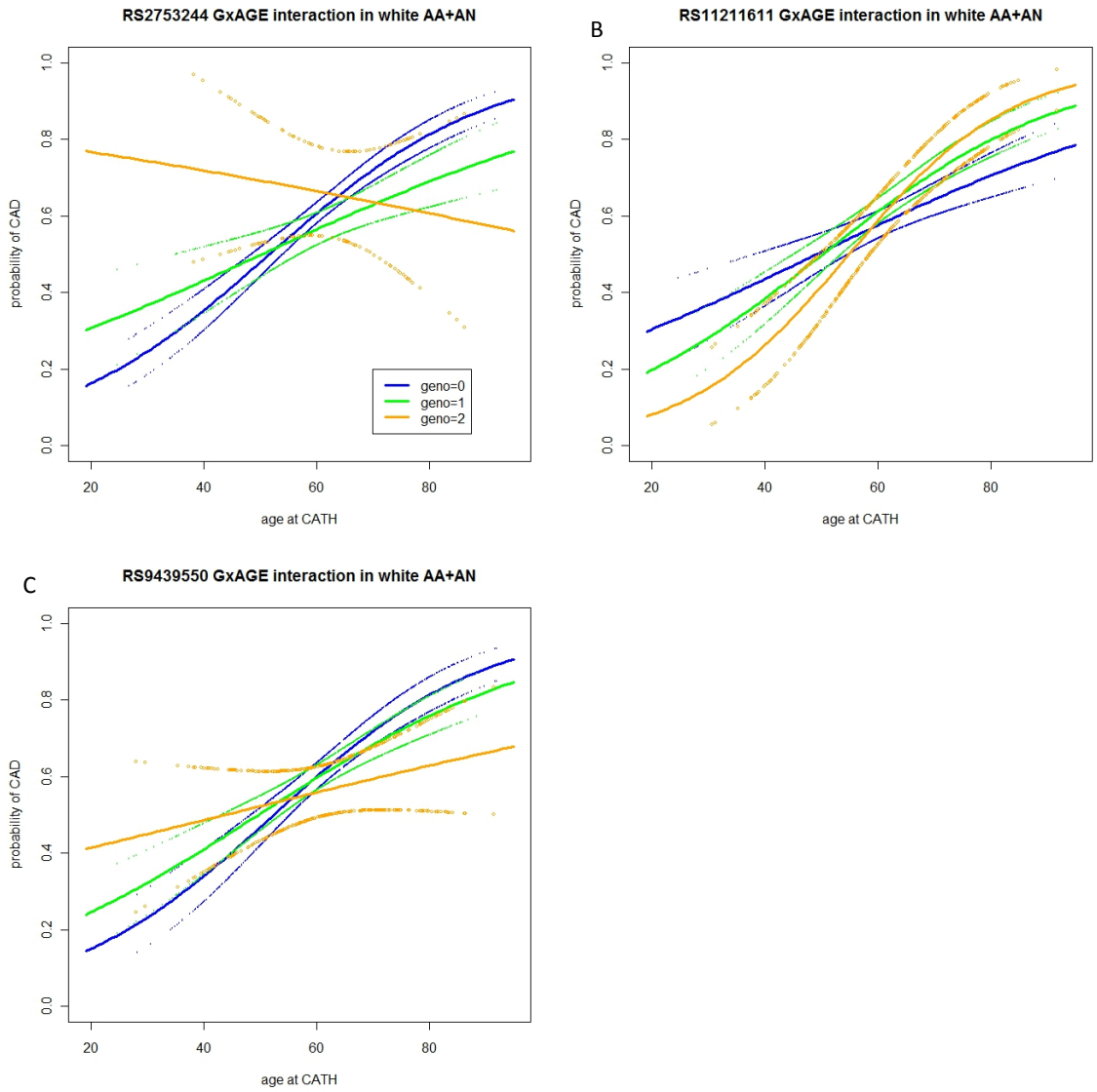

Figure 7: Genotype-modified relationship between age and CAD

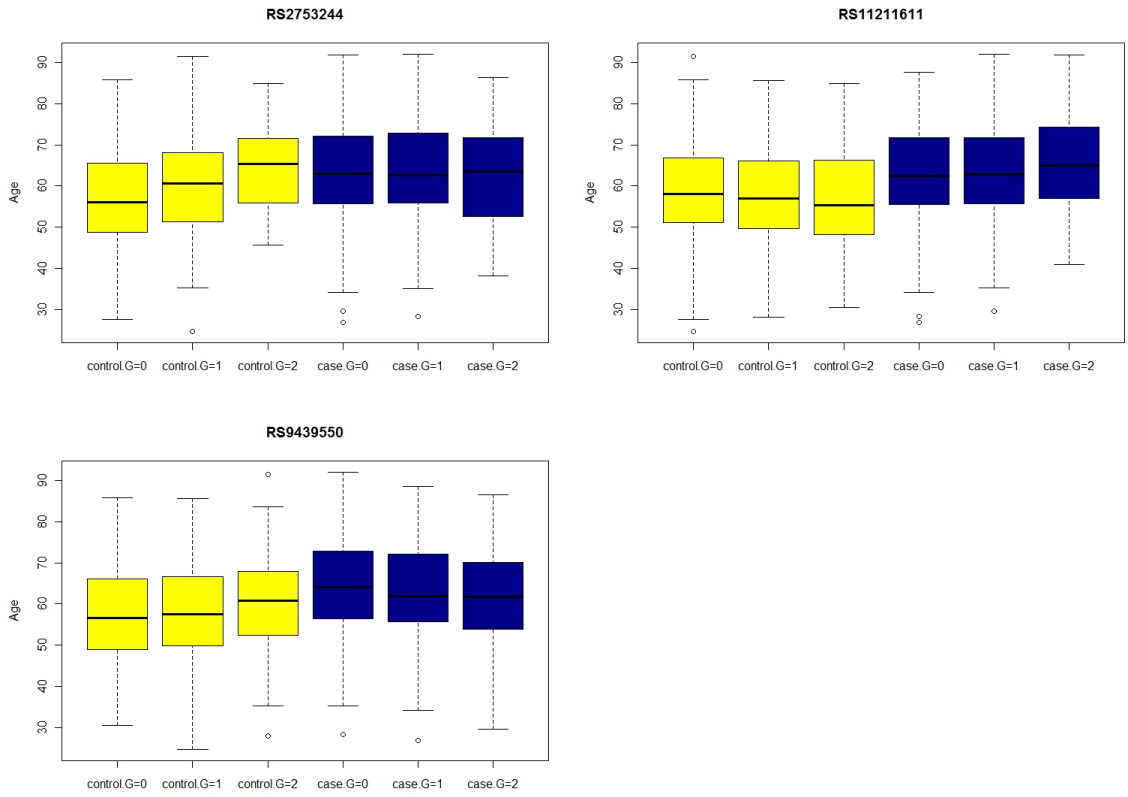

Figure 8: Age-genotype comparisons in cases and controls.

1.4 Discussion

We carried out genome-wide association studies to characterize age-related variants associated with CAD. Modeling GxAGE interaction has discovered loci with age-modified relationship with disease. Most of these loci were not detected via marginal effect association in age-stratified samples. Through our comparison between

young and old, consistent effects were found in SNPs associated in both. Although age might not be as strong an environmental factor and predictor as other covariates such as gender and smoking, our analysis broadly characterizes its effects and relationship with CAD.

Q-Q plots haven shown varying levels of deviation from the theoretical line in most groups. The departure at the right tail might result from the conservative nature of the Wald statistics implemented in PLINK (Hauck et al. 1977). The score test is less conservative and thus a good alternative for association test in future analysis.

A recent meta-analysis consisting of 14 GWAS discovered 13 novel loci associated with CAD (Schunkert et al. 2011). Together with 12 previously published loci, 25 susceptibility loci have been discovered. They further carried out sub-group analysis in young and old, limiting to these main effect loci. They found consistently significant associations in both subgroups, as well as higher odds ratios for early onset than late onset for most of the loci. Our study directly genotyped eleven of these loci and found two loci with $p<.05$. . rs11206510 at 1p32.3 (intergenic region of BSND and PCSK9) shows nominal association in white YA+ON with $p = 0.021$. rs4977574 at 9p21.3 (intron of CDKN2BAS) demonstrates suggestive association in white ($p = 0.031$ in YA+ON, $p = 0.057$ in OA+ON) under a dominant disease model. None of these SNPs show significant interaction in our GxAGE analysis. This finding is consistent with our top interaction

loci in which very few are significantly associated in age-stratified groups. The meta-analysis was not aimed at identification of the difference between age subgroups. Genome-wide detection of age-dependent variants might yield comparable findings as in our studies. Furthermore, the published meta-analysis defined young and old based on the age cutoff of 50 y whereas we used the age of 55 y. In addition, the definitions of cases and controls were not consistent with ours. Thus difference in subgroup definitions might lead to failure in replication. Three out of 11 SNPs show flipped major and minor alleles in whites and blacks. Great disparity of MAF was only observed between racial groups, not within the same race. Consequently, inconsistency of association results were found across races, including different intensities of signals and opposite directions of effects. Furthermore, between-race heterogeneity might demonstrate diversity of genetic ancestries which might result in different LD patterns.

In addition to the 25 SNPs, rs9349379 (intron region of PHACTR1) has been discovered to associate with CAD in European-ancestry and South Asian Ancestry (TCADG Consortium 2011). In our studies, it is nominally associated with CAD in white OA+ON, with $p = 0.043$ and OR = 1.417. Notably only a few SNPs discovered by previous studies are replicable in our analysis. Our limited sample sizes are the major restriction to detect small genetic effects. Small sample sizes have led to oscillation of MAF across case-control groups. However, little evidence of ancestry-specific

associations has been discovered in the GWAS in Europeans and South Asians (Consortium 2011), suggesting the use of combined analyses of different racial groups may result in increased power. Thus, future studies of multi-ethnic groups might overcome the limit of sample size and detect novel variants. Other than sample size, various forms of between-study heterogeneity, like disease ascertainment and age range, might contribute to failure to replicate (Lasky-Su et al. 2008).

We hypothesized that genetic influences on CAD development are age-related. Age itself is an important risk factor of CAD. An increasing trend of CAD prevalence is associated with aging. Moreover, sex-differentiation of CAD incidence grows with age(Roger et al. 2012). We incorporated age into our study design, as we selected cases and controls based on age and CADi (Zhang et al. 2010).

In age-stratified analysis, age was used as categorical variable to decide samples' disease as YA, OA or ON. The age of 55 y differentiated YA and OA. As seen in candidate gene study, at the age of 50s, genetics effects on CADi are obscure, whereas distinction occurs at younger and older ages (Zhang et al. 2010). Therefore, we divided the continuous variable into categorical by a single value cutoff, to account for the dramatic mid-life change. This approach might result in a loss of information, as age is not as straightforward a measurement to define subsets of samples as is gender. In

future studies, we might define young and old at two extremes of ages (i.e. < 50 years of age and > 60 years of age), and recruit more samples to avoid the loss of power.

In GxAGE analysis, we considered the entire age-range to study the interaction that spans from 20 to 80. Age was used as continuous variable and disease status was only based on CADi. Although the definition of cases and controls does not completely overlap with the stratified analysis, comparing theses two types of approaches has provided insights into discovering age-related variants. In spite of this general trend that associates aging with high CAD risk, we found this relationship can be modified by genotypes. An interaction OR less than one was presented in a decreasing relationship between age and CAD in minor homozygotes, or if there was an increasing trend the slope was not as steep as for other genotypes. Minor homozygotes are associated with higher probability of CAD in the young, and reversed in the old. Alternatively , an interaction OR greater than 1.0 was demonstrated in an increasing risk in minor homozygotes. Most of the interactions we discovered by GxAGE do not show significant main effects in stratified analysis. These results suggest that a multi-analytic approach needs to be considered in discovery and validation steps.

Comparing YA+ON with OA+ON, we found excessive SNPs associated with both groups at the threshold of $p < 0.01$. Consistency in effect sizes and directions prompts us to combine young and old cases in the future to search for main effects

independent of age. This consistency might reflect our nested case-control sampling approaches that compared young and old cases to the same group of controls. It might also exhibit characteristics of CAD onset as partly attributable to the same genetic effects persistent through ages. Together with the discovery of age-dependent variants, we have detected the genetic architecture of CAD as the combined effects of age-independent and age-dependent variants. Our studies demonstrate that careful attention should be paid to age range especially in selecting replication samples.

To relate our findings to functional annotation, age-specific pathway analysis is an appealing next step. Gene Set Association Analysis (GSAA) will facilitate this goal (Xiong et al. 2012). Furthermore, integrated studies of GxAGE GWAS, GSAA analysis, Aorta and mouse expressions will help search for related pathways. Expression quantitative trait loci (eQTLs) can correlate sequence variation with transcribed RNA levels, thus adding another layer of evidence to association findings.

2. Detection of GATA2 x Gene interaction in Coronary Artery Disease

2.1 Introduction

Genome-wide association studies of complex traits have often been confounded by a relatively small portion of the variance explained by identified genetic variants. This has led to discussions of where the missing heritability can be found. The presence of rare variants, variation in copy number, and epigenetic effects might be highly variable between cases and controls, thus accounting for a considerable portion of disease heritability(Donnelly 2008). In addition, complex diseases have complex genetic architecture, which is influenced by non-additive genetic effects, such as gene-gene interaction and gene-environmental interaction. Narrow-sense heritability will be inflated if these non-additive genetic effects are not accounted for in these genetic effects (Manolio et al. 2009).

In addition to missing heritability, the findings of GWAS have limitations in discovering functional variants and identifying their roles in biological pathways. Gene-by-gene (GxG) interactions are thought to give important insights into complex traits. Epistatic interactions haven been reported in complex traits, including obesity (Dong et al. 2003; Dong et al. 2005), breast cancer (Cox et al. 2006; Ritchie et al. 2001), autism (Ma et al. 2005) and type 2 diabetes (Cauchi et al. 2008). An epistasis model can detect a genetic variant even without a detectable main effect. Interaction requires that any given variant needs to be considered in the context of other variants (Moore 2003). It is

believed by most investigators that analysis of interactions could facilitate understanding of genetic architecture of complex traits.

If GxG interactions are important disease susceptibility mechanisms, analysis of the underlying biological models may support the findings of statistical interaction. Transcription factor binding is an important regulatory activity that influences gene expression and pathways. Modeling transcription factor (TF) binding events in GWAS will make a clear picture of gene regulation relevant to complex traits or disease endpoints. One way to analyze TF binding in GWAS is to study TF-by-gene interaction, which is a subset of whole genome gene-by-gene interaction (GxG). Modeling specific TF-by-gene interactions avoids that computational burden and multiple testing issues that are particular to genome-wide GxG (Musani et al. 2007). In the absence of validated TFBSs, studies of TFBS prediction with GWAS can potentially elucidate true binding sites, as well as trans-acting regulatory activity.

In spite of the fact that GxG interaction has been supported by multiple lines of evidence as an important contributor to complex traits, interaction detected by a statistical model doesn't necessarily imply true biological interaction (Greenland 2009; Musani et al. 2007). This makes the integration of other evidence necessary, i.e. transcription factor binding, gene expression.

GATA2 is a transcription factor and a regulator of selective endothelial gene expression. It plays an important role in maintaining vascular homeostasis (Menghini et

al. 2005). The disruption of such homeostasis is related to atherosclerosis (Minami et al. 2004). GATA2 has been suspected as a susceptibility gene of CAD. A linkage study of early-onset CAD discovered a locus on 3q13, under which GATA2 is located (Connelly et al. 2006). Expression analysis has shown that GATA2 exhibits regional expression patterns in human aorta that could be responsible for location-related atherosclerosis severity (Seo et al. 2004). Moreover, GATA2 is associated with CAD in our familial study (GENECARD) and cohort study (CATHGEN) (Connelly et al. 2006). Replication in Intermountain Heart. Collaborative Study (IHCS) has shown suggestive association for GATA2 (rs2713604, $p = 0.057$, OR = 1.2) (Horne et al. 2009). The convergence of evidence highlights the importance of analyzing GATA2 binding and its targets. Therefore, we investigated genome-wide GATA2-by-gene interaction. Analysis was carried out in age-stratified cases and controls following the previous analyses. Age has been shown to modify the relationship between disease and gene, causing "age-varying association" (Lasky-Su et al. 2008; Zhang et al. 2010). Considering young and old separately might control for age-varying association and detect age-dependent variants. Our goal is to test the feasibility of using a statistical model in detection of biological interaction, to discover statistically significant binding sites with functional evidence, and to uncover mechanisms joining the endothelial regulatory network and CAD development. We identified 131 loci showing a nominally significant ($p < 1.0E-04$) interaction across all comparison groups. Aligning these loci with ChIP-seq analysis, we found a paucity of

overlapping signals, suggesting statistical significance not agree with physical interaction as measured by ChIP-seq.

2.2 Methods

Sample selection, genotyping and quality control were described in previous methods session. Briefly, our samples were from the CATHGEN study, a cohort study that sequentially enrolled subjects undergoing cardiac catheterization at Duke University Medical Center from 2001 to 2011 (Shah et al. 2010). We selected cases and controls by using different CAD index (CADi) thresholds in young and old . For participants younger than 55 years of age, CADi > 23 was used to define young affected (YA). For older participants, a higher cutoff CADi > 67 was used to define old affected (OA). Both young affected and old affected used a shared control group: participants older than 61 years of age, whose CADi ≤ 23 and without history of cerebrovascular disease, peripheral vascular disease, MI, ICC, or CABG. Due to population stratification, we performed analysis stratified by ethnicity.

Samples were genotyped using Illumina HumanOmin1-Quad array. The same quality controls (QC) were done as described previously.

We selected two SNPs rs3803 and rs2713604 with association identified in family-based sample and case-control sample(Connelly et al. 2006; Horne et al. 2009). These two SNPs were individually tested for GATA2-by-gene interaction in a logistic regression model as implemented in PLINK (Purcell et al. 2007). This model includes all covariates

from the previous model, as well as a single GATA2 SNP (either rs3803 or rs2713604) and GATA2-by-gene cross product. The coefficient of the cross product term reflects the deviation from multiplicativity, whereby the combined effect of two SNPs is greater than the product of individual effects(Knol et al. 2007). An additive genetic model was assumed in all analysis. For SNPs showing suggestive interaction, linkage disequilibrium (LD) was calculated in PLINK (Purcell et al. 2007). We analyzed SNP pairs in a 500 SNP window with a step size of 5 SNPs, and used $r^2 < 0.2$ to determine independence.

Chromatin immunoprecipitation with comprehensive sequencing (ChIP-seq) data were downloaded from UCSC ENCODE Transcription Factor Binding Track. Three cell lines were examined for GATA2 binding: umbilical vein endothelial cells (HUVEC), neuroblastoma clonal subline of the neuroepithelioma cell line (SH-SY5Y), and K562 (Kanki et al. 2011). We searched the binding peaks for SNPs genotyped in our assays. For those SNPs present under the peaks, we compared their association signals to those for genome-wide SNPs to test for significant eenrichment of GxG interactions.

2.3 Results

2.3.1 GATA2 Main effect association

Among the 12 GATA2 SNPs genotyped, calculation of linkage disequilibrium (LD) shows five independent loci in whites and 6 in blacks, based on an r^2 cutoff of 0.5 (Table 3). GATA2 SNPs were tested for association in four case-control groups. Two

significant associations are found in white OA+ON, two associations in black YA+ON, and four associations in black OA+ON. rs3803 and rs11717152 are in strong LD ($r^2 = 1$). They exhibit agreement in OR and p-value across all comparison groups. The same patterns have been identified in LD pair rs1573858 and rs9851497 ($r^2 > 0.9$). Despite that no SNPs withstand correction for multiple comparison ($\alpha = 0.05$, n = 11, p = 0.0045), they still form an appealing set to study interaction, given interaction could be detected in the absence of independent main effects (Moore 2003).

Table 3: GATA2 SNPs associated with CAD in CATHGEN

	white YA+ON				white OA+ON				black YA+ON				black OA+ON			
Illumina SNP	Alleles	frq	OR(95%CI)	p	Alleles	frq	OR(95%CI)	p	Alleles	frq	OR(95%CI)	p	Alleles	frq	OR(95%CI)	p
RS11718692	C/A	0.1987	0.772(0.6,1.0)	0.09025	C/A	0.2124	0.782(0.5,1.2)	0.2699	C/A	0.181	1.976(1.0,3.8)	0.04009	C/A	0.1566	3.959(1.2,13.2	0.02471
RS2713594	G/A	0.4321	0.970(0.8,1.2)	0.7989	G/A	0.446	1.106(0.8,1.6)	0.5622	A/G	0.491	1.512(0.9,2.4)	0.08189	A/G	0.4398	1.100(0.4,2.8)	0.8404
RS1573949	A/G	0.3146	1.200(0.9,1.6)	0.1615	A/G	0.2911	1.024(0.7,1.5)	0.9016	A/G	0.2614	1.152(0.7,1.9)	0.5788	A/G	0.2256	0.263(0.1,1.2)	0.09112
RS3803	G/A	0.4884	0.771(0.6,1.0)	0.08522	G/A	0.4718	0.732(0.5,1.1)	0.1592	G/A	0.4593	1.944(1.0,3.7)	0.04596	G/A	0.4157	3.911(1.2,13.0	0.02605
RS2713604	G/A	0.3362	1.271(1.0,1.6)	0.05983	G/A	0.3169	1.134(0.8,1.6)	0.495	G/A	0.2579	1.307(0.8,2.2)	0.2991	G/A	0.2108	0.307(0.1,1.4)	0.1344
RS11717152	A/C	0.2439	0.860(0.7,1.1)	0.285	A/C	0.2465	0.819(0.5,1.2)	0.3421	A/C	0.2523	1.601(0.9,2.8)	0.09585	A/C	0.2289	3.491(1.0,12.0	0.04744
RS2335052	G/A	0.1601	0.928(0.7,1.3)	0.6606	G/A	0.1479	0.700(0.4,1.2)	0.1611	G/A	0.1991	0.584(0.3,1.0)	0.05165	G/A	0.2229	0.308(0.1,1,2)	0.09199
RS1573858	C/G	0.3591	1.063(0.8,1.4)	0.6277	C/G	0.3439	0.882(0.6,1.3)	0.4939	C/G	0.224	1.279(0.7,2.2)	0.3831	C/G	0.1867	0.265(0.0,1.5)	0.134
RS2860228	G/A	0.4022	0.986(0.8,1.3)	0.9081	G/A	0.3862	1.495(1.1,2.1)	0.0252	A/G	0.405	1.008(0.7,1.6)	0.9692	A/G	0.4036	2.791(1.0,8.1)	0.05846
RS9851497	G/A	0.3564	1.056(0.8,1.4)	0.6631	G/A	0.3415	0.861(0.6,1.2)	0.4143	G/A	0.2262	1.294(0.7,2.2)	0.362	G/A	0.1867	0.265(0.0,1.5)	0.134
RS7629791	G/A	0.4847	0.981(0.8,1.2)	0.8715	A/G	0.4883	0.734(0.5,1.0)	0.07918	A/G	0.4163	0.837(0.5,1.3)	0.4224	A/G	0.4036	0.202(0.1,0.7)	0.01566

2.3.2 GATA2 x Gene interaction

Two SNPs rs3803 and rs2713604 were selected to represent GATA2, and tested for interaction with genome-wide SNPs. They are in linkage equilibrium in both whites ($r^2 = 0.125$) and blacks ($r^2 = 0.070$). Logistic regression models included a GATA2xSNP interaction term in four case-control groups: white YA+ON, white OA+ON, black YA+ON and black OA+ON. Manhattan plots and Quantile-Quantile (Q-Q) plots are shown in Figure 9 and Figure 10. The Q-Q plots show a significant deviation from the expected lines, suggesting that our sample size has limited the power to detect

significant interaction. In particular, the black groups exhibited significant departure from the theoretical line in Q-Q plots. No genome-wide significant interaction was discovered. Next we restricted our analysis to SNPs with nominally significant interaction ($p < 1.0E-04$). After removing SNPs in LD ($r^2 > 0.2$), the number of independent loci ranges from 0 to 35 in four comparison groups. No interaction was detected in either rs3803-by-gene interaction (rs3803xG) or rs2713604-by-gene interaction (rs2713604xG) in black OA+ON, which is the smallest case group. rs2713604 exhibits stronger GxG than rs3803. The number of rs2713604xG trended higher than rs3803xG: 35 compared with 27 in white YA+ON, 24 compared with 16 in white OA+ON, and 23 compared with 6 in black YA+ON. In total, 131 significant interactions were found across all groups. Notably no overlap among the SNPs was found between any two groups. The number of significant main effect SNPs ($p < 1.0E-04$) ranges from 0 to 38 in the four case groups. Based on the number of loci, no stronger evidence of main effects over interaction effects was observed.

Next we used a less stringent threshold of $p < 0.01$. The SNPs shared by white YA+ON and white OA+ON were 415 and 521, in rs3803 and rs2713604 respectively. In contrast, the numbers of SNPs shared across races were 43 and 77, in rs3803 YA+ON and rs2713604 in YA+ON, respectively.

To test whether these interacting variants preferentially lie within regulatory elements, we investigated the locations of 131 loci showing significant ($p < 1.0E-04$)

interaction. Three loci are in proximity with multiple genes thus making it difficult to determine the locations of the SNPs relative to the regulatory region of each gene. One signal maps to the coding region, and three map to UTR. Most loci are located at intergenic (n = 76, 58.0%) or intronic (n = 48, 36.6%) regions. Compared to the distribution of genome-wide SNPs, where 55.7% in intergenic regions and 34.1% in intronic regions, we found slightly higher proportions of interaction SNPs in these locations. Concentration in these regions suggests potential regulatory roles.

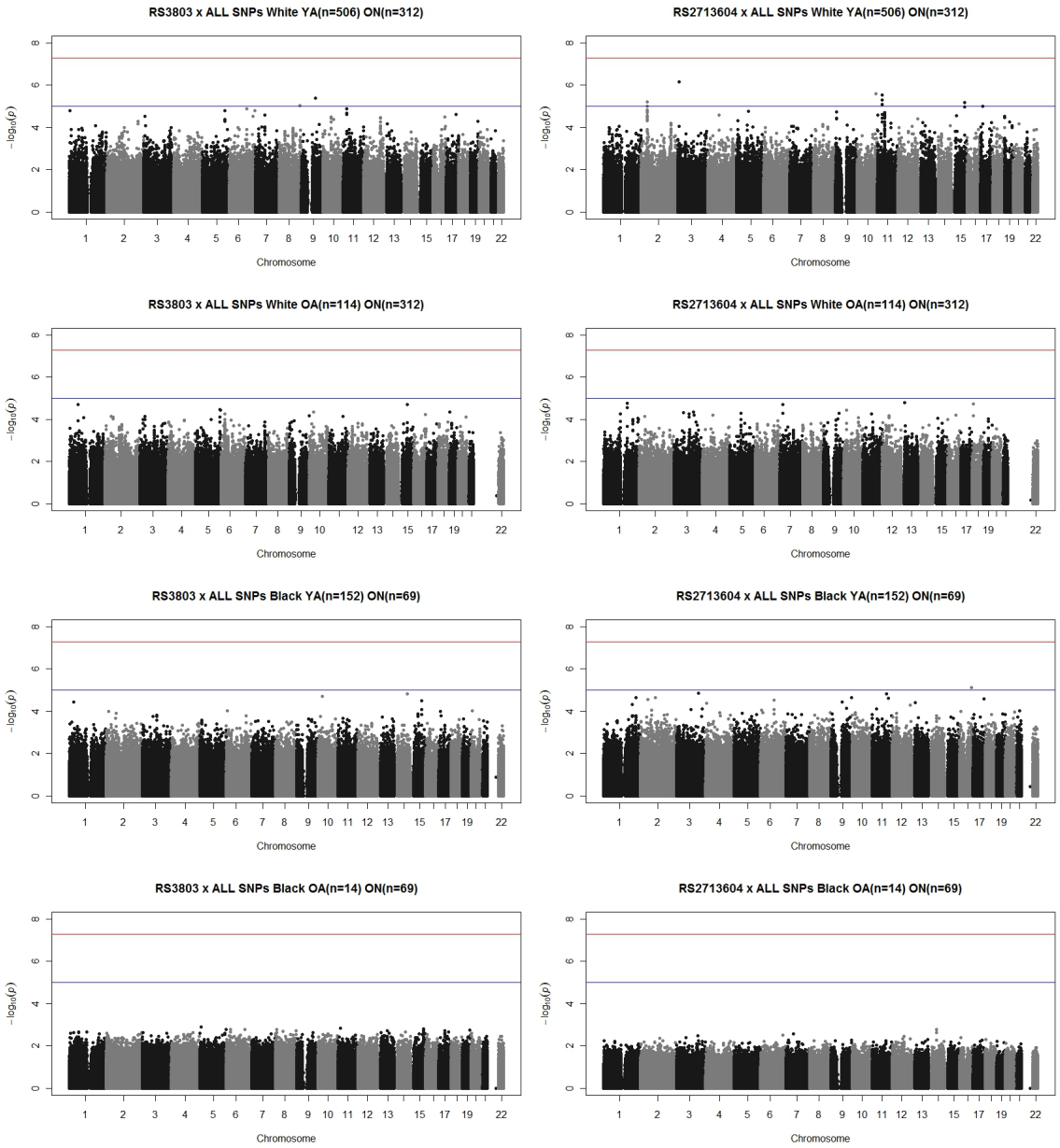

Figure 9: Figure 1. Manhattan plots of GATA2 SNP interacting with genome-wide SNPs. The –log10 of the Wald test p-value of GxG cross product were plotted at chromosomal positions. Two SNPs rs3803 and rs2713604 were interacting with genome-wide SNPs in white YA+ON, white OA+ON, black YA+ON, black OA+ON.

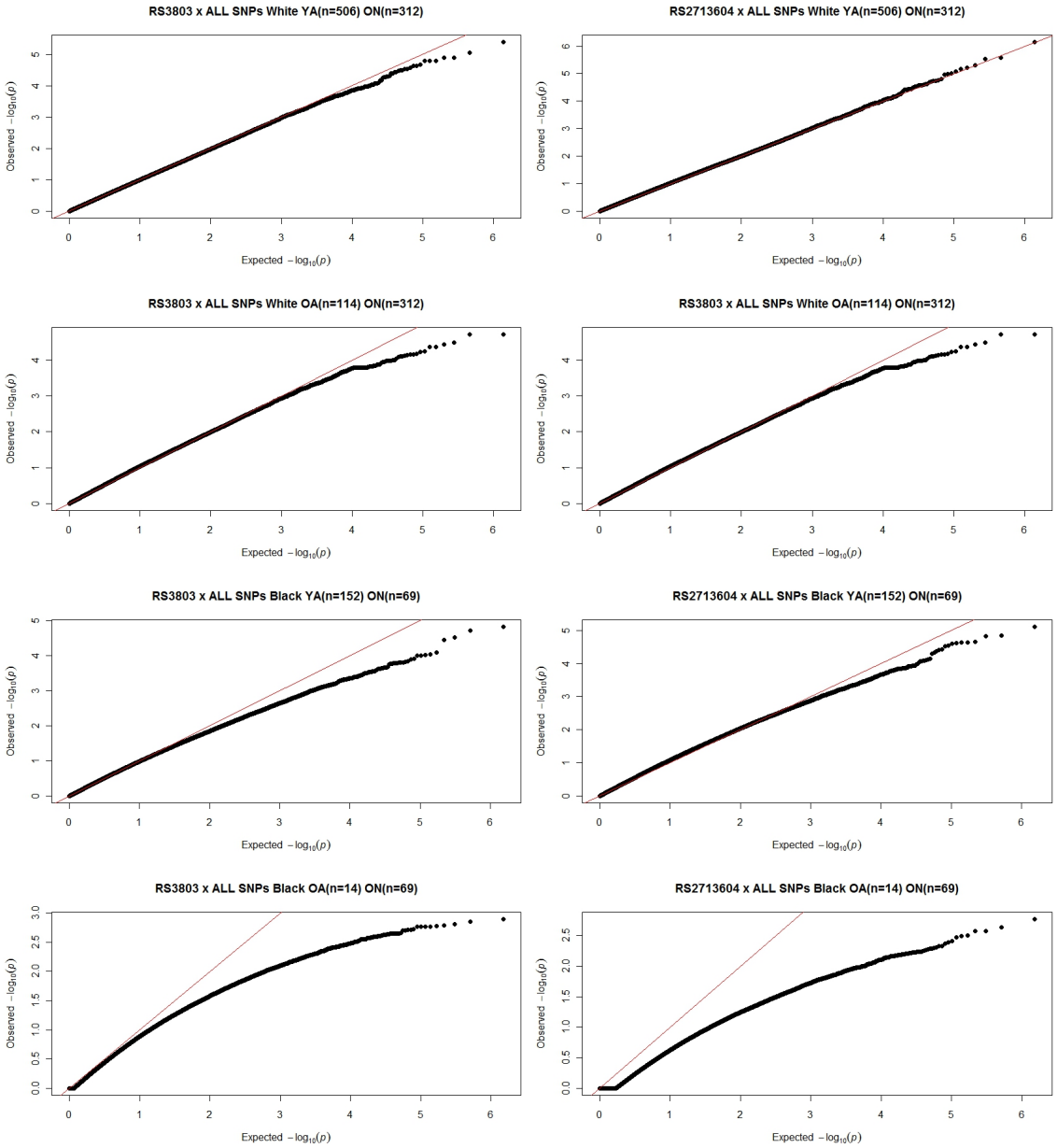

Figure 10: Quantile-Quantile Plots of GATA2 SNP interacting with genome-wide SNPs.

2.3.3 Characterization of GATA2 x Gene interaction

Using a p-value threshold of 1.0E-04, we detected 39 SNPs interacting with rs3808 in white YA+ON, including 6 LD blocks consisting more than one SNP in each.

SNPs in these LD blocks may exhibit stronger evidence of interaction. We selected the first SNP from the 6 LD blocks based on chromosomal coordinate order and studied its odds ratios modified by rs3803. rs34922070 is located in the intergenic region of ICOS and PARD3B. It has an interaction p-value of 5.15E-05 and minor allele frequency (MAF) of 0.46. In a model with rs34922070 alone, its odds ratio is not significantly different from 1, and the marginal association p-value is 0.7422. If we added rs3803 and rs3803-by-rs34922070 into the model, the interaction odds ratio is significantly different from 1. The interaction is exhibited using odds ratios Figure 11. We fitted the logistic regression model and get the estimates of odds ratios of rs3803, rs34922070, the interaction term and covariates in the model. We then fixed BMI to the average value, as we did for the top four eigenvectors in white. For the CAD risk factors in the model, we fixed their values to be 1 to match a high-risk CAD profile. The genotypes of rs3803 and rs34922070 are combinations of 0, 1 and 2, indicating the number of copies of the minor allele. For each combination, an odds ratio from the logistic regression model was calculated by incorporating genotypes and fixed covariates. As seen in Figure 11, antagonistic effects have been discovered if we stratified samples by rs3803 genotype. And rs34922070 shows zero marginal main effect without considering rs3803. This is an example to highlight that the relationship between rs34922070 and CAD has been modified by rs3803. Variants in rs34922070 might react against variants in rs3803 and lead to different disease susceptibility.

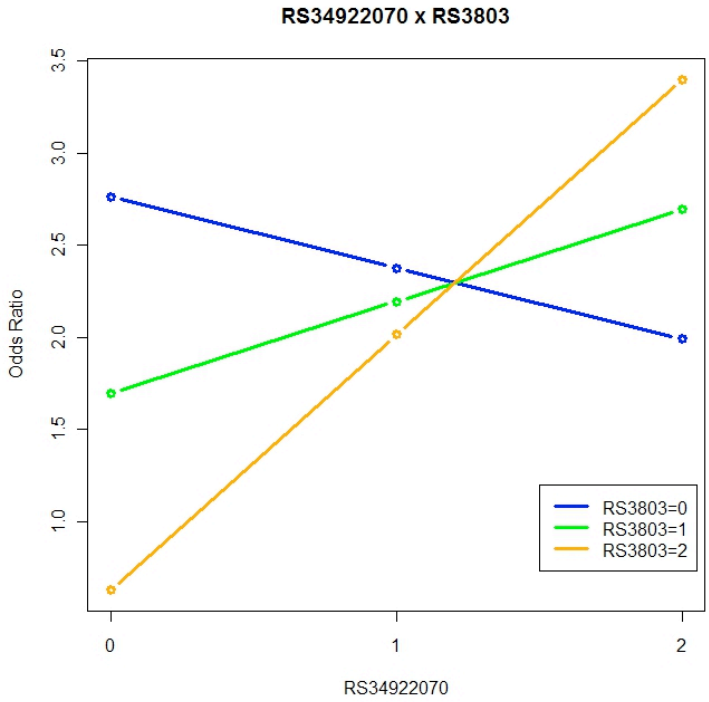

Figure 11: GATA2 genotype modifies rs34922070-CAD association. The full interaction model was fitted, and odds ratios were estimated. Fix all covariates in the model, and set genotypes of two SNPs to be combinations of 0, 1, and 2, indicating the copies of minor alleles. Covariates were set as: use the average of BMI and eigenvectors, and set all CAD risk factors (sex, history of smoking, etc.) to 1. Eight combinations of genotypes, together with fixed covariates were put back into the model to get combined odds ratios. Odds ratios were plotted for rs34922070, stratified by rs3803 genotype.

2.3.4 Comparison of GATA2 x Gene and ChIP-seq

GATA2 binding has been studied by chromatin immunoprecipitation with comprehensive sequencing (ChIP-seq) in umbilical vein endothelial cells (HUVEC), neuroblastoma clonal subline of the neuroepithelioma cell line (SH-SY5Y), and K562. Of the three cell lines, HUVEC most relates to our study of CAD, as GATA2 in endothelial cells may participate in the development of vascular disease. The number of peaks

ranges from 5518 to 50736 in three cell lines, including genes, regulatory elements and intergenic regions without confirmed roles.

As shown in Figure 12, only 807 ChIP-seq peaks are shared by three cell lines. The lack of many shared ChIP-seq peaks suggests that GATA2 binds to targets in a highly cell-type specific manner, and as such may control cell-specific gene expression (Heintzman et al. 2009).

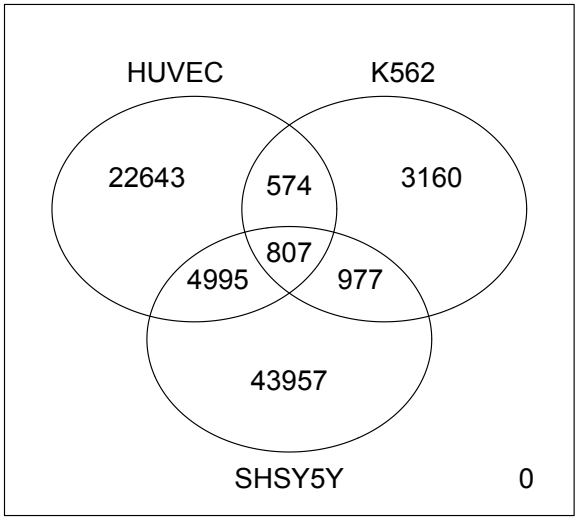

Figure 12: Venn diagram of ChIP-seq peaks discovered in three cell lines.

In order to find functional annotations of GxG discoveries, we assumed that SNPs in ChIP-seq peaks of GATA2 in HUVEC are functional variants. We speculated a correlation between GxG loci and experimentally discovered loci. First, we compared

the 131 SNPs showing significant GxG with the 29019 GATA2 binding sites in HUVEC. We found none located under the peaks, but 11 SNPs are in proximity (±5000 bp). Except for OA+ON, each group has one to SNP discovered this way.

Of the 29019 GATA2 binding peaks found in HUVEC, 7088 coincide with SNPs included on the Illumina HumanOmin1-Quad array. However, the average interaction p-values are close to 0.5 across most comparisons (Table 4). In black OA+ON, the average p-values is 0.734. In general, no comparisons have shown significant interaction in GxG model. These loci were bound by GATA2 at different signal intensity, indicated by signal value. Each locus was assigned a score from 0 to 1000 representing how dark the peak will be plotted in browser. If we limited to score of 1000, we found 2912 out of 7088 loci. These loci presumably illustrated binding events at higher confidence. However, as seen in Table 4, the average p-values are very similar to that of 7088 SNPs. These highly confident loci were not differentiated from the entire ChIP-seq loci by GxG interaction. Consequently, little correspondence was found in comparison of GxG and ChIP-seq analyses.

Table 4: Average p-values of SNPs under GATA2 binding peaks.

	RS3803xG				RS2713604xG			
	white YA+ON	white OA+ON	black YA+ON	black OA+ON	white YA+ON	white OA+ON	black YA+ON	black OA+ON
7088 BS	0.50136748	0.48716568	0.50355746	0.57645989	0.49697413	0.49602825	0.47563231	0.73423868
2912 BS score=1000	0.50010928	0.48228391	0.50344927	0.57609516	0.5095266	0.4957673	0.47655118	0.72971441

2.4 Discussion

Based on evidence of GATA2 as a susceptibility transcription factor in CAD, we took an initial step to discover its association with CAD, and found several nominal associations in our case-control samples. Two SNPs in linkage equilibrium were selected to represent GATA2, and tested for interaction with genome-wide SNPs. We detected 131 SNPs showing significant GxG, in which GATA2 genotypes modified gene-CAD association. Comparison of GxG findings and ChIP-seq did not support our hypothesis that significant GxG SNPs will be enriched in binding sites.

We have discovered that most interacting loci located at intergenic or intronic regions. Similar findings are shown in GATA2 ChIP-seq data that 42% binding regions were located in the intron and 43% at the intergene (Kanki et al. 2011). In addition to GATA2, CEBPA, HNF4A and FXR were shown to bind to the genome primarily in distal intergenic and intronic regions (Schmidt et al. 2010; Chong et al. 2010). These findings suggested that transcription factors might be involved in distal regulatory events. Even though our GxG analysis discovered many fewer sites located in proximity to promoter regions, we have consistent distribution of loci detected by ChIP-seq experiments (Kanki et al. 2011). These GxG loci without confirmed roles can be further integrated with open chromatin and gene expression data to elucidate their functions.

Notably no overlap was found between any two groups at the p-value cutoff of 1.0E-04. YA+ON and OA+ON shared the same group of controls. The absence of shared

loci might indicate age-specific interaction. If we used a p-value cutoff of 0.01, excessive shared SNPs were found. This is comparable to main effect association where we observed a large number of shared overlapped SNPs between YA+ON and OA+ON. This over-representation might manifest because of our nested case-control sampling. It might also provide some insights into the architecture of CAD that consists of both age-dependent and age-specific variants.

Blacks can serve as an independent replication data set for whites. However, between-race heterogeneity of CAD genetics has been observed, which is consistent with ethnic differences in CAD prevalence and risk factors (Chaturvedi 2003; Qi and Campos 2011).

If we applied a p-value threshold of 0.01, the number of shared SNPs across races was close to that of randomly selecting two SNPs from each group.

Furthermore, we found in our samples that allele frequencies vary among racial groups. Failure to replicate might reflect heterogeneity of disease architecture and transcription factor pattern recognition across races. In the context of environmental influences and other genes, the same gene might interact with different variants and impose different functional consequences in different racial groups.

As epistasis is investigated more widely in genetics and genomics studies, the interpretation of statistically significant interactions becomes more crucial. For detected interactions, mechanisms and pathways are assumed to be involved. But there is no

precise link between biological reality and statistically motivated interactions (Cordell 2002). Our analysis tried to align interacting loci with GATA2 ChIP-seq data. However, we have shown these ChIP-seq peaks do not enrich at the top of interaction list. In the logistic regression model, we associated the static genotype with fixed case-control status. In contrast, transcription factor binding is a highly dynamic and cell-type specific event that is prone to environmental influences. Complex disease often involves the complex networks of gene regulation and protein binding. The role of GATA2 is not confirmed in CAD development. It might be one of the many players in the model and contribute only a small effect observed on disease endpoints. Therefore, we could use endophenotypes that link GATA2 and CAD in future analysis to reduce noise from the system. ChIP-seq aims to estimate direct binding of protein to DNA. However, our approach did not look specifically for direct binding. Interacting loci might relate to combinatorial regulation of transcription factor binding, or involvement in the same pathway without any physical binding. Consequently, we could carry out microarray experiments to access whether a particular gene variant interacts with another and leads to the change of gene expression.

Our sample sizes are small so that we have limited statistical power to detect causal variants, as well as replicating findings in published GWAS. GxG interactions require even larger sample size to detect significant interactions. Power analysis will be

our next approach to guide our study design and incorporate more individuals into genotyping and analysis.

We assumed the two GATA2 SNPs we chose were "true" susceptibility SNPs, and GATA2-CAD relationship is mediated through GATA2 SNPs interacting with binding targets. However, rs3803 and rs2713604 variations are not confirmed in GATA2 transcription level, protein conformation, or binding affinity changes. Rs2713604 locates in Intron 5-6, and rs3803 locates in Exon 6. None of them encode protein. Without identified roles in GATA2 functional changes, rs3803 and rs2713604 might not be good candidate variants to study GATA2xG interaction. Alternatively, we could study rs2335052, a nonsynonymous GATA2 SNP in Exon 3. The association results for rs2335052 are not significant. However, inspite of the absence of a main effect and because it may have a functional role, rs2335052 can be studied for for significant GxG interaction.

In addition to our approaches, other tests can by carried out to compare GxG results with ChIP-seq. Genome-wide GxG interactions can be ordered by $-\log_{10}P$ into a ranked list. We will apply Gene Set Association Analysis (GSAA) to explore whether a priori set of ChIP-seq peaks over-represented at the top of the list (Xiong et al. 2012).

Various methods have been applied to detect epistasis. These methods can be divided into two categories: model-based and non-parametric. Most non-parametric methods focus on data reduction and pattern recognition (Ritchie et al. 2001; Nelson et

al. 2001; Ritchie et al. 2003; Moore and Hahn 2002). They are computationally intensive for GWA studies, but can be prospective follow-ups in detection of pathways. Our approach used a parametric logistic regression model which is built on certain assumptions about phenotype, covariates and errors. Violation of these assumptions might lead to invalid statistical inferences (Musani et al. 2007). In the future, we will search for methods that are reliable and powerful in detection of interactions. Finally, even if the detection of GxG interaction lacks biological relevance, allowing for interaction would improve the power to detect association and replicate statistically significant main effect (Greene et al. 2009).

References

Aulchenko YS, Ripatti S, Lindqvist I et al. (2009) Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. Nature genetics 41 (1):47-55. doi:10.1038/ng.269

Cauchi S, Meyre D, Durand E et al. (2008) Post genome-wide association studies of novel genes associated with type 2 diabetes show gene-gene interaction and high predictive value. PloS one 3 (5):e2031. doi:10.1371/journal.pone.0002031

Chaturvedi N (2003) Ethnic differences in cardiovascular disease. Heart 89 (6):681-686

Chong HK, Infante AM, Seo Y-K et al. (2010) Genome-wide interrogation of hepatic FXR reveals an asymmetric IR-1 motif and synergy with LRH-1. Nucleic Acids Research 38 (18):6007-6017. doi:10.1093/nar/gkq397

Connelly JJ, Wang T, Cox JE et al. (2006) GATA2 is associated with familial early-onset coronary artery disease. PLoS genetics 2 (8):e139. doi:10.1371/journal.pgen.0020139

Cordell HJ (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Human molecular genetics 11 (20):2463-2468

Cox DG, Tamimi RM, Hunter DJ (2006) Gene x Gene interaction between MnSOD and GPX-1 and breast cancer risk: a nested case-control study. BMC cancer 6:217. doi:10.1186/1471-2407-6-217

Dong C, Li WD, Li D et al. (2005) Interaction between obesity-susceptibility loci in chromosome regions 2p25-p24 and 13q13-q21. European journal of human genetics : EJHG 13 (1):102-108. doi:10.1038/sj.ejhg.5201292

Dong C, Wang S, Li WD et al. (2003) Interacting genetic loci on chromosomes 20 and 10 influence extreme human obesity. American journal of human genetics 72 (1):115-124. doi:10.1086/345648

Donnelly P (2008) Progress and challenges in genome-wide association studies in humans. Nature 456 (7223):728-731. doi:10.1038/nature07631

Fornage M, Lopez DS, Roseman JM et al. (2004) Parental history of stroke and myocardial infarction predicts coronary artery calcification: The Coronary Artery Risk Development in Young Adults (CARDIA) study. European journal of cardiovascular prevention and rehabilitation : official journal of the European

Society of Cardiology, Working Groups on Epidemiology & Prevention and Cardiac Rehabilitation and Exercise Physiology 11 (5):421-426

Friedlander Y, Austin MA, Newman B et al. (1997) Heritability of longitudinal changes in coronary-heart-disease risk factors in women twins. American journal of human genetics 60 (6):1502-1512

Greene CS, Penrod NM, Williams SM et al. (2009) Failure to replicate a genetic association may provide important clues about genetic architecture. PloS one 4 (6):e5639. doi:10.1371/journal.pone.0005639

Greenland S (2009) Interactions in epidemiology: relevance, identification, and estimation. Epidemiology 20 (1):14-17. doi:10.1097/EDE.0b013e318193e7b5

Hauck WW and Donner A (2004) Wald's Test as Applied to Hypotheses in Logit Analysis. Journal of the American Statistical Association, Vol. 72, No. 360, pp. 851-853

Hauser ER, Crossman DC, Granger CB et al. (2004) A genomewide scan for early-onset coronary artery disease in 438 families: the GENECARD Study. American journal of human genetics 75 (3):436-447. doi:10.1086/423900

Heintzman ND, Hon GC, Hawkins RD et al. (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature 459 (7243):108-112. doi:10.1038/nature07829

Horne BD, Hauser ER, Wang L et al. (2009) Validation study of genetic associations with coronary artery disease on chromosome 3q13-21 and potential effect modification by smoking. Annals of human genetics 73 (Pt 6):551-558. doi:10.1111/j.1469-1809.2009.00540.x

Hunter DJ (2005) Gene-environment interactions in human diseases. Nature reviews Genetics 6 (4):287-298. doi:10.1038/nrg1578

J Mackay GM (2004) The atlas of heart disease and stroke. World Health Organization, Geneva

Kanki Y, Kohro T, Jiang S et al. (2011) Epigenetically coordinated GATA2 binding is necessary for endothelium-specific endomucin expression. The EMBO journal 30 (13):2582-2595. doi:10.1038/emboj.2011.173

Knol MJ, van der Tweel I, Grobbee DE et al. (2007) Estimating interaction on an additive scale between continuous determinants in a logistic regression model. International journal of epidemiology 36 (5):1111-1118. doi:10.1093/ije/dym157

Lanktree MB, Hegele RA (2009) Gene-gene and gene-environment interactions: new insights into the prevention, detection and management of coronary artery disease. Genome medicine 1 (2):28. doi:10.1186/gm28

Lasky-Su J, Lyon HN, Emilsson V et al. (2008) On the replication of genetic associations: timing can be everything! American journal of human genetics 82 (4):849-858. doi:10.1016/j.ajhg.2008.01.018

Ma DQ, Whitehead PL, Menold MM et al. (2005) Identification of significant association and gene-gene interaction of GABA receptor subunit genes in autism. American journal of human genetics 77 (3):377-388. doi:10.1086/433195

Manolio TA, Bailey-Wilson JE, Collins FS (2006) Genes, environment and the value of prospective cohort studies. Nature reviews Genetics 7 (10):812-820. doi:10.1038/nrg1919

Manolio TA, Collins FS, Cox NJ et al. (2009) Finding the missing heritability of complex diseases. Nature 461 (7265):747-753. doi:10.1038/nature08494

Menghini R, Marchetti V, Cardellini M et al. (2005) Phosphorylation of GATA2 by Akt increases adipose tissue differentiation and reduces adipose tissue-related inflammation: a novel pathway linking obesity to atherosclerosis. Circulation 111 (15):1946-1953. doi:10.1161/01.CIR.0000161814.02942.B2

Minami T, Horiuchi K, Miura M et al. (2004) Vascular endothelial growth factor- and thrombin-induced termination factor, Down syndrome critical region-1, attenuates endothelial cell proliferation and angiogenesis. The Journal of biological chemistry 279 (48):50537-50554. doi:10.1074/jbc.M406454200

Moore JH (2003) The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Human heredity 56 (1-3):73-82. doi:10.1159/000073735

Moore JH, Hahn LW (2002) A cellular automata approach to detecting interactions among single-nucleotide polymorphisms in complex multifactorial diseases. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing:53-64

Musani SK, Shriner D, Liu N et al. (2007) Detection of gene x gene interactions in genome-wide association studies of human population data. Human heredity 63 (2):67-84. doi:10.1159/000099179

Murabito JM, Pencina MJ, Nam BH et al. (2005) Sibling cardiovascular disease as a risk factor for cardiovascular disease in middle-aged adults. JAMA : the journal of the American Medical Association 294 (24):3117-3123. doi:10.1001/jama.294.24.3117

Nelson MR, Kardia SL, Ferrell RE et al. (2001) A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. Genome research 11 (3):458-470. doi:10.1101/gr.172901

Price AL, Patterson NJ, Plenge RM et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics 38 (8):904-909. doi:10.1038/ng1847

Purcell S, Neale B, Todd-Brown K et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics 81 (3):559-575. doi:10.1086/519795

Qi L, Campos H (2011) Genetic predictors for cardiovascular disease in hispanics. Trends in cardiovascular medicine 21 (1):15-20. doi:10.1016/j.tcm.2012.01.002

Ritchie MD, Hahn LW, Moore JH (2003) Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. Genetic epidemiology 24 (2):150-157. doi:10.1002/gepi.10218

Ritchie MD, Hahn LW, Roodi N et al. (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. American journal of human genetics 69 (1):138-147. doi:10.1086/321276

Roger VL, Go AS, Lloyd-Jones DM et al. (2012) Heart disease and stroke statistics--2012 update: a report from the American Heart Association. Circulation 125 (1):e2-e220. doi:10.1161/CIR.0b013e31823ac046

Samani NJ, Burton P, Mangino M et al. (2005) A genomewide linkage study of 1,933 families affected by premature coronary artery disease: The British Heart Foundation (BHF) Family Heart Study. American journal of human genetics 77 (6):1011-1020. doi:10.1086/498653

Scheuner MT, Whitworth WC, McGruder H et al. (2006) Expanding the definition of a positive family history for early-onset coronary heart disease. Genetics in medicine : official journal of the American College of Medical Genetics 8 (8):491-501. doi:10.1097/01.gim.0000232582.91028.03

Schmidt D, Wilson MD, Ballester B et al. (2010) Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding. Science 328 (5981):1036-1040. doi:10.1126/science.1186176

Schunkert H, Konig IR, Kathiresan S et al. (2011) Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nature genetics 43 (4):333-338. doi:10.1038/ng.784

Seo D, Wang T, Dressman H et al. (2004) Gene expression phenotypes of atherosclerosis. Arteriosclerosis, thrombosis, and vascular biology 24 (10):1922-1927. doi:10.1161/01.ATV.0000141358.65242.1f

Shah SH, Granger CB, Hauser ER et al. (2010) Reclassification of cardiovascular risk using integrated clinical and molecular biosignatures: Design of and rationale for the Measurement to Understand the Reclassification of Disease of Cabarrus and Kannapolis (MURDOCK) Horizon 1 Cardiovascular Disease Study. American heart journal 160 (3):371-379 e372. doi:10.1016/j.ahj.2010.06.051

TCADG Consortium (2011) A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. Nature genetics 43 (4):339-344. doi:10.1038/ng.782

Xiong Q, Ancona N, Hauser ER et al. (2012) Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets. Genome research 22 (2):386-397. doi:10.1101/gr.124370.111

Zhang L, Connelly JJ, Peppel K et al. (2010) Aging-related atherosclerosis is exacerbated by arterial expression of tumor necrosis factor receptor-1: evidence from mouse models and human association studies. Human molecular genetics 19 (14):2754-2766. doi:10.1093/hmg/ddq172