THE GENETICS OF CARDIOVASCULAR RISK FACTOR CORRELATIONS

By

NURI KODAMAN

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Human Genetics

December 2015

Nashville, Tennessee

Approved:

Professor Douglas P. Mortlock

Doctor Nancy J. Brown

Professor David E. McCauley

Professor Melinda C. Aldrich

Professor Scott M. Williams

In memory of
Mehmet Nuri Kodamanoğlu
1923-2013

ACKNOWLEDGEMENTS

I would like to thank my Committee Chair, Doug Mortlock, and the Members of my Committee, Melinda Aldrich, Nancy Brown and David McCauley. Their intellectually challenging criticism, comments and suggestions made this work possible. David McCauley's dedication through a difficult year has been a source of inspiration, for which, my deepest respect. To Scott Williams, my mentor, and a great educator in the strongest sense of the word, I am forever grateful.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER I

OVERVIEW

Cardiovascular disease (CVD) is the leading cause of death worldwide[1]. Many clinical endpoints of CVD, such as myocardial infarction and ischemic stroke, involve thrombus formation and subsequent blood vessel occlusion. Individuals at risk for thrombotic events can be identified highly effectively by the presence of conditions such as obesity, hypertension, dyslipidemias, and insulin resistance. One recent study, for example, found that over 90% of CVD risk was attributable to only nine such risk factors.[2] However, using these risk factors to infer etiology is not straightforward, in part because no single risk factor or combination of risk factors is a necessary or sufficient condition of CVD. Different patterns of risk factors are known to lead to similar clinical endpoints, while similar patterns may associate with different clinical endpoints by population, implying multiple modes of cellular or systems-level pathogenesis.[3,4] In affected individuals, risk factors also tend to cluster together, making it challenging to discern the causative from the merely epiphenomenal.

This phenotypic complexity, reflecting vast possibilities of interaction between the innumerable genetic and environmental factors that contribute to CVD over many years, can be simplified by identifying conditions (related to environment, sex, or ancestry) that favor the emergence of specific risk factor networks. Characterizing phenotypic heterogeneity in this way can provide insight into the genetic architecture of CVD, improve risk assessment, and increase the power of genetic epidemiologic studies. Thus, before seeking to identify genetic factors that may be involved in shaping observed

patterns of cardiovascular risk factors, we focus in the first part of this manuscript on characterizing their prevalence and interrelatedness.

Towards this end, we have CVD-related data for a large cohort of urban and rural men and women in Ghana, participants of the Hypertension and ARterial Thrombosis (HeART) study.[5] An epidemiological transition is underway in sub-Saharan Africa, marked by a rapid decline in infectious diseases and a rise in chronic non-communicable diseases[6,7]. This has created an urgent need to understand the conditions that promote CVD in low- and middle-income countries, particularly conditions related to urbanization. Rural-to-urban migrations typically lead to less physical activity, greater psychosocial stress, and poorer nutritional habits[6]. Because CVD risk factors, including genetic factors, can vary by ethnicity, an effective global strategy of CVD prevention will require large studies in different populations.[8]

In addition to identifying the prevalence of CVD risk factors and their dependence on age, sex, and urbanization, in Chapter 3 we also assess the comparative relevance of individual risk factors to thrombosis within and across networks, by sex and environment. To gauge this relevance, we focus on two intermediate phenotypes of CVD, namely the plasma proteins plasminogen activator inhibitor type-1 (PAI-1) and tissue plasminogen activator (t-PA), which by virtue of their functional role and quantifiable expression provide a direct link between genetic and environmental variability on the one hand, and thrombotic clinical endpoints on the other. PAI-1 extends the stability and size of developing thrombi by inhibiting t-PA, the enzyme that dissolves the clotting protein fibrin. High plasma concentrations of PAI-1 have been shown to associate with the development of myocardial infarction and other thrombotic disorders (discussed in detail

in Chapter 2). Our premise in this section is that the degree to which quantitative risk factors correlate with PAI-1 and t-PA concentrations serves as a proxy for relevance to thrombotic pathogenesis.

In addition to environmental and behavioral variables, such as those associated with industrialization and urbanization, we know from heritability studies that genetic factors must explain a large fraction of the variation in CVD-related traits. However, the genetic loci discovered by genetic epidemiologic studies so far account for very little of this variation (as discussed in detail in Chapter 4). There is strong evidence that this "missing heritability" is due in part to an intrinsic quality of genetic architecture as such, namely the fact that most complex phenotypes are the product of many genes with often undetectably small physiological effects.[9] However, we argue here that the small statistical effect sizes in genetic epidemiologic studies can also arise from the context-dependence of genetic variants, with "context" comprising any number of environmental and physiological factors and genetic backgrounds. Addressing the fact that the statistical significance of a SNP at the population level may tell us little about its biological significance, we introduce novel, multivariate approaches to SNP discovery in Chapter 4 that leverage (rather than disregard or adjust for) context-dependence; this in contrast to the standard, single-locus models used in most association studies. In the first part of Chapter 4 we lay the theoretical groundwork for our methods, illustrating how context-dependent genetic effects may diminish the power of conventional, single-locus association tests, and how multivariate approaches, including those proposed here, can improve the power not only to detect biologically significant SNPs, but also to improve the biological characterization of SNPs deemed statistically significant, and to elucidate

the networks and pathways in which they are involved.

Finally, in Chapter 5, we turn to real data, and apply our novel multivariate methods to the genotyped exomes of Ghanaian participants of the HeART study described above. Our first task in Chapter 5 is to compare the results and performance of our multivariate statistical methods to existing approaches to establish proof of concept. Thus, restricting genotypic data to SNPs that have previously been reported to have significant effects on biological processes, we assess their effects on lipid traits and the correlations among them. Next, we apply our methods to exome-wide data, seeking to discover genetic variants that mediate the relationship between conventional risk factors and PAI-1. Whereas previous studies seeking to identify genetic factors that influence PAI-1 levels have merely adjusted for other CVD risk factors, such as body mass index and plasma triglycerides, here we seek to identify genetic variants that may increase thrombotic risk by influencing the covariance between these risk factors and PAI-1.

CHAPTER II

INTRODUCTION, BACKGROUND, AND SPECIFIC AIMS

Introduction and Background

Cardiovascular disease (CVD) and thrombosis

Cardiovascular disease is a broad phenotype that generally progresses over decades, encompassing many conditions of vascular origin, including coronary heart disease, stroke, rheumatic heart disease and deep venous thrombosis. The most prevalent CVD conditions worldwide are coronary heart disease and stroke.[10] The pathophysiology of both involves thrombosis, the process of clot formation inside of blood vessels. In a healthy individual, thrombosis contributes to the proper execution of hemostasis, a process that prevents excessive blood loss during mechanical vessel injury. However, certain conditions can promote thrombosis in the absence of mechanical damage, which in turn may lead to clinically adverse sequelae. When thrombi do not degrade properly, owing to an imbalance between the homeostatic processes of coagulation and lysis, they gradually become occlusive, as in a myocardial infarction, preventing the flow of blood through coronary arteries. Thrombi may also detach from the vascular surface on which they form, creating one of the many types of emboli, which can travel to the brain and cause ischemic strokes. A number of insults, such as atherosclerotic deposits on vessel surface, infection, or stagnant blood flow can lead to thromboembolic events, all of which would be associated with a different set of risk factors. The severity of these

prothrombogenic insults and the balance between clot formation and its destruction together determine cardiovascular fitness.

Determinants of CVD risk

Many studies have affirmed the efficacy of various risk factors in predicting CVD.[11] The term "risk factor" (as distinguished from a "risk marker") implies a causal role in etiology in the sense that eliminating a risk factor is expected to reduce risk.[12] In contrast, likely markers include cardiac troponins and homocysteine, as they appear be surrogates for causal physiological processes, and consequently inapt targets for intervention. CVD risk is typically estimated using statistical models such as the Framingham score, which take as inputs continuous measurements or discrete values of risk factors.[13] Apart from age, sex, and family history, most conventional risk factors, such as high blood pressure, cholesterol, obesity, smoking, and diabetes, can be modified or treated. The widely cited study INTERHEART study concluded that over 90% of the global population attributable risk (PAR) of myocardial infarction is attributable to only 9 potentially modifiable risk factors.[2] The total PAR for specific countries ranged from 75% to 100%. Eliminating these risk factors, then, should reduce the proportion of cases in a large population by the corresponding PAR. However, a PAR near 100% does not mean more powerful risk factors cannot be discovered. In fact, PAR can add up to greater than 100%, as when one factor is a necessary component of multiple causal sets of factors. When risk factors have the potential to generate overlapping sets of pathophysiological responses (e.g. dependent on conditions such as sex, environment, and genetic background), risk stratification can generally be improved.[14]

The extent to which genetics contribute to CVD risk is a question of both

scientific and public interest. Certainly, many of the traditional CVD risk factors have strongly inherited bases. Heritability estimates for plasma levels of HDL cholesterol, for example, are 40-60%.[15] Interestingly, genetic variants thought to have a causal role in raising lipid levels have been shown to associate with the first cardiovascular event (and moderately improve risk prediction) even after adjustment for actual lipid levels and all other risk factors.[16] Recently, genome-wide association studies (GWAS) have identified hundreds of single nucleotide polymorphisms (SNPs) with evidence of association with various CVD-related traits; effect sizes are typically very small, as for most complex phenotypes, but nonetheless have substantial influence in aggregate.[17,18] Family history is, accordingly, an important CVD risk factor: a history of premature atherosclerotic CVD in a parent confers a threefold increase in offspring risk,[19] while having a close relative who has experienced a myocardial infarction increases risk sevenfold.[20] Although a fraction of such risk may be attributable to a shared environment, a study of 21,000 Swedish twins found that a male whose monozygotic twin died of CHD before age 55 had a relative hazard of 8.1, while the relative hazard for dizygotic twins was 3.8. The analogous monozygotic/dizygotic relative hazards for women (with twin's age of death increased to 65) were 15.0 and 2.6, respectively. These findings underscored the importance of genetic variation to CVD, while also suggesting that genetic factors may play a greater role in women than in men, and that genetic effects may decrease at older ages.[21]

Indeed, as with most complex diseases, genetic susceptibility to CVD must be interpreted within an environmental context. With respect to its contribution to CVD risk, sex is best interpreted as an "environmental" risk factor in gene-environment space,

because genotype frequencies differ only trivially by sex within a population. As such, sex is perhaps the strongest "environmental" variable: CVD is known to present differently in males and females with respect to onset, prognosis, and response to treatment.[22]

The parallel increase of CVD prevalence with urbanization also reflects the fundamental importance of gene-environment interactions. The repercussions of adopting a "Western lifestyle" (characterized among other things by reduced physical activity and an energy-dense diet) are well illustrated by examining the negative effects of rural-to-urban migrations on cardiovascular health. One systematic review of such migrations in 18 different countries found that CVD risk consistently increased in migrant populations to a level between that of the rural population left behind and the urban population joined.[23] That this risk gradient recurred equally in men and women from many disparate populations (each with roughly homogenous genetic backgrounds) implicates the urban lifestyle as a major and universal risk factor for CVD. However, the consequences of migration in different populations proved to be quite variable with respect to specific risk factors, even when similar in terms of total risk. Most consistent were the effects on body mass index (BMI), total cholesterol, low-density lipoprotein levels (LDL), and blood pressure, while measures of high-density lipoprotein (HDL), hypertension, and fasting glucose exhibited the most variance between migrant populations. This heterogeneity likely stems from both genetic differences and environmental differences between the populations featured in the study.

Just as risk factor measurements vary by population, so do the relative contributions of risk factors to total CVD risk. Noting that low- and middle-income

countries bear 80% of the global CVD burden (measured in disability-adjusted life years), but that most knowledge on CVD risk is derived from European populations, the INTERHEART study explored whether the effects of risk factors varied among countries and ethnic groups around the world. Whereas the 9 major risk factors (mentioned above) were found to account for a similarly high proportion of the PAR in every ethnic group from the 52 countries considered, the relative deleteriousness of every risk factor varied.[2] Other studies have drawn similar conclusions.[24] For example, lipid levels appear to be less consequential in South Asian populations, while blood pressure contributes disproportionately to CVD risk in China.[4]

Similar patterns of risk factors have also been shown to associate with different clinical endpoints by population. For example, in China and much of Africa, stroke risk far outweighs ischemic heart disease risk, all else equal,[25] implying variable modes of cellular and systems-level pathogenesis. (See Figure 2-1).

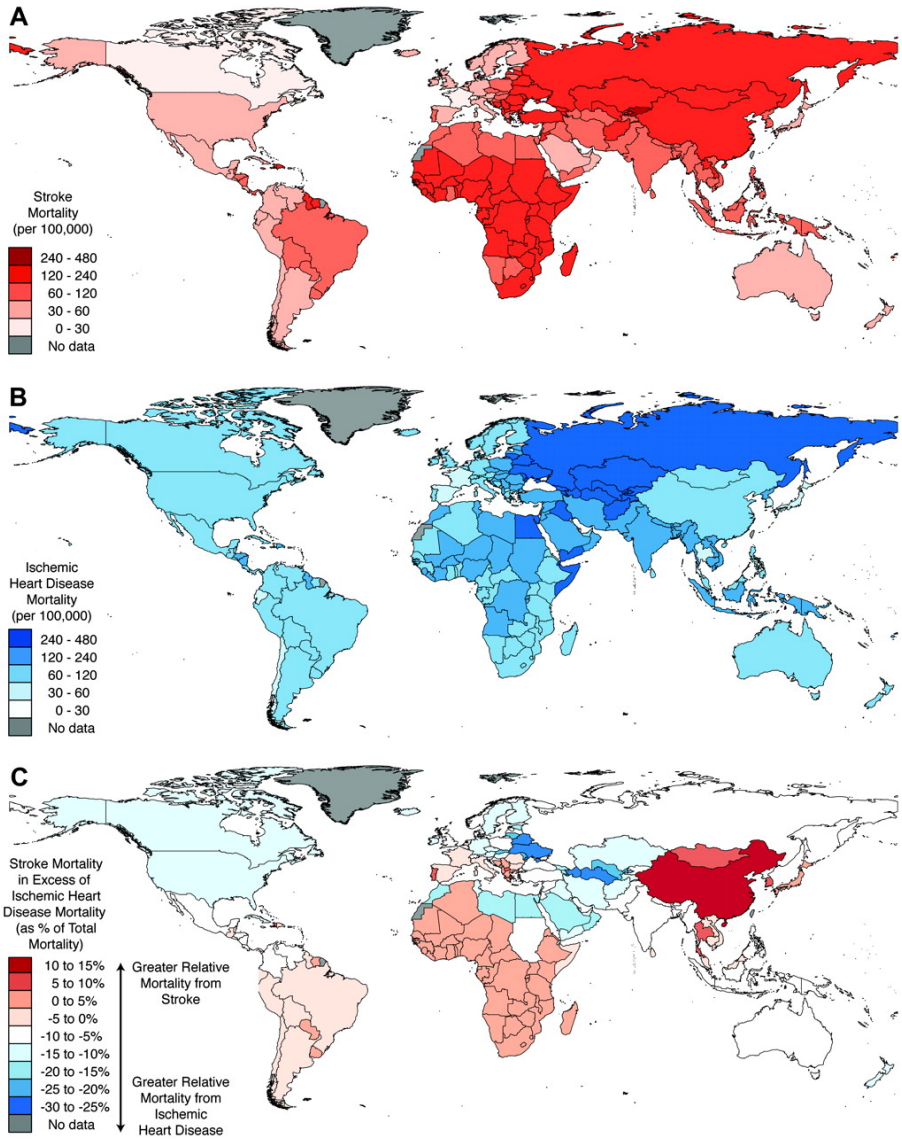


Figure 2-1: Geographic distribution of relative mortality from stroke and ischemic heart disease (World Health Organization Global Burden of Disease Program, 2004). Figure borrowed from Kim *et al*. (2011) [25]

t-PA and PAI-1

Fibrinolysis begins with the release of **t-PA** from endothelial cells:

- **t-PA** converts **Plasminogen** to **Plasmin**
- **Plasmin** degrades the **fibrin** clot, and any intact **fibrinogen**, into **fibrin degradation products**

PAI-1 inhibits the action of **t-PA** by binding to it

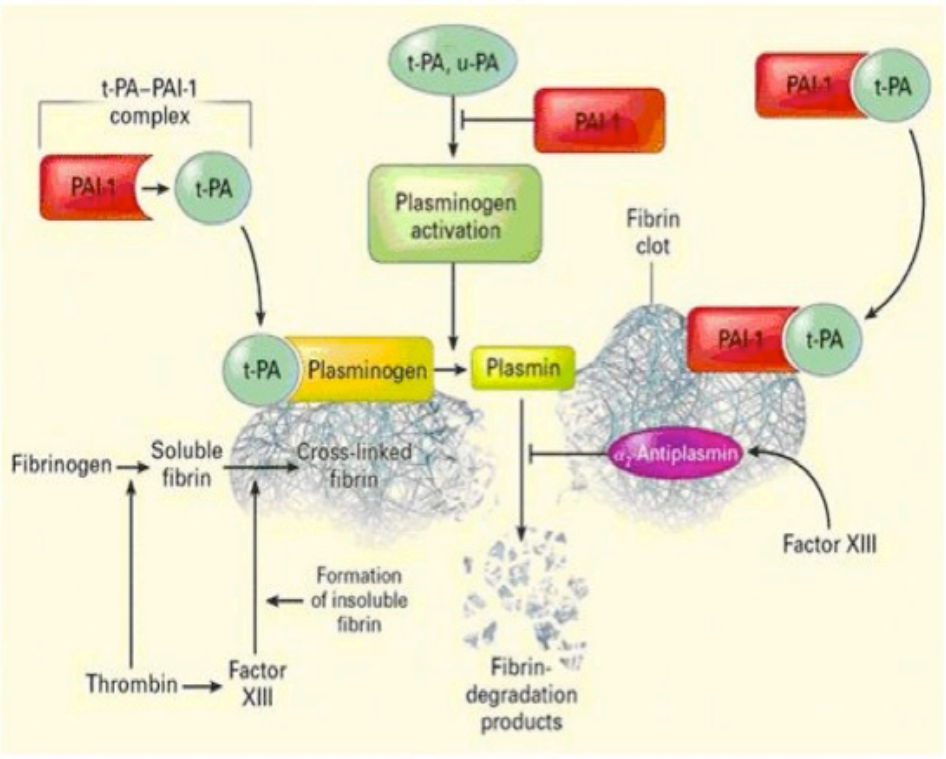


Figure 2-2: Schematic of fibrinolytic pathways. Figure adapted from Kohler *et al* (2000)[26].

Tissue plasminogen activator (t-PA) and plasminogen activator inhibitor-1 (PAI-1) are plasma proteins essential to maintaining hemostasis. Mechanical damage to a blood vessel triggers two concurrent processes to prevent bleeding into surrounding tissue: a temporary platelet plug rapidly forms, functioning to offer immediate protection of the exposed tissue, while the thrombin necessary to create a stable fibrin clot is generated. Damage to a vessel results in vascular smooth muscle spasm, or vasoconstriction, which reduces blood flow, allowing platelets to adhere via glycoprotein receptors to the exposed collagen at the site of

injury and to endothelial von Willebrand factors. Binding with platelets causes a steric change, and this alteration of shape leads to the release of additional vasoconstrictive factors (thromboxane A_2 and serotonin). ADP is also released by activated platelets to recruit more platelets to the site of injury, and fibrinogen is sequestered from plasma to form bonds between activated platelets. These steps temporarily stabilize the platelet plug[27].

Concurrently, thrombin is generated by a separate mechanism. Mechanical damage to blood vessels exposes not only collagen but also tissue factor (TF), and interactions between TF and a circulating coagulation factor, Factor VIIa, initiate a coagulation cascade through the activation of a series of other plasma coagulation factors. The coagulation factors are serine proteases, circulating in plasma as inactive zymogens, but once they are converted into their active form, they activate other downstream zymogens. This cascade eventually leads to the conversion of prothrombin to its active form, thrombin, at high levels[27].

Thrombin is then allowed to act on the fibrinogen that formed bonds between activated platelets in the platelet plug, and convert it to fibrin. The fibrin, in turn, polymerizes and forms a stable clot through cross-linking with Factor XIII[27]. While the formation of this stable fibrin clot is essential during vessel injury, it needs to be tightly regulated so as not to cause myocardial infarctions, pulmonary emboli, ischemic strokes or other thrombotic conditions[28,29]. The main mechanism for controlling the amount of clot formation depends on plasmin's ability to break fibrin down into soluble fibrin degradation products (FDPs).

Plasmin is the active form of plasminogen. Plasminogen circulates in plasma in a closed conformation (an inactive form). However, upon binding to the fibrin in clots, it undergoes steric changes, allowing several enzymes to activate it into plasmin[27]. Tissue plasminogen activator (t-PA) is one of the enzymes that can activate bound plasminogen. The circulating plasma

concentration of t-PA is very low, as most endogenous t-PA is found in vesicles of endothelial cells in its inactivated form. Like the coagulation factors, t-PA is a serine protease, and its production in endothelial cells and its release are stimulated by the presence of active thrombin. t-PA is released by the endothelial cells in its inactive form, and it is converted into its active state once it binds to fibrin[30]. Using fibrin as a cofactor in this capacity enhances the catalytic efficiency of t-PA over 100-fold[31]. However, access to fibrin is modulated by activated Factor XIII, which can mask the t-PA binding site of fibrin upon cross-linking. It has been demonstrated that t-PA is particularly efficient in degrading early stage fibrin, whereas the fully polymerized and cross-linked form is more resistant to t-PA fibrinolysis[26].

On account of its thrombolytic properties, t-PA is used in hospitals to re-perfuse any infarcted zones, with particular clinical applications to myocardial infarctions, pulmonary emboli, and ischemic strokes within 3 hours of first symptoms[27,29].

The main inhibitor of t-PA is plasminogen activator inhibitor-1 (PAI-1), a serine protease inhibitor. PAI-1 also serves to inhibit urokinase plasminogen activator, uPA. PAI-1 serves as a pseudo-substrate for both t-PA and uPA, forming a covalent complex with them in a one to one ratio, thereby preventing them from breaking down the fibrin clot. The t-PA:PAI-1 complex is eventually cleared from circulation by hepatic cells[32]. PAI-1 is found in endothelial cells, adipose tissue, and it is constitutively produced by activated platelets[33]. Its active form is very unstable, with a plasma half-life of 30 minutes; it has been demonstrated that active PAI-1 binds to vitronectin to form a more stable complex. The release of free active PAI-1 by platelets is important at the site and time of the vascular injury, where the concentration of the active form spikes locally and protects the clot from premature lysis. Like t-PA, PAI-1 can also bind fibrin, which stabilizes it while retaining its activity as an inhibitor of t-PA[26,32]. The inhibition of fibrin-

bound t-PA is less efficient because the catalytic domain of t-PA bound to fibrin is less accessible to PAI-1[34]. C-reactive protein, which is released from the liver upon stimulation by macrophages and adipose tissue during inflammatory states, has been shown to increase the levels of PAI-1[35]. Other PAI-1 stimulants include TGF-beta, TNF-alpha, plasma glucose and insulin[26,36]. The vasopressive hormone angiotensin II increases both the production as well as the secretion of PAI-1, and inhibition of angiotensin-converting enzyme has been associated with decreased total plasma PAI-1[26].

t-PA and PAI-1 assays

The mechanisms described above apply mostly to local circulation at the time of the insult, and are thus essential in the response to vessel injury. Understanding the mechanisms of local hemostasis at the time of injury requires insight into the acute plasma dynamics of the active forms of the t-PA and PAI-1. However, the active forms are of less use when studying their contribution to chronic disease. The steady state levels of t-PA and PAI-1 are governed by a different set of factors. Since the Ghanaians in this study were not undergoing cardiovascular insult at the time of the blood draws (aside from the venipuncture), the phenotype of interest needs to be ascertained by a metric that can serve as a proxy for baseline, steady-state plasma levels, as opposed to measures that are important at the time of vascular injury. In this regard, there are three major forms of t-PA in human plasma: active t-PA, t-PA complexed with PAI-1, and t-PA complexed with C1-inhibitor. Therefore, there are three possible measurements of t-PA in the blood: active t-PA, complexed t-PA and total t-PA.

In our protocol, samples of blood were drawn from each subject starting at 8 in the morning to limit the variability due to the diurnal release patterns of serum t-PA and PAI-1. These samples were used for ascertaining the plasma levels of t-PA and PAI-1 with the use of

the enzyme-linked immunosorbent assay (ELISA, Biopool AB, Umea). Samples were drawn in duplicate to provide back-ups. The measurements ascertained total t-PA in plasma; hence, they provided the sum of active t-PA and its complexed forms, whether bound to PAI-1 or C1-inhibitor. There are several reasons for using total t-PA measures as opposed to active t-PA in this study. From an ascertainment perspective, total t-PA is the more reliable measure. It has been shown that active t-PA levels are particularly influenced by the length of veno-occlusion during the venipuncture used to obtain the sample[37]. This is a consequence of the above-described processes in which a mechanical insult to the vessel causes the local release of active t-PA, thereby inflating its steady state value. Furthermore, special processing, such as immediate sample acidification and freezing are required to obtain accurate measures of active t-PA; if those are not adhered to, the levels of the measured protein can decrease by as much as 25%[37].

The pharmacokinetics of t-PA plasma concentrations are contingent on the secretion of t-PA from endothelial cells, inhibition of t-PA by PAI-1 or C-1, and clearance by the liver[38]. It is estimated that in the steady state, less than 20% of plasma t-PA is present as the active form, and that most t-PA in circulation is bound to PAI-1[26,32]. Since active t-PA is present at such low levels, the effects of the phlebotomy issues described above could significantly increase measurement error. Measuring active t-PA alone also ignores the complexed levels covalently bound in the t-PA:PAI-1 form. Since the liver readily clears active t-PA, an elevated measure might be indicative of a temporary local response to any vascular injury, as opposed to chronic thrombotic disease predisposition. A measurement of the complexed t-PA as well as the free form version gives a better idea of the chronic condition in circulation, because it involves both the rapidly metabolized active component as well as the more stable covalently bound forms that take longer to clear.

The antigen PAI-1 assay used in this study measured free PAI-1, PAI-1 complexed with t-PA and urokinase plasminogen activator[37]. Studying total PAI-1 is more appropriate for the objectives of this study than examining free PAI-1 because of the short half-life of the active molecule in plasma. A measure of active PAI-1 alone is a better representation of the enzyme's local, circumstance-dependent effects, because its half-life is less than 30 minutes. Total PAI-1 is a better approximation of the steady state levels over a prolonged period of time, because the t-PA:PAI-1 complex takes longer to metabolize and to clear from circulation.

t-PA and PAI-1 as CVD endophenotypes

Endophenotypes are especially appropriate for the study of complex diseases like CVD in which many potential diagnoses reflecting multiple etiological pathways present with similar symptoms.[39] PAI-1 and t-PA can be considered endophenotypes of CVD on account of their strong association with known cardiovascular factors on the one hand, and their direct biochemical connection to the process of thrombus formation and dissolution on the other. Compared to normal arterial tissue, severely atherosclerotic arterial tissue has been shown to have higher levels of PAI-1 messenger RNA.[40] This may be expected, as PAI-1 extends the stability and size of thrombi.

In addition to biomolecular data, epidemiologic evidence also supports the characterization of t-PA and PAI-1 as endophenotypes Supra-normal PAI-1 levels have been shown to associate with future adverse outcomes not only in CVD patients, but in healthy individuals as well.[41,42] In one study, low levels of plasma fibrinolytic activity at enrollment predicted subsequent incidence of coronary artery disease in young men, suggesting that low fibrinolytic activity precedes heart disease.[43] In an age- and sex-matched case-control study of 520 acute coronary syndrome patients, plasma PAI-1 concentration was shown to be an

independent risk factor for future cardiac events.[44] A prospective study similarly found baseline PAI-1 levels to associate significantly with adverse cardiac events, with a hazard ratio of 1.24 between the highest and lowest third of participants.[41]

Because PAI-1 inhibits the action of free t-PA by binding to it and forming an inactive complex, plasma concentrations of PAI-1 and t-PA correlate positively.[45] Thus, despite its role in breaking down clots, plasma t-PA also associates with increased CVD risk and CVD progression. In one nested case-control study on ischemic heart disease progression, t-PA levels were found to be significantly higher in cases than controls ($p < 0.001$).[46] In a prospective case-control study of 75,343 postmenopausal women without prior CVD, t-PA remained a significant predictor of coronary heart disease even after adjustment for lipid and non-lipid risk factors.[47] High t-PA levels have also been shown to associate with an increased risk of recurrent cardiovascular events in patients with established CVD.[48]

Genetic epidemiology of CVD risk factors

Before the sequencing of the human genome and the onset of genome-wide association studies (GWAS), the great majority of genes implicated in CVD by family studies and candidate-gene studies were of rare Mendelian conditions.[49] Since then, hundreds of genes related to polygenic CVD endpoints, such as myocardial infarction, have been identified by GWAS (Figure 2-3). While a substantial number of these common variants had previously been shown to be involved in rare monogenic disorders,[50] the majority reside in unexpected and often nongenic regions of the genome.[17]

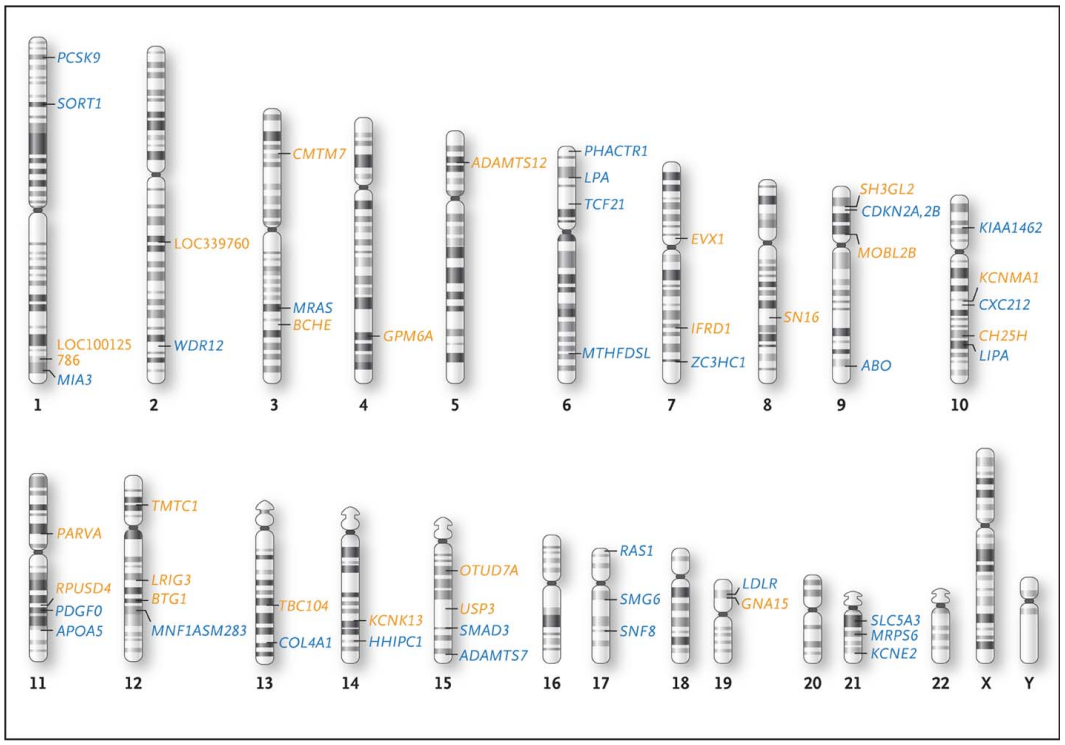


Figure 2-3: Genomic locations of genetic variants associated with the risk of Myocardial Infarction. Figure from O'Donnell CJ et al. (2011).[51]

With complex disease, defining the phenotype of interest is not always straightforward. In most cases, an inherently arbitrary decision has to be made between "lumping" and "splitting"; for, at some level of resolution, no two phenotypes are the same. Whereas the tendency in the field has been to err on the side of "lumping" (mainly because the value of GWAS depends to a large extent on sample size), the power to detect true genetic associations decreases with increased phenotypic heterogeneity. Compounding this problem is the unavoidable issue of genetic heterogeneity, i.e. the fact that the same heritable phenotype, however defined, can often be caused by different genetic loci in different people. The same loci may also have different effects based on extrinsic variables, such as environmental exposures. Cardiovascular genetic epidemiology thus faces the dual challenge of having to untangle the formidable genetic complexity of CVD while navigating the often cryptic phenotypic complexity inherent to it. Yet,

unless these challenges are overcome, insight into biological mechanisms will continue to be limited, while more ambitious goals, such as personalized medicine and drug development, will be even further off. Indeed, a remarkably small number of the hundreds of genome-wide significant loci associated with CVD risk factors (Figure 2-4) have led to functional or mechanistic insights that shed light on the statistical associations.[51] There is, for example, surprisingly little overlap between the genes shown to associate with CVD risk factors and those that associate with CVD-related endpoints and physiological traits.[52]

One way to deal with both phenotypic and genetic complexity is by using endophenotypes, or molecular phenotypic measures that are presumably links in the chain of physiological events leading to a broader phenotype. Because thrombosis is the major pathological step underlying coronary heart disease and stroke, coagulation and fibrinolytic factors in the bloodstream have been considered potential endophenotypes.[53] In previous studies, a fundamental criterion for choosing endophenotypes has been high heritability[54]. PAI-1 and t-PA, for example, have estimated heritabilities of up to 0.83 and 0.67, respectively, in some twin studies.[55,56] Many GWAS have been performed on coagulation factors, and novel genetic loci have been detected, including three each for PAI-1[57] and t-PA[58]. However, there has been virtually no overlap between the replicated genetic associations for these factors and those for other CVD-related traits, including endpoints.[59] Paradoxically, then, although CVD risk factors and t-PA and PAI-1 levels are known to be associated with CVD-related endpoints, as discussed above, genes that influence these phenotypes do not appear to contribute independently to CVD risk.

We maintain that this paradox can be resolved, at least in part, by meeting the challenges posed by both phenotypic and genetic heterogeneity head on, rather than by ignoring them. As a first step, we should accept the growing likelihood that genetic factors acting on single traits

independently of environmental and physiological context play only a minor role in CVD. In this spirit, we explore the possibility that genetic variants that influence the covariance of multiple CVD-related traits may be informative of CVD etiology, and hence clinically significant. Before tackling the genetic complexity of CVD, however, we explore the correlational architecture of cardiovascular phenotypes themselves, keeping in mind that multivariate approaches will profit from a better characterization of the phenotypes studied, and in particular, from a clearer understanding of the conditions under which networks of these phenotypes may vary.

Study	Risk Factor or Clinical Trait	Sample Size		Major Ethnic Group	Selected Major Findings
		Genomewide Association	Replication		
		number of subjects			
Ehret et al., 2011[42]	Systolic and diastolic blood pressure, hypertension	69,395	133,661	European	>25 Loci ,including *CACNB2* and *SH2B3*
Teslovich et al., 2010[43]	Total and LDL cholesterol	100,184	39,875	European, South Asian, East Asian, African	>35 Loci, including *SORT1* and *HMGCR*
Teslovich et al., 2010[43]	HDL cholesterol	100,184	39,875	European, South Asian, East Asian, African	>35 Loci, including *SCARB1* and *CETP*
Teslovich et al., 2010[43]	Triglycerides	100,184	39,875	European, South Asian, East Asian, African	>20 Loci, including *ANGPTL3* and *JMJD1C*
Thorgeirsson et al., 2008[25]	Quantity of cigarettes smoked	31,266	54,731	European	Top loci: *CHRNB3* and 15q25
TAGC, 2010[44]	Quantity of cigarettes smoked	74,053	68,988	European	Top loci: *DBH* and *CYP2A6*
Voight et al., 2010[45]	Type 2 diabetes mellitus	47,117	94,337	European	>25 Loci, including *TCF7L* and *IRS1*
Dupuis et al., 2010[46]	Fasting glucose level	46,186	76,558	European	>20 Loci, including *GCKR* and *ADRA2A*
Speliotes et al., 2010[47]	Body-mass index	123,865	125,931	European	>30 Loci, including *FTO* and *TMEM18*
Heard-Costa et al., 2009[48]	Waist circumference	31,373	38,641	European	Top loci: *NRXN2* and *MC4R*

Figure 2-4: SNPs discovered by GWAS of CVD risk factors, through 2011. Figure adapted from a table in O'Donnell CJ et al. (2011).[51]

Specific Aims

Specific Aim #1

To assess the correlational structure of cardiovascular risk factors and their association with PAI-1 in a Ghanaian population. Before seeking to identify genes that influence the patterns of association among cardiovascular risk factors, our first aim is to develop a thorough understanding of those patterns and the non-genetic forces that shape them. We will assess how cardiovascular risk factors and their networks vary by sex and urban environment in our study population. A better understanding of the relationships among risk factors can also provide insight into their etiologies and the pathogenesis of acute cardiovascular events. The strength of a risk factor's association with thrombogenic factors, for example, can be more informative of its contribution to ischemic risk than its mean level alone. With this in mind, we will examine the relationships between cardiovascular risk factors and PAI-1, and assess how these relationships may vary by sex and environment. We will approach these questions from multiple angles; for example, we will consider age-dependent effects and effects adjusted for body mass index; we will calculate partial correlations to identify possible direct or independent relationships among pairs of risk factors; and we will assess how the clustering of risk factor conditions (as in the metabolic syndrome) affects PAI-1 levels and varies with sex and environment.

Specific Aim #2

To establish a theoretical framework for understanding genetic effects on phenotypic correlations, and to develop multivariate genome-wide association methods that identify them. Few studies have looked for genetic variants that modify correlations between traits, and none has done so on a genome-wide basis. Our aim here is to develop methods towards that end, but

first we will need to explore this largely uncharted territory from a theoretical perspective. Given our present state of knowledge, how common should we expect correlation-modifying genetic elements to be? How would such elements "translate" back to biology? (For, correlation pertains only to populations.) Using these and related questions as a starting point, we will build theoretical models to guide both the development and the assessment of our statistical algorithms. An important consideration will be the pros and cons of methods sensitive only to genetic effects on correlation versus those that can detect genetic effects on individual traits as well.

Specific Aim #3

To identify genetic variants that influence cardiovascular risk factor correlations in the Ghanaian cohort. We will apply the methods developed above to the subset of 1105 Ghanaians genotyped by the Exome chip. We hope not only to discover genetic variants that influence the correlational networks of cardiovascular traits, but also to get an idea of their nature, relative abundance, and potential importance to genetic architecture.

CHAPTER III

PREVALENCE AND INTERRELATION OF CVD RISK FACTORS IN GHANA

CVD Risk Factors in Urban and Rural Ghana: a Cross-Sectional Analysis

Introduction

Urban populations in the developing world are growing rapidly and at an accelerating rate.[60] Rural-to-urban transitions are often associated with marked changes in behavior and lifestyle, such as diminished physical activity, sedentary employment, poorer dietary habits, and increased psychosocial stress.[61] In part because of these emerging risk factors, over 80% of the global burden of cardiovascular disease (CVD) has now shifted to low- and middle-income countries.[2] While proper screening and preventive strategies have reduced CVD in higher income countries, individuals at risk in the developing world are much less likely to be identified and treated, for reasons that include poor infrastructure, inadequate resources, and a lack of awareness regarding CVD and its symptoms in general.[62]

The fastest rate of urbanization worldwide is occurring in sub-Saharan Africa, driven by high fertility rates and rapid industrialization.[60] As in most of the developing world, the transition from pre-industrial to industrialized economies has initiated an epidemiological transition from illnesses related to malnutrition, childbirth, and infection, towards chronic, non-communicable diseases, such as CVD.[63] However, the epidemiological transition in sub-Saharan Africa is still in its early stages. As a consequence, diseases such as HIV and malaria continue to strain limited resources and dominate the public consciousness, while CVD and its often-subclinical symptoms

are overlooked.[64] Thus, populations are becoming older and more vulnerable to CVD at a time when surveillance capacities remain poor and skilled health workers scarce.[64,65]

Our knowledge of CVD epidemiology in sub-Saharan Africa is incomplete.[66] Early surveys revealed that risk factors such as hypertension and diabetes were rare, fueling the hypothesis that CVD was not of substantial public health interest.[67,68] More recently, this view has begun to change.[65,69,70] Nonetheless, variation in study designs and the diversity of the populations being studied have generated an often-confusing picture.[69,71] While some reports suggest that the proportion of disease burden attributed to CVD in sub-Saharan Africa may still be relatively low (primarily on account of persistent infectious disease-related mortality), the average age of death from CVD is the youngest in the world.[72] Thus, all the makings of a CVD epidemic are in place, as both life expectancy and the urban share of the population continue to increase.

Much of our understanding of CVD risk is based on studies of European populations, despite the fact that both the prevalence of risk factors and their relation to CVD endpoints differ among ethnic groups.[8,73] Existing risk assessment algorithms, such as the Framingham score, may consequently be prone to error when applied globally. Moreover, while such algorithms are typically calculated separately for males and females,[74] the effect of sex on CVD incidence and risk profile can also vary with culture and ethnicity.[75] Indeed, sex-specific effects appear to be more pronounced in the developing world, perhaps owing to differences in cultural practices and social behavior.[76,77] For example, men in sub-Saharan Africa are far more likely than women to smoke, whereas women are more likely to be overweight or obese.[65,78]

Given these heterogeneities of CVD risk profiles by sex, environment, and population, a multifactorial approach to CVD assessment and intervention is essential. Here, we describe how

major CVD risk factors, including dyslipidemia, hypertension, obesity, and diabetes, are distributed among urban and rural Ghanaian men and women from a single ethnic group. In addition to the conventional CVD risk factors, we also assess plasma levels of two fibrinolytically active enzymes that may provide deeper insight into CVD risk at the biochemical level. Our overriding goal is to evaluate the prevalence of CVD risk factors in the region and to understand the conditions that give rise to them, establishing a baseline for future comparisons and setting guidelines for appropriate recommendations.

Materials and Methods

Study Population

Unrelated participants were identified from Sunyani, the capital of the Brong Ahafo region of Ghana, population 250,000 as of the 2012 census, and from surrounding rural villages of fewer than 5000 people. Urban recruitment for the study began in 2002 and ended in 2007. Rural participants were recruited in 2008. Participants learned about the study at public venues, including local churches and markets. Individuals were excluded from analyses if they had signs of acute illness (e.g. malarial infection), were under 18 years of age, or were a first or second degree relative of someone already enrolled in the study. Participants provided information via questionnaire regarding their previous medical histories and other demographic and socio-economic variables, including age, sex, education, smoking status, alcohol consumption, and current medications. All participants provided informed consent. Institutional review boards at Vanderbilt University, Dartmouth College, and Regional Hospital, Sunyani approved all protocols.

Anthropometric measurements and biochemical analyses

Standing height and weight were measured to calculate body mass index (BMI). Blood pressure was measured twice; the means for both systolic blood pressure (SBP) and diastolic blood pressure (DBP) were used in subsequent statistical analyses. Blood was drawn between the hours of 8:00 AM and 10:00, after a minimum of 8 hours fast. These samples were used to assess fasting glucose, fasting lipids, and t-PA/PAI-1 levels. Fasting glucose levels were measured using a hand-held Sure Step glucose monitor by LifeScan, using blood drops from the blood

draw needles (LifeScan, Milpitas, California, USA). Total cholesterol (TC), triglycerides (TG) and high-density lipoprotein cholesterol (HDL) levels were measured in plasma; low-density lipoprotein cholesterol levels were calculated using the Friedewald equation (LDL=TC-HDL-TG/5). Plasma samples were stored in liquid nitrogen prior to shipment to Vanderbilt University, where concentrations of t-PA and PAI-1 antigen were measured using a commercially available enzyme-linked immunoassay (ELISA, Biopool AB, Umea).

Categorical outcomes

Hypertension was defined as: SBP ≥140 mm Hg, DBP ≥90 mm Hg, or current use of antihypertensive medication prescribed by a physician.[79,80] Diabetes was defined as a fasting glucose level ≥126 mg/dl or current use of an antidiabetic medication prescribed by a physician.[81] Impaired fasting glucose (IFG) represents an intermediate state of abnormal glucose regulation, associated with abnormal glucose tolerance, and often termed "pre-diabetes." The American Diabetes Association (ADA) now defines IFG as fasting glucose ≥100 mg/dl, having lowered the threshold from ≥110 mg/dl in 2003[82], whereas the World Health Organization (WHO) continues to recommend the 110 mg/dl cut point, citing a lack of evidence that lowering it offers any benefit with respect to reducing adverse outcomes.[83] All analyses below were performed using both cut points, and are referred to accordingly. Total cholesterol (TC), low-density lipoprotein (LDL), and triglycerides (TG) were considered high if they were ≥200 mg/dl, ≥130 mg/dl, ≥110 mg/dl, respectively; while high-density lipoprotein (HDL) was considered low ≤40 mg/dl[84-87]. Obesity was defined as BMI ≥30 kg/m^2, while BMI ≥25 kg/m^2 was deemed overweight.[88] All participants who smoked in the last 30 days qualified as current smokers. Years of education were dichotomized into two different variables, one reflecting whether a

participant had any schooling, and the other, attendance beyond Junior Secondary School (JSS). Ghanaian students typically attend JSS until age 15 in preparation for the "Basic Education Certificate Examination."[89 90]

Statistical Methods

Crude means and standard deviations (SD) or, where appropriate, medians and interquartile ranges (IQR) were calculated for all continuous variables after participants were stratified by sex and urban/rural environment into groups: urban males (UM, N=972), urban females (UF, N=1293), rural males (RM, N=469), and rural females (RF, N=583). Fasting glucose, TG, t-PA, PAI-1, and the ratio of TC to HDL were log transformed to obtain normal or near-normal distributions, and all continuous variables were adjusted for age, after which *t*-tests allowing for unequal variances were used to compare differences in means between sexes stratified by residence (UM vs. UF, RM vs. RF) and between urban and rural residents stratified by sex (UM vs. RM, UF vs. RF). For all continuous variables, differences in age- and sex-adjusted means among urban and rural residents and differences in age- and residence-adjusted means among male and female participants were standardized by dividing by the pooled standard deviations of residuals to estimate the "effect sizes" of urban environment and sex, respectively, on cardiovascular risk factors. These analyses were also performed after adjustment for BMI. The effect size of education beyond JSS among urban residents on CVD risk factors was also assessed, using age- and sex-adjusted residuals. This analysis was not performed on rural residents as too few had such schooling to be included. The "rules of thumb" for interpreting these effect sizes (i.e standard mean differences) are as follows: 0.2 = small, 0.5 = moderate, and 0.8 = large effect.[91] Similar analyses using logistic regression models that controlled for either

age and sex or age and environment were used to estimate the odds ratios of categorical clinical outcomes by environment or sex, respectively. The prevalence estimates of these outcomes were standardized according to the WHO 2000-2025 standard population, using recommended age bins that pertained to our data (18-24, 25-34, 35-44, 45-54, ≥55 years-old).[92,93] Mean values of all categorical and continuous variables were also calculated separately for these age groups. Statistical analyses were performed using STATA (version 12) and JMP (version 11).

Results

In total, 3317 individuals met all eligibility criteria, of which 2265 (68%) were urban dwellers (57% female), and 1293 rural (55% female). Ages ranged from 18-99 and were similarly distributed among urban males (UM), urban females (UF), rural males (RM), and rural females (RM) (p=0.23, Kruskal-Wallis test), with medians of 42.5, 43.5, 42, and 42, respectively (**Table 1**). Smoking was extremely rare among UW (0%), RW (2%) and UM (3%). The 16% of RM who qualified as smokers generally did not smoke cigarettes, but rather their own leaves, presumably tobacco (**Table S1**). Almost all UM (96%) and a similarly large proportion of UW (88%) reported some formal education (**Table S1**). Although this was true for only 64% of RM and 44% of RW, the difference was strongly related to age cohort (**Table S1** and **Figure S1**). With education beyond JSS, the contrast between urban and rural was even greater, with 48% and 30% of UM and UW, respectively, meeting the criterion, but only 5% of RM and 2% of RW (**Table S1** and **Figure S1**).

(Table 1 here)

Blood Pressure and Hypertension

In within-sex analyses (UM vs. RM, UF vs. RF), the age-standardized prevalence of hypertension and age-adjusted mean SBP and DBP were significantly greater in the urban participants (**Table 1** and **Table S1**). In comparisons between sexes stratified by residence (UM vs. UF, RM vs. RF), only SBP differed significantly, and was higher in men (p<0.001) (**Table 1** and **Table S1**). Male sex and urban environment both had small standardized effect sizes on SBP; urban environment had a moderate effect on DBP and hypertension (**Figure 2** and **Figure 3**). There was a marked increase in the prevalence of hypertension with age, which started about

a decade earlier in the urban cohort than the rural (**Figure 4)**. SBP and DBP increased with age as well, but less among rural participants (**Figure S2**).

BMI, Overweight and Obesity

Estimates of age-adjusted, mean BMI and age-standardized prevalences of overweight status ("overweight") and obesity all differed significantly between urban and rural residents stratified by sex (UM vs. RM, UF vs. RF) and between sexes stratified by residence (UM vs UF, RM vs. RF) ($p<0.001$) (**Table 1, Table S1,** and **Figure 1**). The standardized effect size of urban residence on age- and sex-adjusted BMI was large, while that of female sex was moderate (**Figure 2**). These effects were exaggerated at the right tail of the BMI distribution, with the odds of being overweight (BMI ≥25) or obese (BMI ≥30) 4.8 and 7.6 times greater, respectively, among urban residents (**Figure 3A**). Females had 2.9 times greater odds of being overweight and 5.2 times greater odds of being obese than males (**Figure 3B**). BMI, overweight and obesity increased with age chiefly among urban residents, continuing until 45 years of age among UW and until 55 years of age among UM (**Figure 4** and **Figure S3**). By age 45, over 70% of urban women were overweight and over 35% obese.

Fasting Glucose, Impaired Fasting Glucose, and Diabetes

In the pairwise comparisons, differences in age-standardized prevalence of diabetes between urban and rural residents (UM vs. RM, UF vs. RF) were significant ($p<0.001$), and differences between sexes (UM vs. UF, RM vs. RF) were not (**Figure 1** and **Table S1**). Differences in fasting glucose and IFG prevalence (using the ADA's 100 mg/dL cut-point), on the other hand, were consistently significant only between sexes (UM vs. UF, RM vs. RF)

($p \leq 0.001$) (**Table 1**, **Table S1**, and **Figure 1**). Comparisons using the WHO's 110 mg/dL cutpoint yielded less generalizable results, differing significantly only between UM and RM ($p=0.001$) and RF and RM ($p=0.003$) (**Table S1**). The standardized effect size of sex on fasting glucose was small, while that of urban residence was even smaller (by roughly half) (**Figure 2**). However, urban residents had 3.6 times greater odds of diabetes (**Figure 3**). Mean fasting glucose increased similarly with age in all groups (**Figure S4**), whereas the prevalence of diabetes began to increase sharply only by the 35-44 age group among UW, and the 45-54 age group among UM (**Figure 4**). Overall, the age-standardized prevalence of diabetes was 5.7% for urban men (95% CI: 4.4%-7.4%) and 6.6% (95% CI: 5.4%-8.1%) for urban women (**Table S1**).

Lipid traits and Dyslipidemias

Age-adjusted, mean TC and LDL were significantly higher in urban males and females than in their rural counterparts ($p<0.001$) (**Table 1**). The standardized effect size of urban environment on TC and LDL (adjusted for age and sex) was large, approaching one standard deviation. However, urban residence was not significantly associated with increased TG or lower HDL (**Table 1** and **Figure 2A**). Female sex had a small deleterious effect on LDL and TC, and a small beneficial effect on HDL (**Figure 2B**). Among all continuous risk factors, TC and LDL were most robust to adjustment for BMI (**Figure S7**). TG and HDL profiles became significantly worse in the rural residents after such adjustment (**Figure S7**). The effects of sex and environment on dyslipidemias as dichotomous traits (**Figure 3**), and their age-standardized prevalences (**Figure 1**) were broadly similar to results for the continuous measurements. In general, the increase in TC, LDL, and TG with age was steady and similar regardless of sex or environment, with the exception that in men (both urban and rural), TG became inelastic by the

45-54 age group (**Figure 5**). HDL, on the other hand, was the only measure in this study that did not exhibit a significant change with age (**Figure S5**).

PAI-1 and t-PA

All four pairwise comparisons of age-adjusted, mean t-PA (UM vs. RM, UF vs. RF, UM vs. UF, and RM vs. RF) were significant ($p \leq 0.004$), while only one of four tests yielded significant results for PAI-1 (UF vs. RF, $p < 0.001$) (**Table 1**). Analyses of standard mean differences were consistent with these results: urban residence had a moderate effect on t-PA and a small effect on PAI-1, while (male) sex had a small effect on t-PA, but no effect on PAI-1 (**Figure 2**). Mean t-PA increased with age in all groups, while the change in mean PAI-1 with age, though generally increasing, was more variable within and between groups (**Figure S6**).

Discussion

Economic development in sub-Saharan Africa has fostered an epidemiological transition, marked by an increase in the burden of chronic diseases, including cardiovascular disease. In Ghana, where more than half of the population now lives in urban areas,[94] recent epidemiologic studies have reported a rise in the prevalence of conditions such as hypertension, diabetes, and obesity.[69,71,95,96] Here we have taken a broad view of these and other risk factors to present a more complete picture of cardiovascular disease risk in the region, both as it currently stands and as we may expect it to increase with continued urbanization.

Once considered virtually absent in the sub-Saharan African region, hypertension has quickly emerged as a major epidemic.[97-99] In the absence of adequate infrastructure for screening, prevention and control, high blood pressure is rarely diagnosed in its early stages, when it is most modifiable[100]. Untreated, it can lead to renal failure, coronary heart disease, and stroke, particularly hemorrhagic stroke, which is the leading cause of cardiovascular disease-related death in people of African descent.[72] In our study population, the age-standardized prevalence of hypertension among urban residents (33%) was in the upper range of estimates previously reported for West African cities, including Accra, the capital of Ghana (30%).[69,95] Also in keeping with previous reports, hypertension prevalence was the same in urban men and women (34% and 32% respectively).[69] Urban residents were at significantly greater risk for hypertension than rural residents, primarily because mean urban DBP was almost one-half standard deviation greater than mean rural DBP. In contrast, urban residence was associated with a small increase in SBP. Male sex had a slightly larger effect than urban residence on SBP, but no significant effect on DBP. Notably, hypertension appeared to increase at similar rates among urban and rural participants by age 45.

Few large studies have assessed hypertension prevalence in rural West African populations. Our age-standardized estimates of 20% prevalence for rural men and 21% for women were similar to results from recent cross-sectional studies of rural Nigerian populations[101-103] (**Figure 3-1** and **Table S1**), although those studies did not present age-standardized, sex-specific prevalences, complicating comparison. Nonetheless, taken together with previous studies, including several smaller ones,[99,104-106] our data indicate that hypertension should no longer be considered rare in rural West Africa.

In contrast to early studies in Ghana, which estimated the prevalence of diabetes to be 0.4%[112] and 0.2%.[113] the age-standardized prevalence in the urban Ghanaians in our study was much higher, and comparable to the estimated world prevalence among all adults of 6.4[111] Importantly, awareness of diabetes throughout sub-Saharan Africa is low, and undiagnosed cases common, such that affected individuals are at higher risk for complications than in the developed world.[114] Thus, preventing a substantial escalation in diabetes-related morbidity and mortality in the face of continued urbanization and demographic ageing will be a major challenge in coming years.[70]

Throughout this study, we analyzed not only dichotomous clinical outcomes, but also the continuous risk factors that underlie them. While dichotomous outcomes may have more interpretable clinical significance, continuous measures such as BMI, blood pressure, fasting glucose, and total cholesterol are also associated with clinical endpoints,[107-110] and can provide complementary, often clinically useful information. and can provide complementary, often clinically useful information. This was evident in our analyses of diabetes and fasting glucose. Whereas urban residents were significantly more likely to have diabetes than rural residents, they did not have higher fasting glucose levels; in fact, median fasting glucose levels were highest

among rural females. Mean glucose was not significantly different between urban and rural participants even over the age of 55, when diabetes prevalence diverged considerably (15% urban vs. 4% rural, for both men and women). The increased risk of diabetes is therefore driven by the greater variance in the fasting glucose levels of urban participants (p-value = 0.0001; Levene's test), as well as the joint effects of fasting glucose with other correlated risk factors in urban environments (see Section B of this chapter).

Importantly, whether urban residence was a risk factor for impaired fasting glucose or not depended on which threshold (i.e. ADA or WHO) was used. When the ADA value (≥100 mg/dl) was used, urban and rural residents had roughly the same odds of impaired fasting glucose, or "pre-diabetes". Thus, if the ADA criterion for intermediate hyperglycemia is taken as a reliable predictor of future diabetes, rural Ghanaians are more at risk than generally thought. In addition, our results indicate that using only a continuous measurement may underestimate the clinically important differences between groups with respect to glucose metabolism.

Epidemiologic differences between sexes can be expected to reflect not only underlying pathophysiological and sociocultural factors, but also their interactions. To assess whether urbanization had sex-specific effects on the distribution of cardiovascular risk factors, we stratified participants in our initial analyses by both sex and environment, rather than categorically adjusting for sex. In fact, the changes in lifestyle that accompany urbanization are unlikely to be uniform across sexes, and even the same exposures may affect men and women differently.[107,115-118] Our results for BMI support this; we found that the age-standardized prevalence of overweight or obesity among urban women was singularly high, at 60%. These results are consistent with those of the Women's Health Study of Accra (64.9%).[119] Stratifying study participants by age as well as by sex and residence revealed that by age 45, over 70% of

urban women were overweight and over 35% obese. While urban residents (independent of sex) had 7.6 times higher odds of being obese than their rural counterparts, and women had 5.2 times higher odds than men, the combined effects of urban residence and female sex were greater than additive. Correspondingly, in a logistic regression model of obesity controlling for age, sex and residence, the sex-by-residence interaction term was highly significant ($p<0.0001$). The effect modifier here is likely sociocultural; it has been noted that increased body mass in sub-Saharan Africa has traditionally been recognized as a sign of social status and female attractiveness.[78,120] However, as excessive adiposity and its associated comorbidities impose enormous costs on quality of life and health-care systems[121], this situation warrants sustained intervention and calls for strategies for prevention.

The age-standardized prevalence of obesity among urban men in our study population was not particularly high (7%), but the percentage of those overweight was substantially greater, at 35%. This contrasted strongly with the rural men, among whom obesity was practically non-existent, and the proportion of overweight was low (11%) and did not increase significantly with age. Obesity and overweight were also low among rural women, indicating that these conditions are driven almost entirely by factors related to urbanization. Therefore, the fact that obesity is increasing faster in Ghana than in any other West African nation can probably be attributed to the rapid rate of urbanization there relative to other nations in the region.[122,123]

Few studies have assessed lipid traits in West Africa, and to our knowledge, no large study ($N>1000$) has done so in Ghana.[124,125 126] This may be partly because infectious and inflammatory causes of cardiovascular disease are relatively more common than atherosclerosis in sub-Saharan Africa, in contrast to other world regions.[72] The relatively small number of studies that have measured lipid traits have also reported generally favorable profiles, creating

the impression that dyslipidemia is not a problem.[127] For example, a 2011 survey of serum total cholesterol in 199 countries and territories found sub-Saharan Africa to have the lowest mean level among all world regions (158 mg/dl).[126] However, currently only about one-third of sub-Saharan Africa is living in urban areas. Because that number is rapidly increasing, with an inflection point projected for 2035,[128] understanding the effects of urbanization on lipid profiles may be more important than estimating their present levels.

Indeed, we found that urban residence had a stronger effect on age- and sex-adjusted total cholesterol and low-density lipoprotein cholesterol than on any other risk factor, raising them 0.85 and 0.9 standard deviations, respectively. Moreover, total cholesterol was higher in both male and female urban participants in every age group. Remarkably, the urban/rural differences in total cholesterol appear to be driven entirely by differences in low density lipoprotein, as there were no differences in HDL or triglycerides among participants from urban or rural settings. Rates of hypercholesterolemia (total cholesterol ≥ 200 mg/dl) and high LDL (≥130 mg/dl) were likewise consistently higher among urban men and women across all age groups (data not shown). The effects on total cholesterol and LDL were fairly robust to a BMI adjustment, indicating that some, but not all of the observed differences can be attributed to the disparity in BMI. Differences in quality of diet (e.g. the consumption of highly processed foods or unhealthy macronutrient ratios) may therefore also contribute to differences in lipid profiles.

Hypercholesterolemia, and in particular, high LDL levels, are strongly implicated in the pathogenesis of atherosclerotic plaque formation and consequent symptomatic cardiovascular disease. Within the context of oxidative stress and modification, oxidized LDL permits both intimal macrophage uptake (creating "foam cells") and induces cytotoxic damage to surrounding endothelial and smooth muscle cells.[129] Correspondingly, measures to decrease LDL have been

shown to reduce coronary events, peripheral vascular disease, and strokes, while slowing progression of atherosclerosis.[130 131,132] With pharmacologic improvements in cardiovascular risk profiles, higher LDL in this population represents a modifiable risk factor regardless of triglyceride or HDL levels.

Although mean total cholesterol levels among rural men (142.2 mg/dl) and women (152.3 mg/dl) were low, median triglyceride levels were unexpectedly high (82.5 and 82 for men and women, respectively). The difference between urban and rural age- and sex-adjusted triglyceride levels was not significant. High-density lipoprotein profiles were even less favorable among rural participants, with close to 40% of men and women having levels ≤40 mg/dl (Figure 3-1 and Table S1). In fact, when adjusted for BMI as well as age and sex, the rural triglyceride and HDL profiles were both about 0.3 standard deviation worse than the urban, indicating the presence of underlying causal factors unique to the rural environment. Although we have seen no explicit references to this trend, we note that other studies have also reported poor triglyceride and HDL profiles in rural populations globally, including in India,[133] Nigeria,[134] Peru,[135] Mexico,[87] and Guatemala.[136] This phenomenon deserves further study.

The screening of risk factors for subclinical cardiovascular disease can help identify individuals at high risk of myocardial infarction and stroke.[137] We assessed two such novel risk factors, plasminogen activator inhibitor-1 (PAI-1) and tissue plasminogen activator (t-PA). PAI-1 impedes the removal of thrombi from the vascular system by bonding to t-PA and neutralizing its thrombolytic properties. Clinical evidence indicates that spikes in PAI-1 increase the risk of thromboembolic events, whereas t-PA is clinically administered after ischemic stroke to clear arterial occlusions (Chapter 2). Importantly, PAI-1 is a highly pleiotropic risk factor, also playing a role in atherosclerosis and participating in biochemical pathways related to inflammation and

the deterioration of metabolic homeostasis.[138] Although t-PA and PAI-1 have opposite roles at the physiological level, plasma t-PA in epidemiologic studies is in fact positively correlated to PAI-1, because assays that measure t-PA typically detect it in bound form to PAI-1 (see Chapter 2). Accordingly, both t-PA and PAI-1 levels have been shown to associate with cardiovascular risk factors and clinical endpoints. It is not, in fact, clear which is the better predictor.[48]

Although rural men generally had the healthiest cardiovascular profiles in our study their plasma PAI-1 levels were comparable to those of urban men and women. This was unexpected, because PAI-1 is released by adipose tissue, making elevated plasma PAI-1 one of the hallmarks of obesity, whereas mean BMI among rural men was lower than that of urban men, and the prevalence of obesity was essentially zero. PAI-1 is also released by platelets and endothelial cells, and its expression is directly influenced by triglyceride levels[82], which were high among rural men; however, there was no significant difference in triglyceride levels between rural men and women, and rural women had the lowest PAI-1 levels among all groups. Thus, the reasons for higher PAI-1 among rural men are not clear, although they appear to be sex-specific.

Interestingly, t-PA appeared to be substantially more sensitive to both urban residence and sex than PAI-1. Urban residents had a small increase in PAI-1 but moderately increased plasma t-PA concentrations, and mean t-PA levels were higher among urban men and women of every age group. This may be of interest as prior results have indicated that t-PA may be a better predictor of CVD risk than PAI-1 (ref). However, the results for PAI-1 and t-PA were, to a large extent, directionally consistent, allowing us to conclude that urbanization is likely increasing cardiovascular risk because of pro-thrombotic and pro-inflammatory risk factors in addition to the conventional risk factors described above.

The limitations of our study are primarily those that pertain to cross-sectional designs and convenience sampling. These include restrictions on our ability to elucidate causal relationships and the possible introduction of biases. Some of the prevalence changes we observed with age, for example, may be due to a birth cohort effect, but repeated measurements or information on secular trends would be required to confirm this. However, broadly speaking, our results were consistent with those reported for similar populations. Additionally, because we sampled from a relatively small city, the estimates of the effects of urbanization presented here are likely to be conservative. By sampling from rural villages of fewer than 5000 people (in most cases, much fewer), where subsistence farming is still the main occupation, we also hope to have limited the potentially confounding factors introduced by technological advances into more semi-rural settings.[139] Finally, our conclusions have implicitly assumed that the risk factors we measured affect disease risk in African populations much as they do in populations of European descent, for whom most clinical studies thus far have been conducted.

Our results, taken on the whole, underscore the dramatic role of urbanization in changing CVD risk profiles in Ghana. We note that urbanization appears to be the dominant factor in producing the less favorable risk profiles related to blood pressure, BMI, fasting glucose, lipids, PAI-1 and t-PA. However, there are important exceptions, such as triglycerides, and particularly HDL. Prospective studies in multiple venues will be required to clarify and build upon the results presented here, with the ultimate goal of understanding how CVD risk factors can act together to affect clinical disease. Nonetheless, we have described key transitions that are central to chronic disease etiology, the understanding of which will become increasingly important in sub-Saharan Africa.

Table 3-1. Physiologic and metabolic variables in the Ghanaian cohort.

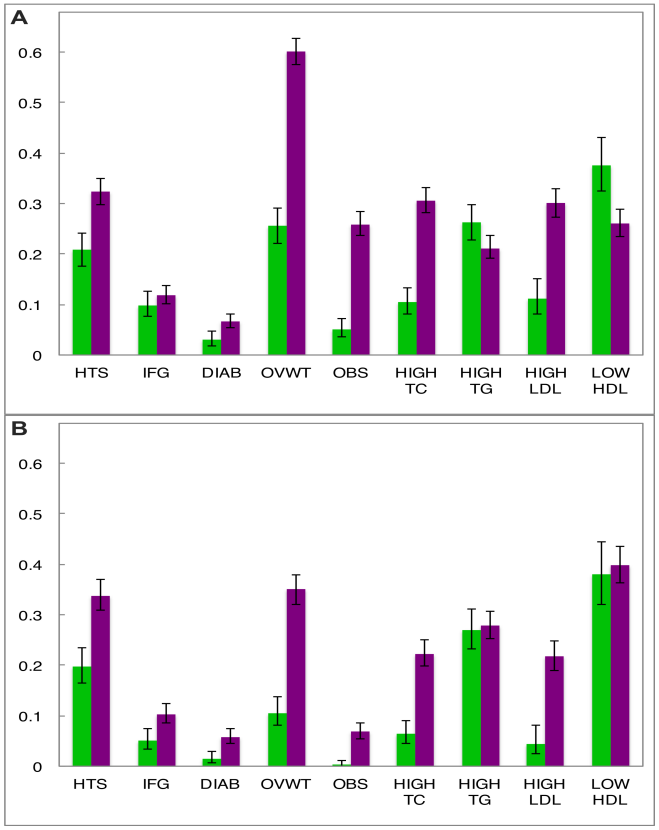
	Females			Males			Urban	Rural
	Urban	Rural	p-value	Urban	Rural	p-value	p-value by sex	p-value by sex
N	1293	583		972	469			
Age (years)	42.1 (11.3)	43.9 (15.9)	0.005	42.9 (12.6)	44.9 (17.2)	0.005	0.113	0.333
BMI (kg/m^2)	26.9 (5.6)	22.9 (3.9)	<0.001	24.0 (3.9)	21.5 (2.7)	<0.001	<0.001	<0.001
SBP (mm Hg)	125.1 (18.3)	123.8 (20.2)	0.002	130.2 (18.9)	127.3 (16.9)	0.002	<0.001	<0.001
DBP (mm Hg)	77.7 (10.7)	73.7 (11.6)	<0.001	78.0 (12.4)	73.5 (10.8)	<0.001	0.694	0.623
TC (mg/dL)	181.8 (42.1)	152.3 (36.9)	<0.001	170.6 (42.5)	142.2 (36.4)	<0.001	<0.001	<0.001
LDL-C (mg/dL)	113.9 (37.6)[1]	88.6 (32.3)[2]	<0.001	106.3 (34.1)[3]	76.4 (27.3)[4]	<0.001	<0.001	<0.001
HDL-C (mg/dL)	49.2 (14.6)[1]	46.5 (15.9)[2]	0.002	43.5 (13.3)[3]	44.5 (14.7)[4]	0.002	<0.001	0.212
TC/HDL-C	3.8 (1.6)[1]	3.3 (1.8)[2]	<0.001	3.9 (1.7)[3]	3.2 (1.5)[4]	<0.001	<0.001	0.097
TG (mg/dL)	77 (47)	82 (52)	0.103	83 (57)	82.5 (53)	0.103	<0.001	0.084
Glucose (mg/dL)	93 (15)	94 (14)	0.371	91 (15)	90 (14)	0.371	<0.001	<0.001
t-PA (ng/mL)	6.4 (4.6)	4.3 (3.4)	<0.001	6.7 (5.3)	5.6 (4.3)	<0.001	0.004	<0.001
PAI-1 (ng/mL)	3.9 (6.3)	2.9 (4.4)	<0.001	3.7 (6.3)	3.5 (4.8)	<0.001	0.282	0.253

[1]n=955, [2]n=317, [3]n=722, [4]n=225

Data shown as: crude mean (standard deviation), except for TC/HDL-C, TG, glucose, t-PA, and PAI-1, shown as: median (interquartile range)

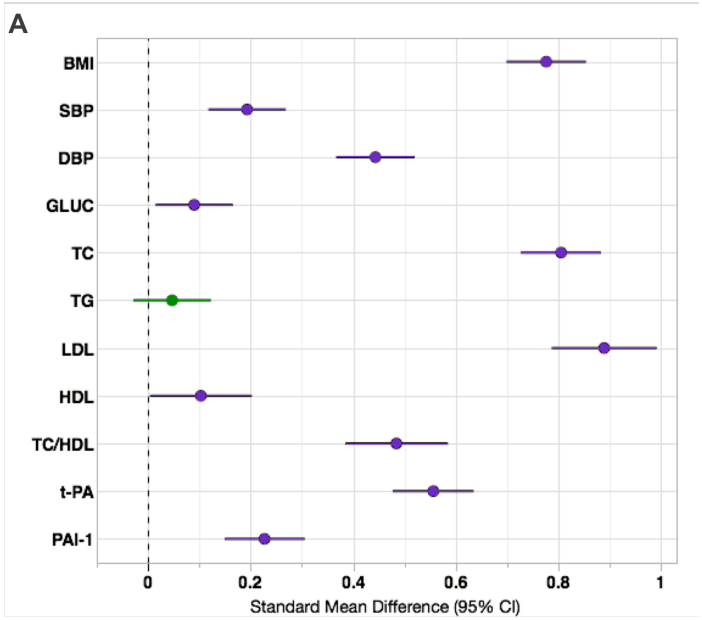
BMI - body mass index; **SBP** - systolic blood pressure; **DBP** - diastolic blood pressure; **TC** - total cholesterol; **LDL-C** - low density lipoprotein cholesterol; **HDL-C** - high density lipoprotein cholesterol; **TG** – triglycerides; **Glucose** – fasting plasma glucose; **t-PA** - tissue plasminogen activator; **PAI-1** - plasminogen activator inhibitor;
p-value: *t*-test (allowing for unequal variances) was performed on age-adjusted residuals to evaluate significance of difference between means; TC/HDL, TG, glucose, t-PA, and PAI-1 were first log-transformed.

Figure 3-1. **Age-standardized prevalence rates of dichotomous clinical outcomes by sex and urban/rural environment in Sunyani, Ghana. (A)** Urban (purple) and rural (green) females; **(B)** urban (purple) and rural (green) males. Error bars denote 95% confidence intervals of estimates.



Abbreviations: HTS – hypertension (SBP ≥140 or DBP ≥90); IFG – impaired fasting glucose (using the WHO cut-point of 110 mg/dL); DIAB – diabetes (glucose ≥126 mg/dL); OVWT – overweight (BMI ≥25); OBS – obesity (BMI ≥30); HIGH TC – hypercholesterolemia (cholesterol ≥200 mg/dL); HIGH TG – elevated triglycerides (≥ 110 mg/dL); HIGH LDL – elevated low-density lipoprotein cholesterol (≥130 mg/dL); LOW HDL – low high-density lipoprotein cholesterol (≤40 mg/dL). For UF, RF, UM, and RM, N=1293, 583, 972, and 469 (except for HIGH LDL and LOW HDL: N=955, 317, 722, 225), respectively. All data age-standardized to the WHO 2000-2025 standard population.

Figure 3-2. The effect of urban/rural environment and sex on cardiovascular risk factors in Sunyani, Ghana. (A) Absolute differences between urban and rural standardized means (with 95% confidence intervals); colors represent the group with the higher mean (purple: urban; green: rural). Data were adjusted for age and sex. (B) Absolute differences between male and female standardized means (with 95% confidence intervals); colors represent the group with the higher mean (red: female; blue: male). Data were adjusted for age and urban/rural residence.


A
BMI
SBP
DBP
GLUC
TC
TG
LDL
HDL
TC/HDL
t-PA
PAI-1
0
0.2
0.4
0.6
0.8
1
Standard Mean Difference (95% CI)

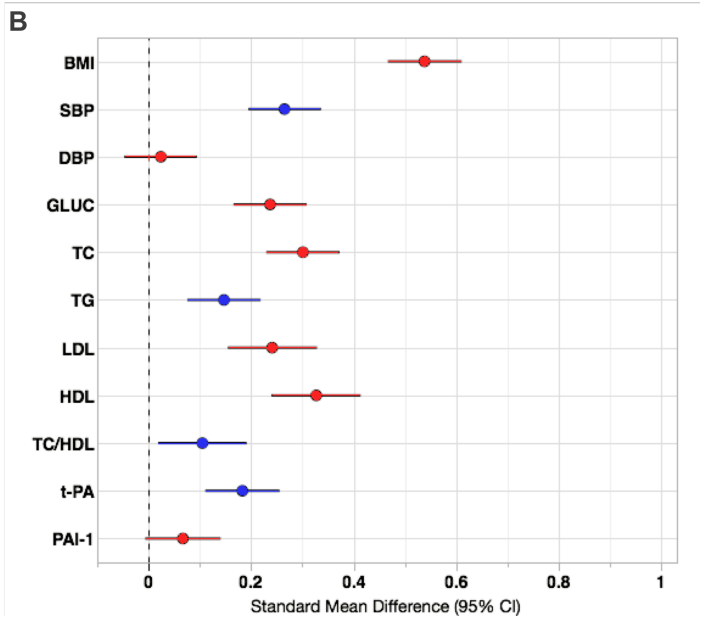
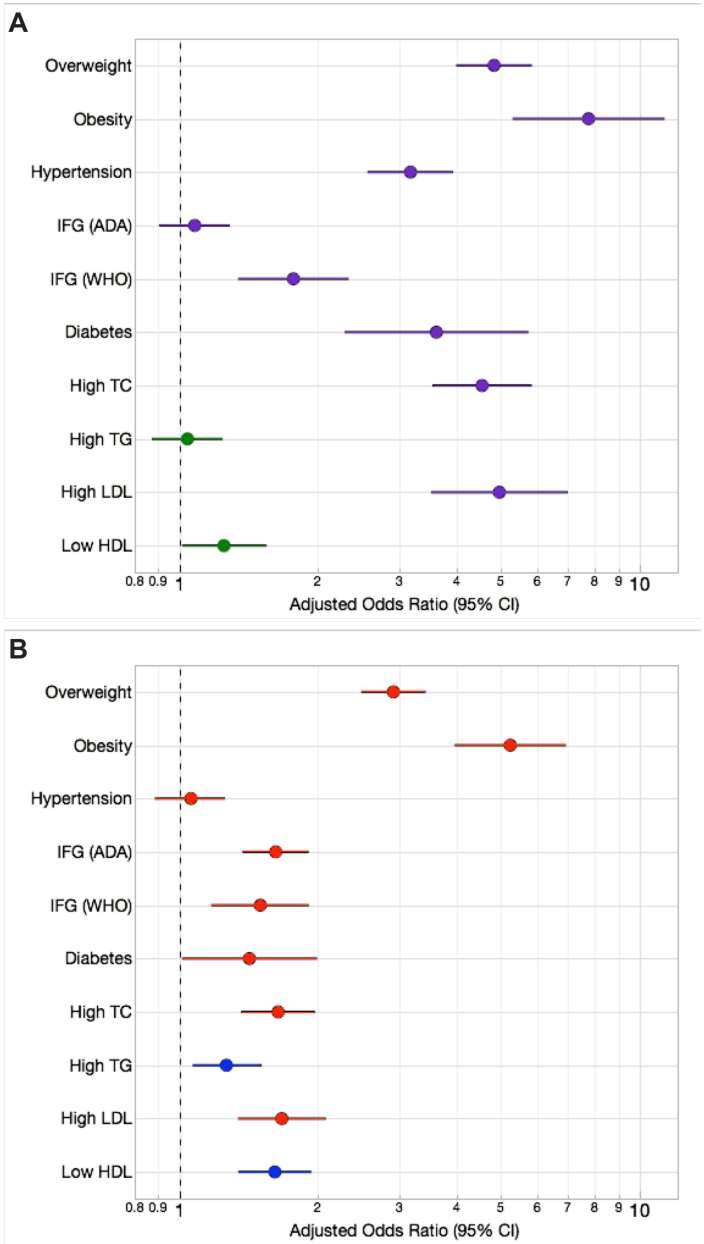
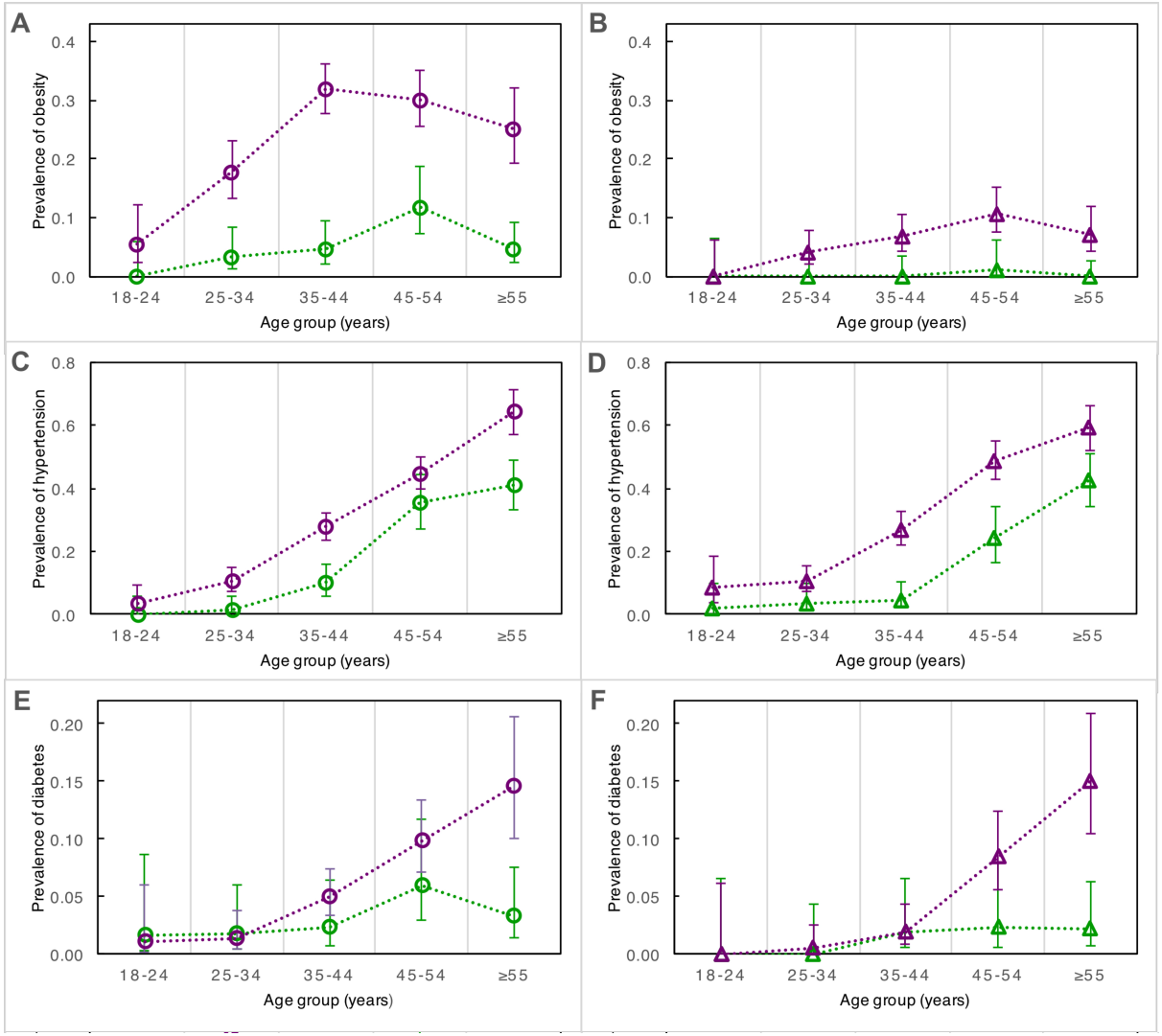

B
BMI
SBP
DBP
GLUC
TC
TG
LDL
HDL
TC/HDL
t-PA
PAI-1
0
0.2
0.4
0.6
0.8
1
Standard Mean Difference (95% CI)

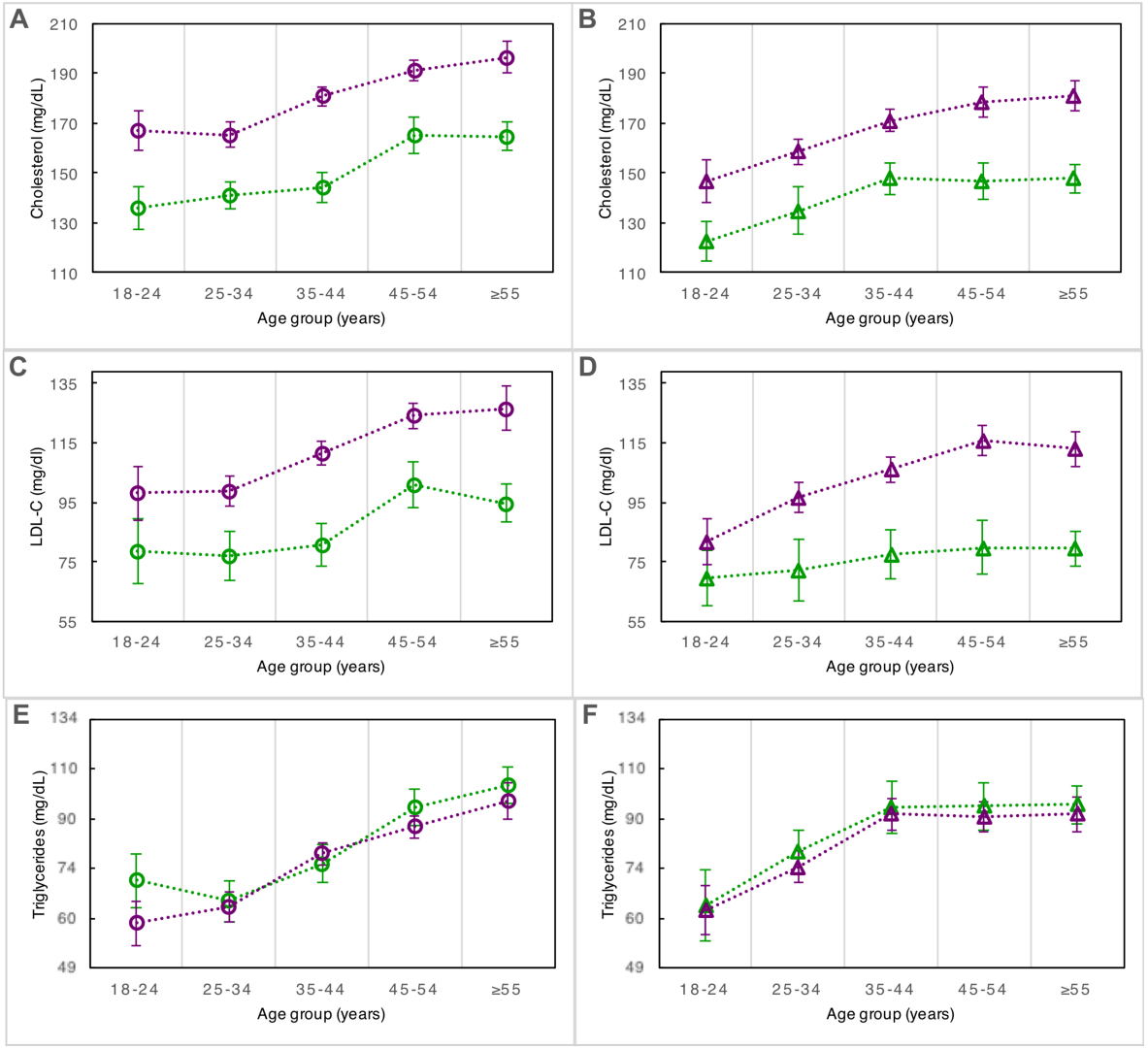
Figure 3-3. The effect of urban/rural environment and sex on dichotomous cardiovascular risk factors in Sunyani, Ghana. (A) The increased odds of each outcome (with 95% confidence intervals) are depicted for the group with the higher odds (urban: purple; rural: green). Data were adjusted for age and sex. (B) The increased odds of each outcome (with 95% confidence intervals) are depicted for the group with the higher odds (female: red; male: blue). Data were adjusted for age and environment.


A
Overweight
Obesity
Hypertension
IFG (ADA)
IFG (WHO)
Diabetes
High TC
High TG
High LDL
Low HDL
0.8 0.9 1 2 3 4 5 6 7 8 9 10
Adjusted Odds Ratio (95% CI)
B
Overweight
Obesity
Hypertension
IFG (ADA)
IFG (WHO)
Diabetes
High TC
High TG
High LDL
Low HDL
0.8 0.9 1 2 3 4 5 6 7 8 9 10
Adjusted Odds Ratio (95% CI)

<u>Figure 3-4</u>. Prevalence by age group of obesity, hypertension, and diabetes in urban and rural men and women from the Sunyani region of Ghana. Females (circles) are depicted in the left panels (A), (C), and (E); males (triangles) in the right panels (B), (D), and (F). Purple = urban; green = rural. Error bars denote 95% confidence intervals of estimates.



<u>Figure 3-5</u>. Mean lipid levels by age group in urban and rural men and women from the Sunyani region of Ghana. Females (circles) are depicted in the left panels (A), (C), and (E); males (triangles) in the right panels (B), (D), and (F). Purple = urban, green = rural. Error bars denote 95% confidence intervals of estimates. Note: in (E) and (F), vertical axis is logarithmic.



PAI-1 and the Risk Factors of the Metabolic Syndrome

Introduction

The metabolic syndrome (MetS) is a set of cardiometabolic abnormalities that tend to co-occur in people at increased risk for coronary heart disease and type 2 diabetes. Although MetS is a complex disorder with no single factor for a cause, obesity typically plays a prominent role in its etiology, particularly in the form of metabolically active visceral fat.[142,143] Less clear are the mechanisms by which adipocytes accelerate conditions such as insulin resistance, hypertension, and dyslipidemia, and how they, in turn, aggravate atheromatous degeneration and promote acute phase cardiovascular disease.[140,141]

Given the uncertainty of its etiology and the still-evolving nature of its definition, the usefulness of MetS as a diagnosis, from both a clinical and an epidemiologic standpoint, has been questioned.[144,145] Recent debate has centered on whether the increased cardiovascular risk associated with MetS can be wholly accounted for by the additive contributions of its component risk factors.[146-148]If so, it has been argued that the additional insight gained by studying the risk factors together rather than in isolation may be limited. However, quite apart from the question of whether MetS increases cardiovascular risk additively or exponentially is the question of why its component conditions co-occur in the first place. Elucidating the mechanisms of their co-occurrence, including the environmental, behavioral, and genetic factors that give rise to them, will require a collective assessment.[149-151]

MetS may also be distinguishable from isolated cases of its component conditions by a supra-normal level of clotting and anti-fibrinolytic factors, which generates a hypercoagulable state in the blood. [152,153]Among the most prominent of these factors is plasminogen activator inhibitor-1 (PAI-1), which impairs the degradation of clots by inhibiting tissue-type plasminogen

activator (t-PA), a major thrombolytic enzyme in the fibrinolytic pathway.[26] Clinical evidence suggests that spikes in PAI-1 trigger ischemic events,[154] and epidemiologic studies have convincingly demonstrated a positive association between elevated plasma PAI-1 levels and cardiovascular risk.[26,155,156] Because MetS not only accelerates atherosclerosis, but also increases the risk of thrombosis and subsequent thrombotic events, PAI-1 may be a natural link between the syndrome and acute phases of cardiovascular disease.[157,158]

In fact, the role of PAI-1 in MetS likely goes beyond that of inhibiting fibrinolysis. It is a highly pleiotropic enzyme that predisposes patients to premature atherosclerosis by interfering with cell migration and promoting a chronic state of low-grade inflammation.[159,160] Many of the component conditions of MetS increase PAI-1 gene expression and appear to be influenced by PAI-1 expression in turn.[157,158] Adipocytes release PAI-1 directly, while some evidence indicates that PAI-1 itself may promote the accumulation of visceral fat.[161] Adipocytokines also increase PAI-1 expression indirectly, such as via inflammatory pathways, making elevated plasma PAI-1 one of the hallmarks of obesity.[162] By mediating insulin signaling, PAI-1 may also provide a biochemical link between obesity and insulin resistance, the two most salient features of MetS.[138,163,164] Finally, insulin resistance itself increases PAI-1 expression by accelerating lipolysis and releasing an excess of free fatty acids into the blood.[165]

The PAI-1 promoter is also responsive to several other metabolic and endocrine factors, including very low-density lipoprotein (VLDL), which likely explains the association between PAI-1 and hypertriglyceridemia.[166] A similar connection exists between PAI-1 and the renin-angiotensin system, involved in hypertension[167]. Yet, despite the fact that no clinical measurement appears to be more strongly associated with the components of MetS than PAI-1,[149,152]only a few large epidemiological studies have assessed the relationship between PAI-1

and multiple CVD risk factors in a systematic way. Moreover, all have done so using samples primarily from populations of European descent.[149,168,169]

Here, we present a multivariate analysis of MetS and PAI-1, using cardiovascular data from 3331 men and women in Ghana from both urban and rural locales. We assess the correlational architecture of MetS risk factors, and estimate risk factor contributions to thrombotic endpoints, using intensity and independence of association with PAI-1 as a proxy. We also evaluate whether conditions related to sex and urbanization influence these relationships.

Materials and Methods

Study Cohort Description

Please see Section A of this chapter.

Anthropometric measurements and biochemical analysis

Height, weight, systolic and diastolic blood pressure, fasting plasma lipids, glucose, and plasma PAI-1 were measured as described in Section A of this chapter.

Study variables

Five categorical metabolic risk factors (hypertriglyceridemia, low HDL, hypertension, hyperglycemia, and obesity) were defined according to the updated National Cholesterol Education Program Adult Treatment Panel-III (NCEP ATP-III) criteria,[171] as follows: triglycerides (TG) ≥ 150 mg/dl; high-density lipoprotein cholesterol (HDL) <40mg/dl in males or <50 mg/dl in females; systolic (SBP) and diastolic (DBP) blood pressure ≥ 130/85 mm Hg or on anti-hypertensive medication; fasting glucose (GLUC) ≥ 100 mg/dl or on antidiabetic medication; and body mass index (BMI) ≥ 30. MetS was defined as the presence of three of those five conditions.[141] None of the study participants reported taking statins. Mean arterial pressure (MAP) was calculated using the formula: MAP = DBP + [(SBP-DBP)/3], which approximates the average arterial pressure during a single cardiac cycle. MAP was used in the correlational analyses instead of SBP and DBP to maintain a one-to-one correspondence between quantitative traits and the components of MetS. However, it is worth noting that MAP has been

shown to predict future metabolic syndrome more accurately than SBP, DBP, or pulse pressure.[172]

Statistical methods

Summary data of the quantitative risk factors used in this study are presented in Section A of this chapter. For calculating the prevalence of MetS by sex and urban/rural setting ("residence"), study participants for whom no data was missing (N=2220) and who had at least three of the five conditions (as per the NCEP ATP-III guidelines above) were deemed cases. Prevalence rates of MetS were age-standardized to the World Health Organization (WHO) standard population (see Section A of this chapter). Relative risks of MetS (urban vs. rural; female vs. male) were also estimated, with 95% confidence intervals derived using the anti-log of the formula

$$\log(p_1/p_2) \pm z_{\alpha/2}\sqrt{\frac{1-p_1}{n_1 p_1} + \frac{1-p_2}{n_2 p_2}}$$

Where n_i is the sample size for group i, p_i the sample probability of MetS, and $z_{\alpha/2}$ the critical z-value, 1.96, for α=0.05.

Log-transformation improved the approximations to normality of all of the quantitative variables in this study (among which only PAI-1, glucose, and triglycerides were highly skewed). Therefore, because statistical tests of Pearson's product-moment correlations can be sensitive to violations of normality,[173,174] all data were log transformed for the correlational analyses. For clarity of presentation, however, references in the text to the variables are not emended to reflect these transformations (e.g., as "ln-glucose," "ln-PAI-1" etc.).

All pairwise correlations between BMI, MAP, GLUC, TG, HDL, and PAI-1 were calculated after each measurement was adjusted for age, sex, and residence. The same pairwise correlations were calculated for (1) subjects stratified by residence, with variables adjusted only

for age and sex; and for (2) subjects stratified by sex, with variables adjusted only for age and residence. The homogeneity of correlation between groups (urban vs. rural, male vs. female) was then assessed by *t*-test after Fisher transformation of correlation coefficients.

Partial correlation measures the strength of association between two variables after controlling for a set of other variables. Partial correlations were calculated for every pair of variables in the set of BMI, MAP, GLUC, TG, HDL, and PAI-1 after controlling for the remaining variables in the set (as well as age, sex, and residence, as above). These pairwise correlations were also calculated separately for subjects first stratified by sex and subjects first stratified by residence, as above, for evidence of heterogeneity of correlation.

The following approach was used to assess visually whether the strengths of association between MetS traits (BMI, MAP, GLUC, TG, and HDL) and PAI-1 were consistent over the entirety of their respective ranges, and to identify possible patterns of non-linear association therein. First, all six variables were adjusted for age, sex, and residence, and the residuals were standardized (variable names below refer to the standardized residuals). For each of the five MetS traits, values were ranked in ascending order and paired with their corresponding PAI-1 values. The 25th percentile, median, and 75th percentile of PAI-1 (period=100) was then plotted against the corresponding median (period=100) of each MetS trait.

To capture the global features of these relationships, smooth curves of the plots were created using a cubic spline method. Briefly, for *n* observations, where $\tilde{x}_i$ is the i^{th} standardized median (period 100) of a risk factor, $i \in [1, n-99]$, and $\tilde{P}_i$ the corresponding quantile value of PAI-1, such that $\tilde{P}_i = \mu(\tilde{x}_i)$, the smoothing function $\hat{\mu}$ estimated μ by minimizing

$$\sum_{i=1}^{n-99} \left(\tilde{P}_i - \hat{\mu}(\tilde{x}_i) \right)^2 + \lambda \int_{x_1}^{x_2} \hat{\mu}''(\tilde{x})^2 d\tilde{x}$$

The first term represents the sum of squares error and the second term the penalty for "roughness." The parameter λ controls the bias-variance tradeoff, and was set to 10.[175]

To evaluate how severity of MetS influences PAI-1 levels, two approaches were taken. First, an ordinal measure of severity of MetS was defined as the number of component conditions with which an individual was diagnosed; the effect on mean PAI-1 was then assessed. Next, a continuous measure of severity of MetS was defined as the first principal component (PC) of the five quantitative risk factors, TG, HDL, MAP, GLUC, and BMI. The moving medians of the first three PCs versus the corresponding moving medians of PAI-1 were then graphically depicted, using the method described above.

Statistical analyses were performed using JMP (version 11) and STATA (version 12).

Results

Of the 3331 participants in this study, 2276 (68%) were urban residents, of whom 1298 (57%) were female and 978 male. Of the 1055 rural residents, 583 (55%) were female and 472 male. Ages ranged from 18-99 and were similarly distributed among urban and rural men and women (p=0.23, Kruskal-Wallis test). For analyses based on diagnoses of dichotomous risk factors associated with MetS (hypertriglyceridemia, low HDL, hypertension, hyperglycemia, and obesity), participants with missing data were excluded, lowering sample size to 2220 (see **Table S2** in Appendix A for the breakdown by group).

Urban residence increased the relative risk (RR) of MetS for both men (1.61; 95% CI: 1.02-2.53) and women (1.72; 95% CI: 1.28-2.32). Urban women, who had by far the highest prevalence of obesity among all groups (**Table S2**), also had a significantly higher risk of MetS than urban men (RR: 1.68; 95% CI: 1.36-2.07). Among rural participants, risk did not differ significantly by sex. The age-standardized prevalences of MetS and the distribution of MetS risk factors are shown by group in **Table S2** and **Figure S6**.

Mean PAI-1 (log normalized and adjusted for age, sex, and residence) rose exponentially as the number of MetS risk factor conditions increased linearly. The quadratic fit to the means was virtually perfect ($R^2 = 0.996$) (**Figure 3-6**). PAI-1 was normally distributed within all five categories (Shapiro-Wilk test; $p>0.05$ for all). Participants diagnosed with ≤1 MetS risk factor did not have significantly different mean PAI-1 levels, and the 95% confidence intervals for both groups fell below the mean PAI-1 level for the study cohort as whole. In contrast, among participants who had two or more diagnoses of MetS risk factors, mean PAI-1 increased significantly with each incremental diagnosis, and 95% confidence intervals were above the mean PAI-1 level for the study cohort as whole (**Figure 3-6**).

All pairwise correlations between quantitative risk factors associated with MetS (adjusted for age, sex, and residence) were highly significant ($p<0.0001$), except for the correlation between MAP and HDL ($p=0.14$) (**Table 3-2** and **Figure 3-7A**). Despite being statistically significant, however, most of the correlations were relatively weak; those greater than 0.30 were between PAI-1 and BMI ($r=0.43$; 95% CI 0.40-0.45); PAI-1 and TG ($r=0.35$; 95% CI 0.32-0.38); and BMI and MAP ($r=0.30$; 95% CI 0.27-0.33). The strongest negative correlation was between TG and HDL (r=0.-27), with 95% confidence interval (-0.31, -0.23) (**Table 3-2** and **Figure 3-7A**).

To assess whether urban residence had an effect on these correlations, the above analysis was repeated after stratifying participants by residence and adjusting for age and sex. The pairwise correlations between the quantitative risk factors that define MetS (BMI, MAP, HDL, TG, and GLUC) were similar for both urban and rural populations. Tests assessing heterogeneity of correlation were significant only for BMI-HDL ($p=0.047$) and BMI-TG ($p=0.012$) (**Table 3-3** and **Figure 3-7B**). However, three of the five correlations between PAI-1 and MetS risk factors exhibited highly significant heterogeneity: PAI-1 and BMI ($p=4.9\times10^{-9}$), PAI-1 and GLUC ($p=9.9\times10^{-4}$), and PAI-1 and TG ($p=2.1\times10^{-3}$) (**Table 3-3** and **Figure 3-7B**). In all these cases, the correlation between risk factors was stronger in the urban population (**Table 3-3**). When participants were stratified by sex, no p-value for heterogeneity of correlation was below 0.01 (**Table S3** and **Figure 3-7B**). Two of the three comparisons that were significant at the 0.05 level featured PAI-1; the correlation between PAI-1 and MAP was stronger in men (0.27 vs. 0.20; $p=0.024$), whereas the correlation between PAI-1 and glucose was stronger in women (0.23 vs. 0.15; $p=0.012$) (**Table S3**).

Partial correlations were also calculated to quantify the strength of association between every pair of MetS-related risk factors independent of the other risk factors. Although all partial correlations were statistically significant (the majority at $p<0.0001$), only four pairs of partial correlations were greater than 0.15 in magnitude: PAI-1 and BMI, r=0.33 (0.29, 0.37); PAI-1 and TG, r=0.24 (0.20, 0.28); BMI and MAP, r=0.22 (0.18, 0.26); and TG and HDL, r= -0.22 (-0.26, -0.18) (**Table 3-4** and **Figure 3-8A**). When this analysis was repeated after stratifying participants either by sex or residence (as above), only one of fifteen partial correlations was significantly different between sexes (PAI-1 and GLUC, p=0.023), and three of fifteen partial correlations were significantly different between urban and rural residents (PAI-1 and GLUC, p=0.001; PAI-1 and BMI, p=0.0002; and BMI and GLUC, p=0.0054) (**Table 3-5**, **Table S4**, and **Figure 3-8B**). The partial correlational analyses were repeated for MetS risk factors without PAI-1, and yielded comparable results (**Table S5**, **S6**, and **S7**).

To complement the correlational analyses, which can mask non-linear or non-continuous patterns of association between variables, the change in median PAI-1 was assessed over "sliding windows" of the other risk factors. All variables were first adjusted for age, sex, and residence, and standardized (values reported below are Z-scores). Over the entire ranges of both BMI and triglycerides, median PAI-1 rose from about -0.5 to 1.0. However, for BMI less than one standard deviation below its mean, median PAI-1 did not change (**Figure 3-9A**). The relationship between PAI-1 quartiles and glucose displayed the most abrupt shift in association patterns: PAI-1 rose rapidly until glucose reached ~1.5 standard deviations above its mean, after which it was flat. Thus, the relatively low correlation between glucose and PAI-1 reported above (r=0.20) likely reflects the composite effects of two entirely different patterns of association, one strong, one weak, with the shift occurring at the far right tail of glucose values (**Figure 3-9A**).

The moving quartiles of PAI-1 were also evaluated for data stratified by sex and residence. The discordant behavior of PAI-1 at the right tails of the MAP and glucose distributions likely explains the male/female heterogeneity of correlation between those variables (**Figure 3-10A**). Overall, however, the male and female trajectories of PAI-1 quartiles were almost indistinguishable, regardless of risk factor (**Figure 3-10A**). In contrast, the moving quartiles were noticeably more discordant among urban and rural populations (**Figure 3-11A**). Median PAI-1 barely increased in the rural population when both BMI and glucose values below their respective means, median. The upper range of glucose values was also much greater in the urban population, likely driving the observed heterogeneity of correlation as well (**Figure 3-11A**).

The above analyses were also carried out for PAI-1 and each MetS risk factor using the standardized residuals after adjustment for all other MetS risk factors, providing insight into the partial correlations reported above. Whereas the partial correlations between PAI-1 and MAP, PAI-1 and HDL, and PAI-1 and glucose were all relatively weak in magnitude ($r<0.10$) and not significantly different from each other (**Table 3-4**), PAI-1 quartiles appeared to be more sensitive to glucose than MAP or HDL (**Figure 3-9B**). When participants were stratified by sex, there was a strong correspondence between male and female moving PAI-1 quartiles (**Figure 3-10B**). A similar correspondence was observed between the urban and rural PAI-1 quartiles for MAP, HDL, and TG. However, there were pronounced urban/rural differences for BMI and glucose (**Figure 3-11B**), explaining the heterogeneity of correlation observed therein.

Discussion

It is well known that the cardiovascular risk factors associated with the metabolic syndrome tend to cluster, and in so doing increase the risk of coronary heart disease and stroke. Why they cluster, however, and the pathophysiological mechanisms by which their co-occurrence increases ischemic risk, are not well understood. To what extent is the clustering of risk factors the effect of causal relationships among the risk factors themselves? Do factors such as ancestry, sex, and urban lifestyle, known to influence cardiovascular risk factors individually, also influence the patterns of association among them? Do co-occurring risk factors have joint effects on ischemic risk greater than the sum of their individual contributions, and if so, by what biochemical means? These and related questions motivated our present study.

Apart from cardiometabolic abnormalities, one of the hallmarks of the metabolic syndrome is a prothrombotic state, characterized by elevated plasma levels of the anti-fibrinolytic enzyme, PAI-1. Because PAI-1 plays a direct biochemical role in thrombosis, a risk factor's relevance to the thrombotic stages of MetS strength of correlation with PAI-1 may be indicative of its. We found that the correlations between PAI-1 and the continuous risk factors used to define MetS (BMI, MAP, HDL, TG, and GLUC) were stronger, on average, than the correlations between the risk factors themselves. For example, BMI, TG, and GLUC were all most strongly correlated with PAI-1. Given that MetS is essentially a descriptive term for the stochastically improbable correlation of five conventional risk factors, their consistently strong correlations with PAI-1 (regardless of sex or residence) raises the question of whether PAI-1 should also be considered a definitional component of the syndrome.

Whereas the distributions and mean values of the risk factors assessed in this study typically varied significantly with sex and/or residence, only a few of the pairwise correlations

among them did. The correlations with PAI-1 were the main exceptions; those with BMI, TG, and GLUC were markedly stronger in the urban population. To the extent that elevated PAI-1 increases the risk of thrombotic endpoints, abnormally high BMI, TG, and GLUC may therefore confer greater risk in urban than in rural environments. In the same way, hypertension may pose a greater risk in Ghanaian men, and hyperglycemia in women, owing to the significantly stronger MAP-PAI-1 and GLUC-PAI-1 correlations in men and women respectively.

A deeper understanding of the etiology and hierarchical architecture of MetS requires that we distinguish direct interactions among risk factors from merely incidental associations. Crude correlations among risk factors can provide insight into the topology of risk factor networks, but because the correlation between two risk factors may be driven entirely by their mutual association with other risk factors, such insight is limited. On the other hand, a strong partial correlation between two risk factors, independent of others, would indicate that they either cause each other directly, or share a "private" set of causal factors. We therefore also assessed the partial correlations between each pair of MetS risk factors conditioned on all others. Partial correlational analysis is more appropriate here than multivariate regression, because designating variables as "independent" and "dependent" would fail to take into account the dynamic relationships and feedback loops characteristic of metabolic systems.

The statistical significance of all partial correlations (in analyses both with and without PAI-1) indicated that the relationships between all pairs of risk factors were partly independent. In general, PAI-1 had the strongest such relationships, and these larger partial correlations likely reflect known biochemical and physiological connections. For example, the strongest relationship was between PAI-1 and BMI ($r=0.33$), consistent with the fact that adipose tissue releases PAI-1. The second strongest was between PAI-1 and TG ($r=0.24$), likely influenced by

the fact that the PAI-1 promoter is responsive to VLDL. The only risk factor for which there is no evidence of a direct biochemical link with PAI-1, HDL, had the weakest partial correlation with PAI-1, with a 95% confidence interval between -0.09 and -0.01. We emphasize that independent associations with PAI-1 may be particularly indicative of a risk factor's proximate relevance to thrombosis.

When PAI-1 was excluded from analyses, there was no heterogeneity of partial correlation by either sex or residence for any of the pairs of MetS risk factors. Thus, to whatever extent these risk factors have direct causal relationships with each other, they appear to be consistent over a wide range of values, physiological backgrounds and environmental exposures. In contrast, partial correlations with PAI-1 were more likely to exhibit heterogeneity by sex and, particularly, residence. Although the p-values were consistently larger, the patterns of heterogeneity were similar to those observed for the crude correlations (i.e. correlations adjusted only for age and sex/residence). The reduced significance could be partly due to the smaller sample sizes (by ~1/3) in these analyses, since the higher order partial correlations required full data.

Obesity is generally considered the primary causal component of MetS,[140] but the partial correlations between BMI and the other continuous risk factors that define MetS were unexpectedly weak. Cause cannot be inferred from correlation, but it can, in a sense, be ruled out; if adiposity alone had strong independent effects on the other risk factors, the partial correlations with BMI should have been stronger. Thus, the connection between obesity and MetS may not be as straightforward as commonly accepted, and likely involves multiple simultaneous factors of susceptibility. The relationship between BMI and MAP may be somewhat of an exception, as the partial correlation was similar to the crude correlation. (The

association between TG and HDL also appeared to be similarly unrelated to other risk factors). The relatively weak partial correlations with BMI were surprising, because the connection between obesity and MetS in our study was evident in the small percentage (<1%) of participants who had isolated obesity, i.e. obesity in the absence of any other risk factor conditions. Moreover, among the 764 participants who had only a single risk factor condition, that condition was obesity for only 2.7%. In contrast, 27% had hypertension, and 55% low HDL. Also to expectations, given that the behavioral changes associated with urbanization are major factors in the emerging global obesity epidemic, we found that the urban population in our study was at significantly greater risk of MetS than the rural, and that urban women in particular, who had by far the highest prevalence of obesity, were at greater risk of MetS than urban men.

On the other hand, the age-standardized prevalence of MetS among rural men (7.8%) was unexpectedly high, insofar as no participant had a BMI above the obesity-threshold of 30 kg/m^2. While "metabolically obese" individuals of normal weight are not uncommon,[177] often MetS without obesity reflects insulin resistance caused by means other than adipose tissue.[178] However, 40% of the rural men with MetS in our study did not have hyperglycemia in addition to hypertension and/or dyslipidemia (low HDL or high TG), making it unlikely that they were insulin resistant. MetS in the absence of obesity can also be caused by irregularly distributed adipose tissue, as when an excess of visceral fat is masked by waist circumference in the normal range,[179,180] but there is no evidence for this phenomenon among the rural Ghanaian men with MetS. Thus, how comparable these rural participants are to others diagnosed with MetS is unclear, either from the standpoint of either pathophysiology or clinical prognosis.

For any two variables, the correlation coefficient is the expected change in either variable as the other increases by one standard deviation. Because the correlation coefficient extracts a

single, linear relationship from bivariate data, it becomes less informative as the true relationship between those variables displays patterns of non-linearity or non-continuity. However, such patterns can be a particularly significant concern in the study of clinical conditions, such as MetS, for which risk often pertains only to the uppermost quantiles of variables (e.g., if ischemic risk is amplified when PAI-1 exceeds a certain threshold). We therefore complemented our correlational analyses by graphically depicting the moving quartiles of PAI-1 over increasing values of MetS risk factors. The rationale for centering this analysis on PAI-1 quartiles rather than mean values was, first, to minimize the influence of relative outliers caused by the characteristic kurtosis of the PAI-1 distribution, and second, to gain insight into whether PAI-1 levels in individuals at upper or lower quartiles responds differently to changes in the other risk factors. To our knowledge, this question, which can have clinical relevance, has not been previously addressed.

Interestingly, although the correlation between PAI-1 and BMI was significantly greater than that for TG, the increase in PAI-1 quartiles with TG appeared to be more consistent and, on average, slightly greater in slope. In contrast, the median or 75^{th} percentile of PAI-1 did not increase at all when BMI was less than -1σ, and began to decrease in slope when BMI was greater than 1σ. Similarly, median PAI-1 increased rapidly with glucose over most of its range, but stopped increasing as glucose approached 2σ. Because many glucose values exceeded 2σ (even after transformation and adjustment), the near-zero slope of median PAI-1 over that range likely had an outsized effect on the unexpectedly low GLUC-PAI-1 correlation. In theory, correlations can be attenuated not only by a weakening of the median PAI-1 slope but also by a greater dispersion of values around median PAI-1. In that respect, the relatively weak correlation

between PAI-1 and MAP (0.23) was likely influenced by the notable increase in the interquartile range of PAI-1 as MAP increased above 1σ.

But correlation already said. The male and female trajectories of PAI-1 quartiles by risk factor were practically indistinguishable, indicating that the relationships between these traits are robust to a wide range of physiological differences. However, the two heterogeneities of correlation by sex that we did observe (GLUC-PAI-1 and MAP-PAI-1) clearly reflecting a tight relationship at the right tales of the appeared to be driven by shifts at the right tails of the GLUC and MAP distributions. As noted above, this may be particularly relevant to the etiology of MetS.

The urban and rural moving quartiles of PAI-1 with BMI and with GLUC were extremely discordant, as expected, given the highly significant heterogeneities of correlation observed for those pairs of risk factors. Glucose had a much wider range of values in the urban population (despite standardization after stratification and age- and sex-adjustment), which likely influenced the urban-rural GLUC-PAI-1 heterogeneity of correlation to some extent. However, a pronounced difference in slopes between urban-rural quartiles of PAI-1 was also evident for glucose values below the mean.

The moving quartile analyses using the residuals of the partial correlations showed that the strongest independent relationships involving PAI-1 were with BMI, TG, and GLUC. That between PAI-1 and TG was particularly strong for TG values greater than 1σ. Importantly, all independent relationships with PAI-1 were relatively weak when risk factors were less than approximately -1σ.

While we could not find similar studies with which to compare our higher order partial correlations, a recent meta-analysis of 85,000 people (of whom 7.8% were African American) reported pairwise correlations between MetS risk factors adjusted for age and sex.[149] Among all of our pairwise correlations, only BMI-GLUC and BMI-HDL here (adjusted for age, sex, and residence) fell outside the 95% confidence intervals reported by the meta-analysis. That study reported r= -0.33 for BMI-HDL and r= 0.28 for BMI-GLUC, in contrast to our results of r= -0.15 and r= 0.15, respectively. Why these two pairs of risk factors were anomalous in our study population, even among urban participants only, is not clear. However, many Ghanaians had low HDL in general (see Section A of this chapter), and the plot of the moving HDL median vs. BMI confirmed that low HDL levels are common in Ghana even with healthy BMI. More specifically, the observed (and expected) negative relationship between HDL levels and BMI disappeared when BMI was less than one standard deviation below its mean, as the interquartile range expanded. These trends likely weakened the correlation. The moving quartiles of GLUC vs. BMI, on the other hand, increased linearly throughout their entire ranges, making the weak correlation between BMI and GLUC in Ghanaians more enigmatic.

We also examined the effect that the clustering of risk factor conditions had on PAI-1 levels. We found that mean PAI-1 levels increased exponentially with the number of conditions present (regardless of which). In fact, the quadratic fit to the means was virtually perfect (R^2=0.996). The greatest increase in mean PAI-1 occurred as the number of risk factor conditions increased from two to three, reflecting an intimate association between PAI-1 and MetS (as defined by NCEP ATP-III). With each incremental increase in the number of conditions over one, mean PAI-1 increased significantly. In contrast, mean PAI-1 was not significantly different

between participants with one risk factor or none, and was below the population mean for both groups.

A much-discussed topic has been whether the clustering of risk factors associated with MetS increases total ischemic risk in an additive or exponential way.[181-184] However, if mean PAI-1 increases exponentially as the number of risk factors increases linearly, as observed here, then the question of whether MetS is "more than the sum of its parts" rests to a large degree on the clinical consequences of elevated PAI-1. In fact, both clinical and epidemiologic studies have convincingly demonstrated a connection between elevated PAI-1 levels and increased cardiovascular risk, as discussed above. The observation here that PAI-1 rises in an exponential manner in patients most at risk for ischemic events also supports our recent recommendation that future epidemiologic studies of PAI-1 consider the risk profiles of patients in the upper quartile of the distribution.[185]

Because the NCEP ATP-III definition of MetS is combinatorial, such that no risk factor is necessary and all are interchangeable, MetS can naturally be transformed from a binary variable into a continuous one by adding the Z-scores of the continuous risk factors underlying the condition. While several studies have adopted this approach,[186-188] we found that using the first principal component (PC1) of the five risk factors offers important advantages. The two approaches were, in fact, very similar (leading to values correlated at r= 0.82 in men and r=0.80 in women), because the loadings of all five risk factors for PC1 were relatively equal. However, using the first principal component has the additional merit of extracting the most information from the dataset (by definition). Moreover, because PC1 is uncorrelated to subsequent PCs, the loadings of the subsequent PCs reveal (in order of importance) the various combinations of risk factors that most differentiate the population with respect to how MetS (defined continuously as

PC1) is composed. We see, for example, that our study participants differed most with respect to the relative contributions of BMI and MAP vs. TG and (low) HDL, or more intuitively, obesity/hypertension versus dyslipidemia (see PC2).

Confirming the strong connection between PAI-1 and MetS. we observed that the relationship between PC1 and PAI-1 was extremely strong. Specifically, their correlation was 0.56, and PAI-1 quartiles increased by more than 2 standard deviations over the range of PC1. Notably, the relationship between median PAI-1 and PC1 was practically identical in men and women, and median PAI-1 did not vary at all with PC2 or PC3, signifying that PAI-1 increases with MetS *per se*, regardless of the physiological background or specific risk factor composition involved.

Conclusion

We have explored the role of PAI-1 in the metabolic syndrome from multiple angles. Although we have previously published and also report here that urban residence and, to a lesser extent, sex have dramatic effects on the mean values of cardiovascular risk factors, our correlational analyses here reveal that the *relationships* among the risk factors remain remarkably robust. The relationships between risk factors and PAI-1, however, appear to be far more sensitive to differences in sex and environment. It will be interesting to see if the patterns we have identified here, such as the exponential relationship between mean PAI-1 and MetS diagnoses, and the non-linear relationships between PAI-1 and some of the MetS risk factors, are replicable and generalizable to other populations.

Figure 3-6. Mean PAI-1 concentrations (and 95% confidence intervals) by the number of components of the metabolic syndrome. PAI-1 levels were adjusted for age, sex, and urban/rural residence, then standardized**.** The best quadratic fit to the means is also depicted (R-square = 0.996).

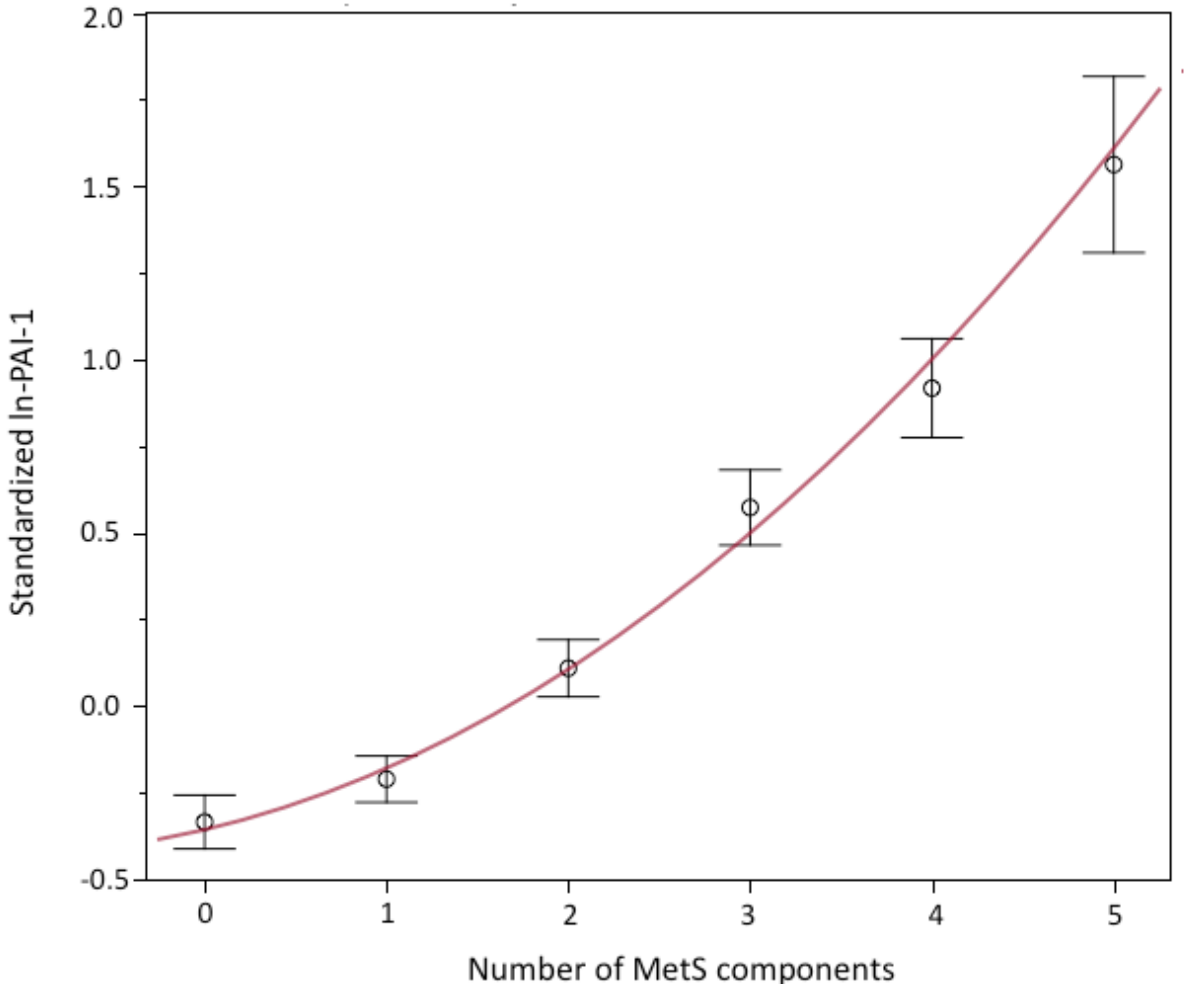


Table 3-2. Pairwise correlations between cardiovascular risk factors associated with the metabolic syndrome.

Trait 1	Trait 2	*r*	CI	N	p-value
BMI	PAI-1	0.43	(0.40, 0.45)	3331	<.0001
MAP	PAI-1	0.23	(0.20, 0.26)	3331	<.0001
HDL	PAI-1	-0.17	(-0.21, -0.13)	2225	<.0001
TG	PAI-1	0.35	(0.32, 0.38)	3321	<.0001
GLUC	PAI-1	0.20	(0.16, 0.23)	3331	<.0001
BMI	MAP	0.30	(0.27, 0.33)	3331	<.0001
MAP	HDL	0.03	(-0.01, 0.07)	2225	0.1376
HDL	TG	-0.27	(-0.31, -0.23)	2220	<.0001
TG	GLUC	0.17	(0.14, 0.21)	3321	<.0001
BMI	HDL	-0.15	(-0.19, -0.11)	2225	<.0001
MAP	TG	0.16	(0.13, 0.20)	3321	<.0001
HDL	GLUC	-0.12	(-0.16, -0.08)	2225	<.0001
BMI	TG	0.25	(0.22, 0.28)	3321	<.0001
MAP	GLUC	0.13	(0.10, 0.17)	3331	<.0001
BMI	GLUC	0.15	(0.12, 0.18)	3331	<.0001

r = Pearson correlation coefficient, calculated using residuals after adjustment for age, sex, and residence;
CI = 95% confidence interval;
p-value = probability of *r* if true correlation is zero;
Note: p-values > 0.05 have been grayed out.

Table 3-3. Pairwise correlations between cardiovascular risk factors associated with the metabolic syndrome, by urban or rural residence.

		Rural				Urban				Homogeneity of Correlation p-value
Trait 1	**Trait 2**	***r***	**CI**	**N**	**p-value**	***r***	**CI**	**N**	**p-value**	
BMI	**PAI-1**	0.29	(0.23, 0.34)	542	<.0001	0.47	(0.44, 0.51)	2276	<.0001	4.9E-09
MAP	**PAI-1**	0.24	(0.18, 0.29)	1055	<.0001	0.23	(0.19, 0.27)	2276	<.0001	0.7878
MAP	**BMI**	0.28	(0.22, 0.33)	1055	<.0001	0.31	(0.27, 0.34)	2276	<.0001	0.4186
HDL	**PAI-1**	-0.17	(-0.25, -0.08)	543	0.0001	-0.18	(-0.22, -0.13)	1682	<.0001	0.8302
HDL	**BMI**	-0.08	(-0.17, 0.00)	543	0.0522	-0.18	(-0.23, -0.13)	1682	<.0001	0.0472
HDL	**MAP**	0.05	(-0.03, 0.14)	543	0.2110	0.03	(-0.02, 0.07)	1682	0.2752	0.5818
TG	**PAI-1**	0.28	(0.22, 0.33)	1051	<.0001	0.38	(0.35, 0.42)	2270	<.0001	2.1E-03
TG	**BMI**	0.18	(0.13, 0.24)	1051	<.0001	0.27	(0.23, 0.31)	2270	<.0001	0.0122
TG	**MAP**	0.19	(0.13, 0.25)	1051	<.0001	0.15	(0.11, 0.19)	2270	<.0001	0.2561
TG	**HDL**	-0.30	(-0.37, -0.22)	542	<.0001	-0.26	(-0.31, -0.22)	1678	<.0001	0.4726
GLUC	**PAI-1**	0.11	(0.05, 0.17)	1055	0.0004	0.23	(0.19, 0.27)	2276	<.0001	9.9E-04
GLUC	**BMI**	0.20	(0.14, 0.25)	1055	<.0001	0.13	(0.09, 0.17)	2276	<.0001	0.0621
GLUC	**MAP**	0.17	(0.11, 0.23)	1055	<.0001	0.11	(0.07, 0.16)	2276	<.0001	0.1416
GLUC	**HDL**	-0.14	(-0.22, -0.05)	543	0.0014	-0.12	(-0.16, -0.07)	1682	<.0001	0.6887
GLUC	**TG**	0.20	(0.14, 0.26)	1051	<.0001	0.16	(0.12, 0.20)	2270	<.0001	0.2878

r = Pearson correlation coefficient, calculated using residuals after adjustment for age and sex, by residence;
CI = 95% confidence interval;
p-value = probability of *r* if true correlation is zero;
Homogeneity of Correlation, p-value = probability of these data if true correlation is equal for urban & rural populations;
Note: p-values > 0.05 have been grayed out.

Table 3-4. Partial correlations between components of the metabolic syndrome, including PAI-1, for the 2220 study participants with no missing data.

Trait 1	Trait 2	*r*	CI	p-value
BMI	PAI-1	0.33	(0.29, 0.37)	<.0001
MAP	PAI-1	0.09	(0.05, 0.13)	<.0001
HDL	PAI-1	-0.05	(-0.09, -0.01)	0.0156
TG	PAI-1	0.24	(0.20, 0.28)	<.0001
GLUC	PAI-1	0.10	(0.06, 0.14)	<.0001
BMI	MAP	0.22	(0.18, 0.26)	<.0001
MAP	HDL	0.12	(0.08, 0.16)	<.0001
HDL	TG	-0.22	(-0.26, -0.18)	<.0001
TG	GLUC	0.08	(0.04, 0.12)	0.0001
BMI	HDL	-0.09	(-0.13, -0.04)	<.0001
MAP	TG	0.08	(0.04, 0.12)	<.0001
HDL	GLUC	-0.08	(-0.12, -0.04)	0.0003
BMI	TG	0.07	(0.03, 0.11)	0.0007
MAP	GLUC	0.08	(0.04, 0.12)	0.0002
BMI	GLUC	0.04	(0.00, 0.08)	0.0564

r = Pearson partial correlation coefficient, calculated using residuals after adjustment for age, sex, and residence;
CI = 95% confidence interval;
p-value = probability of *r* if true partial correlation is zero;
Note: p-values > 0.05 have been grayed out.

Table 3-5. Partial correlations between components of the metabolic syndrome, including PAI-1, by urban or rural residence.

		Rural				Urban				Homogeneity of Correlation p-value
Trait 1	**Trait 2**	***r***	**CI**	**N**	**p-value**	***r***	**CI**	**N**	**p-value**	
BMI	**PAI-1**	0.21	(0.12, 0.29)	542	<.0001	0.37	(0.33, 0.41)	1678	<.0001	0.0002
MAP	**PAI-1**	0.15	(0.07, 0.23)	542	0.0004	0.07	(0.03, 0.12)	1678	0.0024	0.1068
MAP	**BMI**	0.20	(0.12, 0.28)	542	<.0001	0.23	(0.19, 0.28)	1678	<.0001	0.4607
HDL	**PAI-1**	-0.11	(-0.20, -0.03)	542	0.0085	-0.03	(-0.08, 0.01)	1678	0.1626	0.1090
HDL	**BMI**	-0.03	(-0.12, 0.05)	542	0.4282	-0.12	(-0.16, -0.07)	1678	<.0001	0.1003
HDL	**MAP**	0.14	(0.05, 0.22)	542	0.0014	0.12	(0.07, 0.17)	1678	<.0001	0.7363
TG	**PAI-1**	0.18	(0.10, 0.26)	542	<.0001	0.26	(0.22, 0.31)	1678	<.0001	0.0810
TG	**BMI**	0.05	(-0.03, 0.14)	542	0.2241	0.07	(0.03, 0.12)	1678	0.0027	0.6733
TG	**MAP**	0.13	(0.04, 0.21)	542	0.0034	0.06	(0.02, 0.11)	1678	0.0094	0.2049
TG	**HDL**	-0.26	(-0.33, -0.18)	542	<.0001	-0.21	(-0.25, -0.16)	1678	<.0001	0.3053
GLUC	**PAI-1**	-0.01	(-0.09, 0.07)	542	0.8199	0.15	(0.10, 0.20)	1678	<.0001	0.0010
GLUC	**BMI**	0.13	(0.05, 0.21)	542	0.0023	-0.01	(-0.05, 0.04)	1678	0.7965	0.0054
GLUC	**MAP**	0.11	(0.02, 0.19)	542	0.0113	0.07	(0.02, 0.11)	1678	0.0058	0.3994
GLUC	**HDL**	-0.09	(-0.18, -0.01)	542	0.0299	-0.07	(-0.12, -0.02)	1678	0.0030	0.6723
GLUC	**TG**	0.12	(0.04, 0.20)	542	0.0046	0.06	(0.01, 0.11)	1678	0.0126	0.2171

r = Pearson partial correlation coefficient, calculated using residuals after adjustment for age and sex, by residence;
CI = 95% confidence interval;
p-value = probability of *r* if true partial correlation is zero;
Homogeneity of Correlation, p-value = probability of these data if true partial correlation is equal for urban & rural;
Note: p-values > 0.05 have been grayed out.

Figure 3-7. Heat maps of risk factor correlations and heterogeneity. (A) The strength of correlation between cardiovascular risk factors associated with the metabolic syndrome and (B) the significance of differences in correlation by sex (below diagonal) and urban/rural residence).

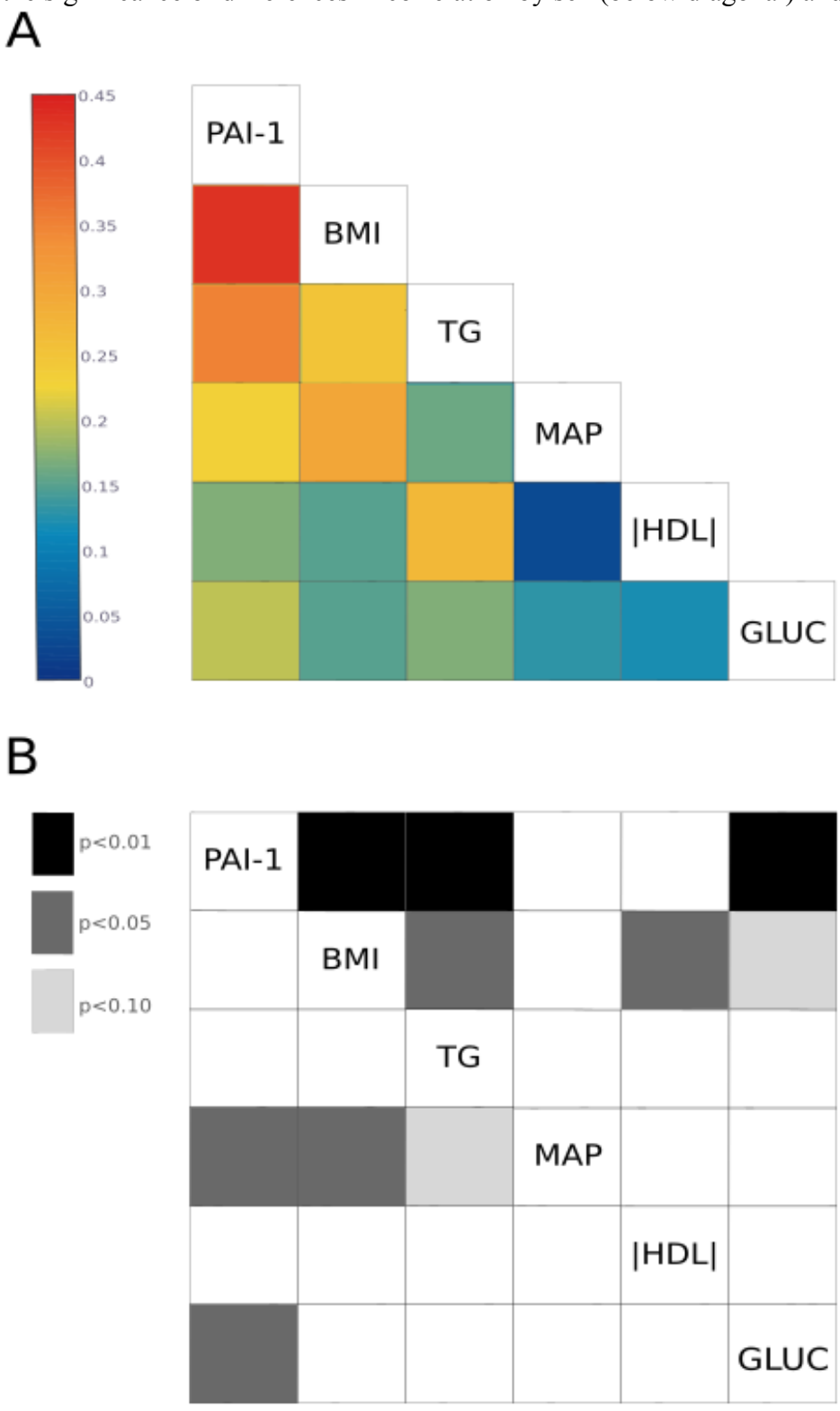


Figure 3-8. Heat maps of risk factor partial correlations and heterogeneity. (A) The strength of partial correlations between cardiovascular risk factors associated with the metabolic syndrome and (B) the significance of differences in partial correlation by sex (below diagonal) and urban/rural residence (above diagonal).

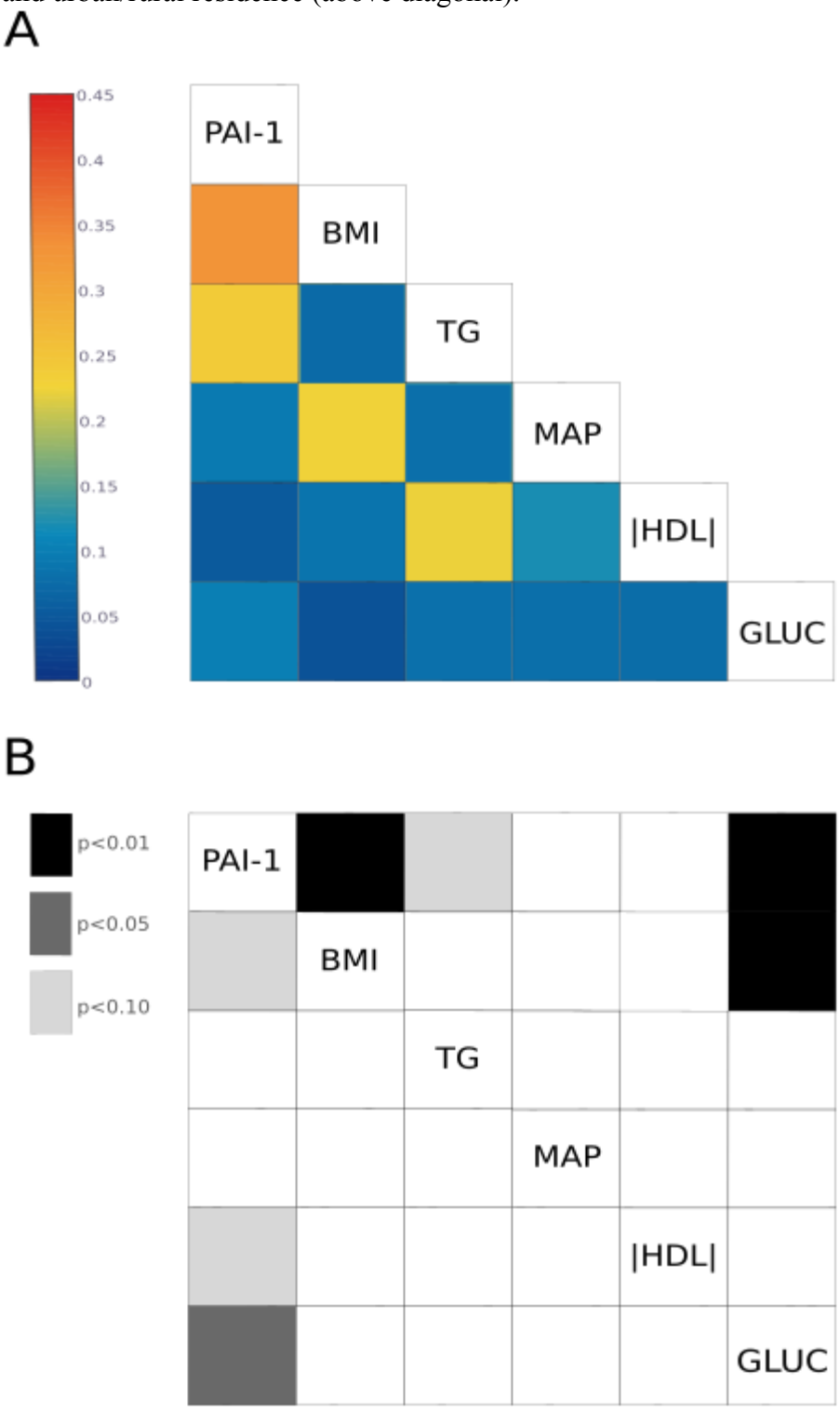
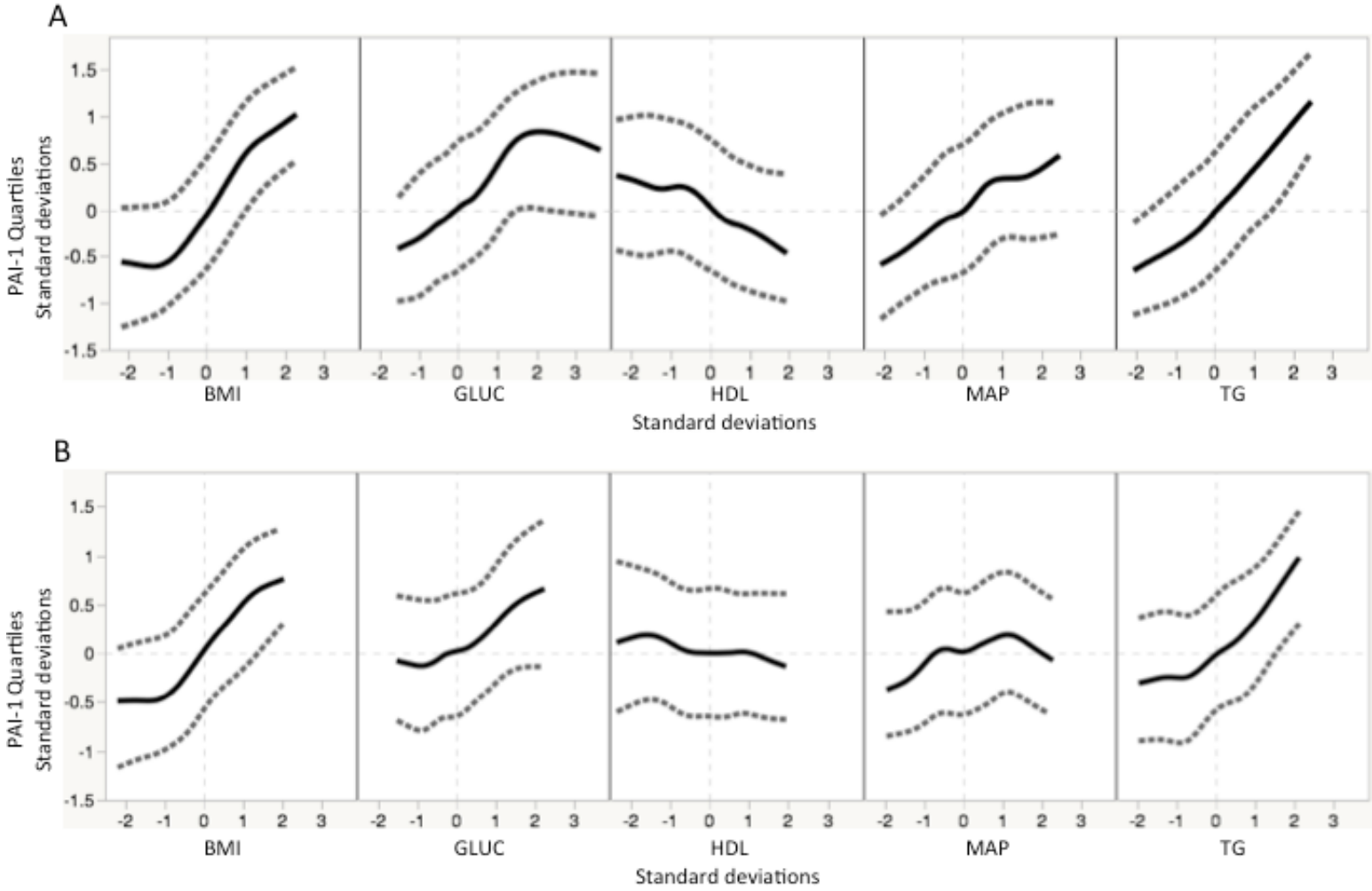


Figure 3-9. **Moving medians and 1st and 3rd quartiles of standardized PAI-1 values as a function of standardized MetS risk factor values**. (A) PAI-1 and MetS risk factors adjusted for age, sex, and residence; (B) PAI-1 and each risk factor also adjusted for the other 4 risk factors. Period for median = 100. Data smoothed using cubic spline (see Methods).



<u>Figure 3-10.</u> Moving medians and 1st and 3rd quartiles of standardized PAI-1 values as a function of standardized MetS risk factor values for men (blue) and women (red). (A) PAI-1 and MetS risk factors stratified by sex and adjusted for age and residence; (B) PAI-1 and each risk factor also adjusted for the other 4 risk factors. Period for median = 100. Data smoothed using cubic spline (see Methods).

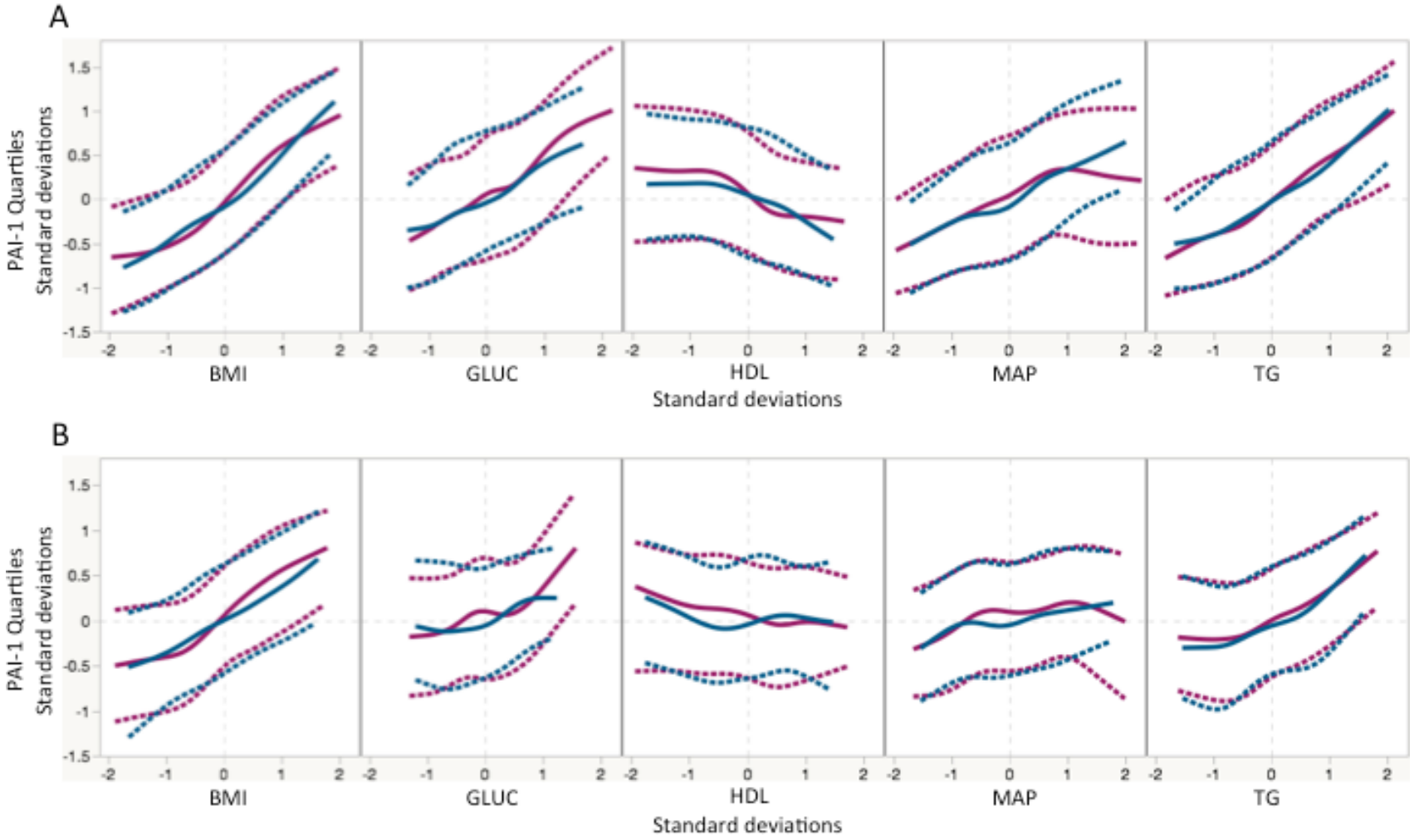
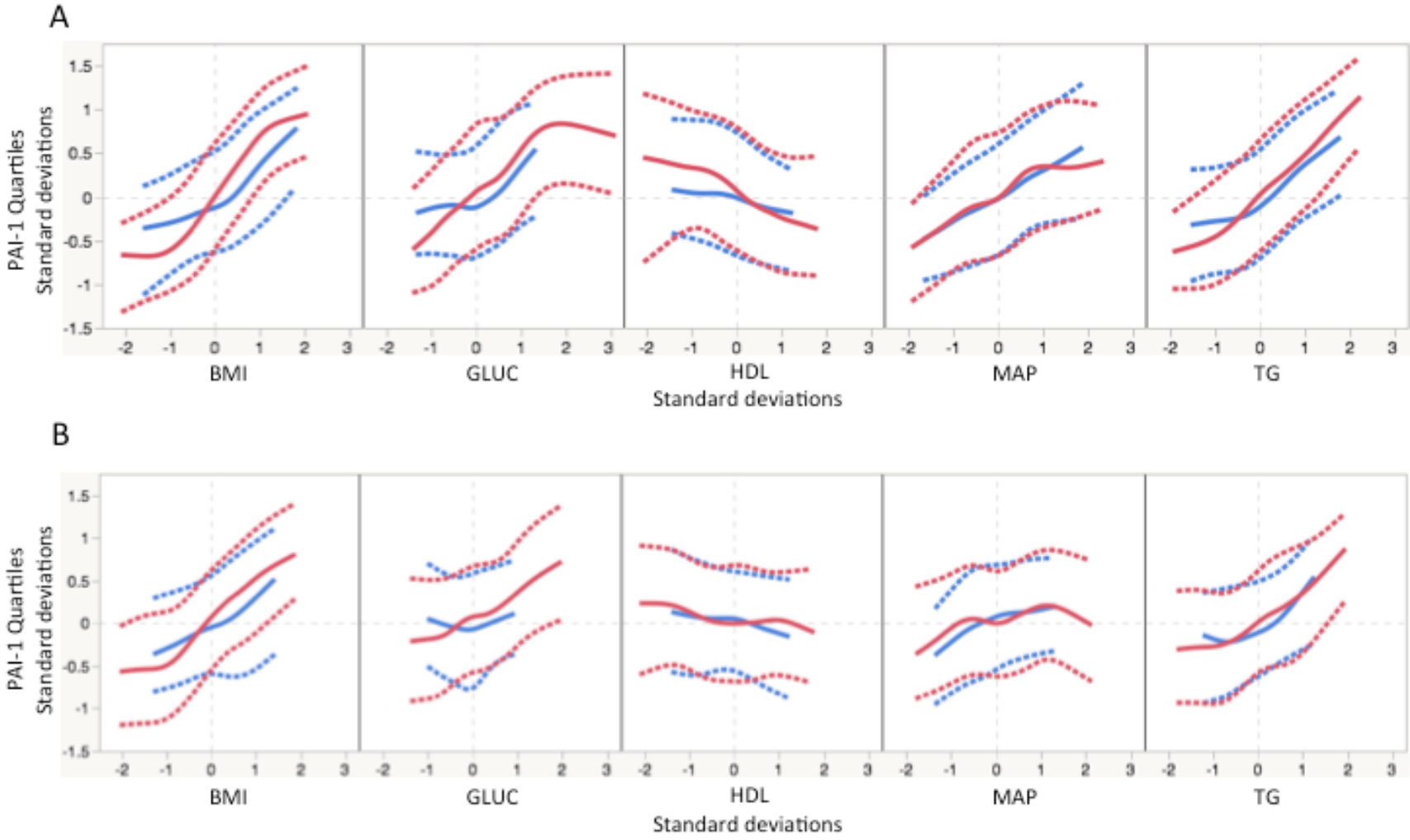


Figure 3-11. Moving medians and 1st and 3rd quartiles of standardized PAI-1 values as a function of standardized MetS risk factor values for urban (red) and rural (blue) participants. (A) PAI-1 and MetS risk factors stratified by residence and adjusted for age and sex; (B) PAI-1 and each risk factor also adjusted for the other 4 risk factors. Period for median = 100. Data smoothed using cubic spline (see Methods).



CHAPTER IV

MODELS OF CONTEXT-DEPENDENT GENETIC EFFECTS

Overview of statistical models

Throughout, we assume an additive, bi-allelic SNP in Hardy-Weinberg equilibrium that is distributed binomially with minor allele frequency p. For genotype $X \in [0,1,2]$:

$$\Pr(X = g) = \binom{2}{g} p^g (1-p)^{(2-g)}$$

$$Var(X) = 2p(1-p) = k$$

Because the variance of X, $2p(1-p)$, will appear regularly in the following models and derivations, it will be denoted by the constant k for simplicity.

With no loss of generality, we can center X by subtracting mean genotype, $2p$, so that

$$X \in [-2p, 1-2p, 2-2p]$$

Consider the linear regression model for a quantitative trait Y with expected value equal to zero, which includes an additive interaction between genotype X and a covariate, $Z \sim N(0,1)$, such that the *i*th individual has phenotype

$$Y_i = \beta_X X_i + \beta_Z Z_i + \beta_{XZ} X_i Z_i + \varepsilon_i$$

where $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ is an independent error term, and β_X, β_Z, and β_{XZ} represent the marginal effects of X, Z, and the interaction of X and Z, respectively.

Then,

$$Var(Y) = \beta_X^2 k + \beta_Z^2 + \sigma_\varepsilon^2$$

To aid interpretability and/or simplify calculation, we can occasionally set $Var(Y)$ equal to 1 and adjust σ_ε^2 accordingly. Then, for example, the narrow-sense heritability of Y with respect to X simplifies to

$$h_Y = 1 - \left(\beta_Z^2 + \sigma_\varepsilon^2\right) = \beta_X^2 k$$

Also, the correlation between trait Y and the covariate Z becomes

$$cor(Y,Z) = \frac{Cov(Y,Z)}{\sqrt{Var(Y)\cdot Var(Z)}} = Cov(Y,Z) = \beta_Z$$

because the covariance of Y and Z (regardless of the variance of Y or ε) is

$$\begin{aligned} &Cov(Y,Z) = E(YZ) - E(Y)\cdot E(Z) \\ &= E(\beta_X XZ + \beta_Z Z^2 + \beta_{XZ} XZ^2 + Z\varepsilon) - 0 \\ &= \beta_X E(X)\cdot E(Z) + \beta_Z \cdot E(Z^2) + \beta_{XZ} E(X)\cdot E(Z^2) + E(Z)\cdot E(\varepsilon) \\ &= \beta_Z \end{aligned}$$

Note that above, $E(X) = 0$ because we have centered X, and that $E(Z^2) = 1$, the variance of the standard normal distribution.

Much of the following discussion centers on ordinary least squares regression, primarily because CVD risk factors are quantitative traits. However, many of the ideas presented here can be extended to other general linear models, including logistic regression models of case/control data.

Overview

Evidence for context-dependent genetic effects

Cardiovascular disease risk factors have strongly inherited bases: estimates of heritability for conventionally measured traits such as lipid levels, fasting glucose, and blood pressure typically range from 40-60%[15,189,190]. Genetic epidemiological studies have for many years attempted to identify the genetic variants that explain a meaningful fraction of this heritable variation. Despite bringing several advantages to this search, genome-wide association studies (GWAS) have been unable to discover many SNPs that account for >1 per cent of the population variance of typical complex traits[9]. However, GWAS findings have allowed us to draw some firm conclusions where previously only speculation was possible. Three conclusions that concern us here are that (1) a large number of common variants, distributed unpredictably across the genome, have small (but not infinitesimal) effects on complex phenotypes; (2) these common variants often associate with multiple (and sometimes seemingly unrelated) phenotypes[17]; and (3) common variants *in aggregate* do indeed capture a large proportion of the heritable variance of complex phenotypes. Small individual effect sizes, however, make identifying relevant loci a major challenge.[191]

In theory, the power of GWAS to detect SNPs of biological interest is limited only by sample size, but inferences about genetic architecture from GWAS results are intrinsically limited. One reason for this is that genetic architecture, or a "genotype-phenotype map," makes sense only at the level of the individual. So, for example, if a SNP's contribution to a phenotype varies with environment or genetic background, its population-level effect-size may provide little insight into its physiological significance. Moreover, as sample size increases (as in meta-analyses), genotypic and phenotypic heterogeneity often increases with it, making the translation of GWAS results back to the level of the individual more challenging, even as the number of significant

hits increases. In the case of rare variants, we expect even highly deleterious SNPs to have negligible effect sizes at the population-level, so we exclude them from GWAS *a priori* on that account. In contrast, common variants susceptible to environmental or background-specific influences are cryptic. Thus, the degree to which small effect sizes are merely artifacts of heterogeneous genetic architecture remains an open question.

The poor replicability of association studies and the failure of even replicated SNPs to offer much predictive power are certainly consistent with the hypothesis that context-dependent genetic effects are pervasive[192]. However, the most conclusive evidence in favor of the phenomenon can be found in studies of model organisms, particularly those of wild-derived, inbred lines raised under controlled environmental conditions, such as the Drosophila Genetic Reference Panel (DGRP)[193]. The large number of flies per isogenic line in the DGRP enables highly precise mean and heritability estimates for multiple phenotype, while the characterization of millions of fixed SNPs per line has made large association studies possible. In one such study on starvation resistance, 115 significant SNPs were used to derive a multiple regression model that explained >80% of the phenotypic variance[193]. When the inbred *Drosophila* lines were allowed to interbreed, however, and a follow-up association study was conducted on starvation resistance, not one of the 267 significant SNPs in the new study overlapped with the 115 loci of the previous study, even at the nominally significant p-value of 10^-5 [194]. These findings clearly underscored the importance of genetic background effects on the role of SNPs in genetic architecture.

Similar conclusions about genetic architecture have been reached in many closely studied model organisms.[195] In yeast, GWA-type studies have had the additional benefit of enormous sample sizes and uniform allele frequencies of 50% (obtained by crossing haploid isogenic

strains)[196]. Here, too, when sample sizes are sufficiently large, the additive contribution of genetic factors to ecologically relevant traits can often be entirely attributed to detectable loci[197]. So, for example, one study reported that only 4 SNPs accounted for 100% of the heritability in yeast sporulation efficiency[198]. However, a follow-up experiment found that the variance explained by these 4 SNPs depended to a large extent on the precise combination of substrate (glucose, fructose, etc.) and strain background (oak vs. vineyard) tested[199]. Simply put, the phenotypic impact of SNPs (and SNP combinations) could not be predicted without taking both environment and genetic background into account.

Cardiovascular disease (CVD) in humans does not appear to be an exception in this regard. Epidemiological and quantitative genetic studies tell us that genetic factors taken together are a major component of CVD risk, while phenomena such as the parallel increase of CVD prevalence with urbanization indicate that genetic susceptibility must be understood within an environmental context. Sex, too, is a well-recognized modifier of genetic effects[200,201]. **Figure 4-1** depicts sex-based differences in heritability for a number of CVD-related risk factors. CVD-related endpoints also present differently in males and females with respect to onset, prognosis, and response to treatment.[22]

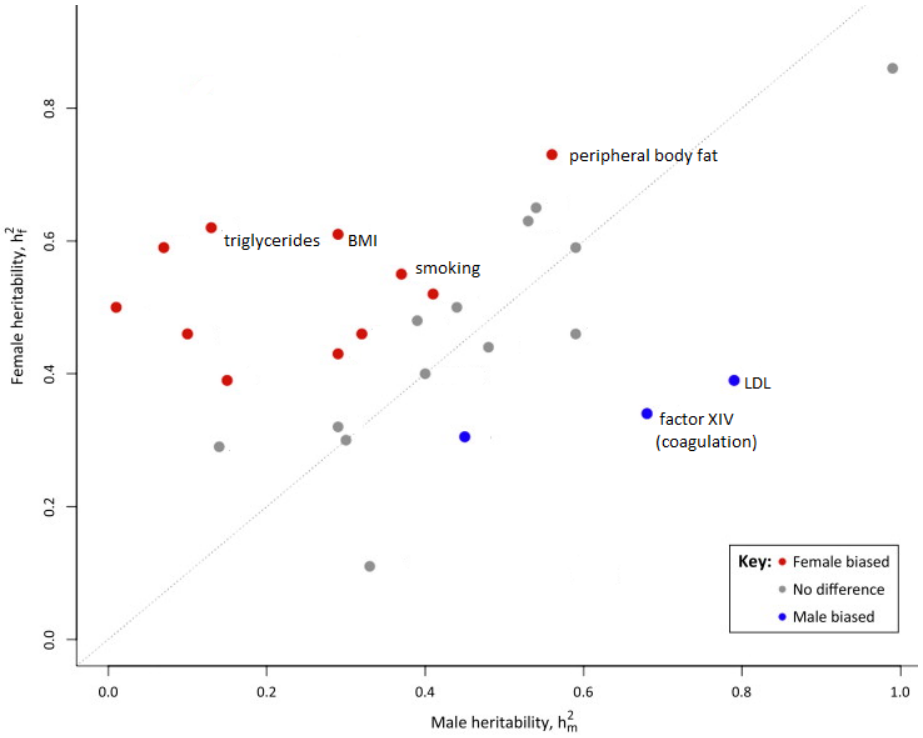


Figure 4-1: Comparison of narrow-sense heritability estimates for human traits by sex. Only those traits related to CVD are labelled. Statistically significant differences in heritability between sexes are red (higher in women) or blue (higher in men). Figure adapted from Gilks et al (2014).[202]

As a modifier of genetic effects, sex illustrates why "context" should be understood in the widest sense of the term. For, it is neither an environmental covariate, nor, strictly speaking, is it a genetic one, since loci involved in sex determination presumably have little direct connection to CVD risk downstream of development. Rather, sex represents a phenotypic background analogous to a genetic background—a physiological environment in which gene-"environment" interactions occur, leading to variation in trait expressivity and penetrance. It is unlikely that sex is unique among complex phenotypes in this regard, for the dynamic interdependence of physiological traits in general suggests that much genetic control occurs at the gene-phenotype interface. Nevertheless, genetic epidemiological studies have typically disregarded the possibility of gene-phenotype interactions. For example, the Catalog of Published GWAS hosted by the

National Human Genome Research Institute (NHGRI) lists ~100 entries on SNP-covariate interactions significant at the p< 10^-5 level, but only a handful of those covariates comprise physiological phenotypes. Two studies report interactions with BMI, three with age.

Typical genetic association studies do, of course, typically adjust for other phenotypes as covariates, but it has been shown that adjusting for phenotypes that are themselves heritable adds bias to associations.[203] Moreover, in the case of SNPs with covariate-contingent effects, mere adjustment without an allowance for possible interactions can weaken signals substantially. As a case in point, there is evidence that the dearth of significant loci related to insulin resistance in GWAS of type 2 diabetes (T2D) can be attributed in part to the standard procedure of adjusting for BMI[204]: by narrowing the search to loci that affect T2D independent of BMI, such studies have unwittingly diminished their power to detect the adiposity-mediated SNPs involved in insulin resistance.

Another clue that gene-phenotype interactions (as distinct from gene-gene (GXG) or gene-environment (GXE) interactions) are common and consequential can be found in the growing number of SNPs reported to display quantile-specific penetrance, including those associated with BMI and lipid traits.[205-207] That the effect sizes of many well-characterized SNPs are larger at the tails of phenotypic distributions suggests that the phenotypes themselves may increase SNP expression, or serve as surrogates (or modifiers) of others that do. Thus, an implicit assumption of GWAS—that the genotype-phenotype relationship holds across phenotypic distributions —is suspect.

In contrast, statistical interactions have been sought frequently between pairs of SNPs. However, these efforts have met with little success[208] and the search appears to rest on weak theoretical ground.[209] Research on GXE interactions in relation to complex disease is active and

growing, but information about environmental exposures can be inconsistent (with respect to timing, duration, measurement, etc.) and difficult to collect prospectively.[210,211] Phenotypic measurements, on the other hand, are commonly collected within single studies, following standardized protocols. Moreover, phenotypes such as CVD-related traits are often correlated, indicating that they share common influences that may also change gene expression. Thus, like sex and age, a large number of anthropometric and physiologic measurements can be considered proxies for a complex network of factors that direct gene expression and influence phenotypic plasticity.

Limitations of existing methods

The routine statistical approach to detecting context-dependent genetic effects is to fit a regression model with a cross-product term (genotype*covariate) to phenotypic data. A significant nonzero coefficient for the interaction term is interpreted as evidence of departure from the linear model. In the context of genetic studies, however, the interaction beta coefficient can be interpreted more specifically, which we present here as motivation for what follows. Namely, if a SNP (*X*) has no effect on a covariate (*Z*), then the interaction coefficient is a measure of the expected change in covariance between *Z* and the outcome variable with each additional allele copy:

Given the linear regression model, with parameters as defined in the Statistical Overview

$$y_i = \beta_0 + \beta_X x_i + \beta_Z z_i + \beta_{XZ} x_i z_i + \varepsilon_i$$

in which coefficients represent actual effects and not estimates, the covariance between outcome *Y* and covariate *Z*, conditioned on genotype *g* is

$$\begin{aligned} &Cov(Y,Z \mid X=g) = E(YZ \mid X=g) - E(Y \mid X=g)\cdot E(Z \mid X=g) \\ &= E(\beta_X gZ + \beta_Z Z^2 + \beta_{XZ} gZ^2 + Z\varepsilon) - 0 \\ &= \beta_X g \cdot E(Z) + \beta_Z \cdot E(Z^2) + \beta_{XZ} g \cdot E(Z^2) + E(Z)\cdot E(\varepsilon) \\ &= \beta_Z + \beta_{XZ} g \end{aligned}$$

With each additional allele, $\Delta\, Cov(Y,Z)$ is:

$$\begin{aligned} &Cov(Y,Z \mid X=g+1) - Cov(Y,Z \mid X=g) \\ &= \left(\beta_Z + \beta_{XZ}(g+1)\right) - \left(\beta_Z + \beta_{XZ} g\right) \\ &= \beta_{XZ} \end{aligned}$$

We see that a SNP that has a context-dependent effect on an outcome (with *Z* as "context" and *Y* as outcome) is here conceptually equivalent to a SNP that additively changes the covariance

between the variables *Y* and *Z*.

However, an additive change in covariance represents only a special case of a more general phenomenon. There are, in theory, a number of ways SNPs can influence covariance, yet not be detected (or only weakly detected) by the interaction coefficient of the standard linear regression model. Importantly, in the above derivation, the coefficients represent actual, not estimated effects; in practice, the ability of the SNP*covariate interaction coefficient to estimate changes in covariance between *Y* and *Z* will be sensitive to the "correct" characterization of the dependent versus the independent variable. When dealing with biological phenotypes (as opposed to, e.g., environmental exposures) this characterization is not always straightforward, and may in fact be unwarranted. For example, a genetic effect on the covariance between Y and Z can occur even when neither phenotype is biologically dependent on the other. It should also be evident that β_{XZ} becomes a less effective estimator of changes in covariance as the dominance deviation from additivity increases. If the heterozygote genotype changes the covariance between *Y* and *Z*, for instance, while the homozygote genotypes do not (i.e. incomplete dominance), the expected coefficient of interaction would be zero. In Chapter 5, we introduce a way to detect genetic effects on covariance without assuming additivity.

There is also the case of pleiotropic SNPs, which influence the covariance among multiple traits in a way that the interaction coefficient β_{XZ} and, in fact, the standard linear regression approach altogether, have no power to detect. Recently, several multivariate methods have been developed to address such cross-phenotype associations, with the basic objective of identifying genetic variants that associate with multiple traits simultaneously. These will be discussed in Chapter 5, where we introduce a statistical approach that expands upon them.

Having (1) emphasized the likely importance of context-dependent genetic variants from

the standpoint of both pathophysiology and epidemiology; (2) presented a conceptual link between context-dependence and covariance modification; and (3) shown that the conventional interaction models used in association studies are limited in their ability to assess genetic effects on trait covariance, we can now ask whether improving the sensitivity of statistical models to such effects may increase our power to identify biologically meaningful genetic variants. The presentation of such methods will be the focus of the next chapter. Here, we have set for ourselves the preliminary goal of developing biologically plausible models of genetic context-dependence. These models will serve as both a theoretical framework for assessing the performance of novel and conventional statistical methods, as well as a source of data generation to make such assessments empirically.

Introduction

Genetic association studies that test for interaction typically have one of two aims. In some studies, the motivation is to enhance the genetic signal of a single-locus test by allowing for the possibility of an interaction. Such analyses require a joint test of both main and interaction effects. Other studies focus on the interaction term itself, with the objective of detecting SNP-covariate interactions that are statistically significant even in the absence of marginal genetic effects.

It is worth illustrating just how exceptional significant interactions in the absence of marginal effects likely are in the context of genetic epidemiology. To do so, we can simulate data using the standard regression model with an interaction term

$$Y_i = \beta_X X_i + \beta_Z Z_i + \beta_{XZ} X_i Z_i + \varepsilon_i$$

setting the proportion of variance explained by the interaction equal to 1.5% (an unusually strong effect, for illustrative clarity), and the marginal effect of the SNP to zero. In **Figure 4-2**, we see graphically that there is no "good" or "bad" genotype, but context (i.e. covariate Z) makes it so.

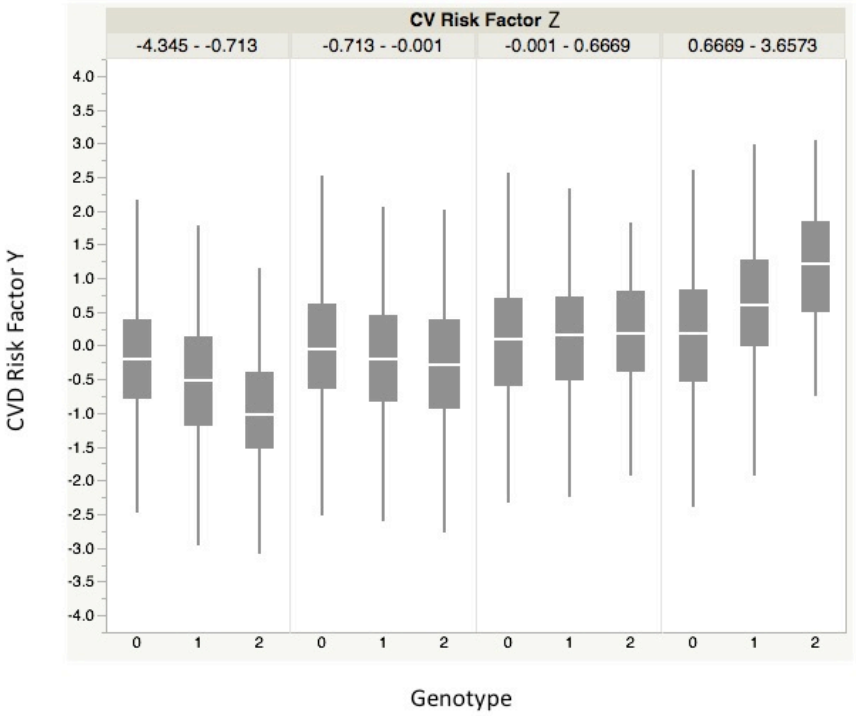


Figure 4-2 Simulated data illustrating the effect of a strong interaction effect with no marginal effect, arranged by genotype and by covariate-quartile. The covariate *Z* is standardized, normally distributed, and correlated to *Y* at r=0.10; N=5000.

In the next chapter we discuss a class of theoretical SNPs that would in fact leave such a fingerprint, namely genes directly involved in regulating the covariance between two biological traits. For an example of a more commonly observed type of interaction, however, we can turn to a recent study, which found that a lack of physical activity "accentuates" the effect of a variant in the gene *FTO* on BMI.[212] The variant had already been shown by multiple studies to associate with increased BMI and obesity.[213,214] If it had displayed only interaction effects and no marginal effects, individuals with 2 copies of the allele who exercised frequently would have had lower BMI than individuals with 1 or 0 copies who exercised the same amount. In other words, the *FTO* allele would reverse roles, switching from a risk-conferring factor to a risk-reducing one, depending on level of activity. We might expect such cases to be relatively rare, and in fact that expectation has led to recommendations that joint tests should be used when testing for interactions.[215] [208]

In contrast, **Figure 4-3** depicts a strong SNP-covariate interaction that also displays a marginal effect. Here, the relative status of the SNP as a "risk" allele does not change with the value of the covariate *Z*, as before. In these more "normal" scenarios, a marginal effect will accompany an interaction effect.

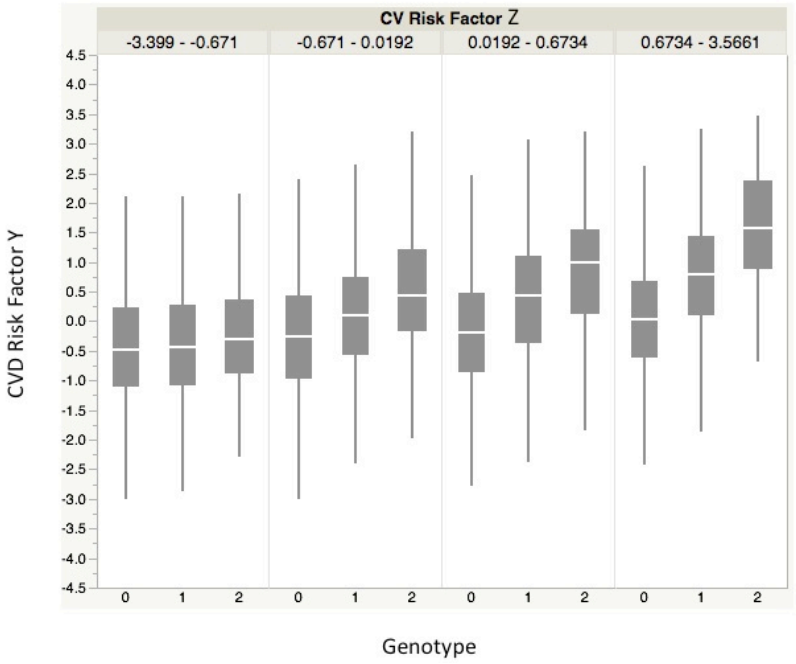


Figure 4-3 Simulated data illustrating the effect of a strong interaction effect with marginal effect, arranged by genotype and by covariate-quartile. Note that the relative status of the SNP as a "risk" allele does not change with Z, as in **Figure 2**. When that does not happen, a marginal effect will accompany an interaction effect. The covariate *Z* is standardized, normally distributed, and correlated to *Y* at 0.10; N=5000.

However, the statistical marginal effects that frequently accompany significant gene-covariate interactions do not necessarily tell us anything about genetic architecture. In other words, however strong the marginal effect, one cannot infer that at the biological level, the gene has an effect independent of the covariate. It is entirely possible that reported main effects are merely the statistical artifacts of pure SNP-covariate interactions at the biological level. In fact, the data depicted in **Figure 4-3** was simulated from a model of genetic effects completely dependent on a covariate's value, which we will describe below.

At present, it is commonly accepted that most statistical interaction effects will display marginal effects, but mostly in vague terms.[208] It is not possible to estimate, for example, the degree to which the relative "share" of a biological gene-by-covariate interaction is captured by marginal or interaction effects in a regression analysis. It is thus impossible to quantify the advantage of using joint interaction tests over single tests, or to elucidate the conditions under which each may be most applicable. Moreover, to our knowledge, there is at present no way to simulate phenotypic data based on a realistic and *specifiable* basis of covariate-dependent genetic architecture. Thus, it is difficult to weigh the pros and cons of various methods for detecting such SNPs. Here, we present three general models based on more realistic mechanisms of genetic architecture.

Theoretical Model 1

[Note: Although the terms "main effect" and "interaction effect" are statistical concepts and thus do not necessarily translate back to biology, we have borrowed them here to define underlying mechanisms. In brief, a genetic variant's "main effect" as used here represents its baseline contribution to a quantitative trait at the individual level. A gene with only a main effect therefore has the same effect regardless of physiological context. An "interaction effect" represents an addition to the genetic variant's "main effect" owing to some covariate.]

Consider the scenario of a standardized covariate Z that amplifies (multiplicatively) a SNPs effect, but only after it (the covariate) passes a certain threshold. For mathematical tractability, we can set this threshold to the mean, zero. The covariate has no effect on the SNP otherwise. The SNP X may also have a "main effect" β_X on outcome Y, where "main effect" is defined as an effect not influenced by Z

$$Y_i = \begin{cases} \beta_X X_i + \beta_Z Z_i + \varepsilon_i, & Z_i \leq 0 \\ \beta_X X_i + \beta_Z Z_i + \beta_{XZ} X_i Z_i + \varepsilon_i, & Z_i > 0 \end{cases}$$

The expected proportion of variance explained by the marginal effect of the SNP is

$$R_X^2 = \frac{\left[Cov(X,Y)\right]^2}{Var(X)\cdot Var(Y)} = k\left(\beta_X + \frac{\beta_{XZ}}{\sqrt{2\pi}}\right)^2$$

where $k=Var(X)=2p(1-p)$

(see Appendix B-1 for derivation)

We see here that even if the SNP has no main effect ($\beta_X = 0$), meaning that the effect of SNP *X* depends completely on a covariate's value, a regression analysis would nonetheless detect a marginal effect, which would account for an expected $\frac{\beta_{XZ}^2 k}{2\pi}$ of the variance of *Y*.

We can similarly derive the expected proportion of variance explained by the interaction cross product (see Appendix B2):

$$R_{XZ}^2 = \frac{\left[Cov(XZ,Y)\right]^2}{Var(XZ)\cdot Var(Y)} = \frac{\beta_{XZ}^2 k}{4}$$

Note that the ratio $R_X^2 : R_{XZ}^2$ tells us how the "significance" of the biological interaction would likely be distributed among the marginal and interaction effects in a standard regression model.

$$\frac{R_X^2}{R_{XZ}^2} = \frac{2}{\pi}$$

We see that when a SNP's effect on a quantitative trait completely depends on the value of a covariate, such that the SNP has no effect if the covariate is below its mean, and a multiplicative effect as the covariate increases above its mean, the expected proportion of variance explained by the marginal effect of the SNP is roughly 2/3 of the proportion of variance explained by the SNP-covariate interaction effect in a standard regression analysis.

Using the formulae for R_X^2 and R_{XZ}^2, we can now specify one or the other to simulate data

structured by an underlying genetic architecture. The fully parameterizable R code for such simulations can be found in Appendix D **Figure 4-4** was generated by setting β_X =0; the correlation between *Y* and *Z* to 0.10; and R^2_{XZ} to 1.5%, to match **Figure 4-2**.

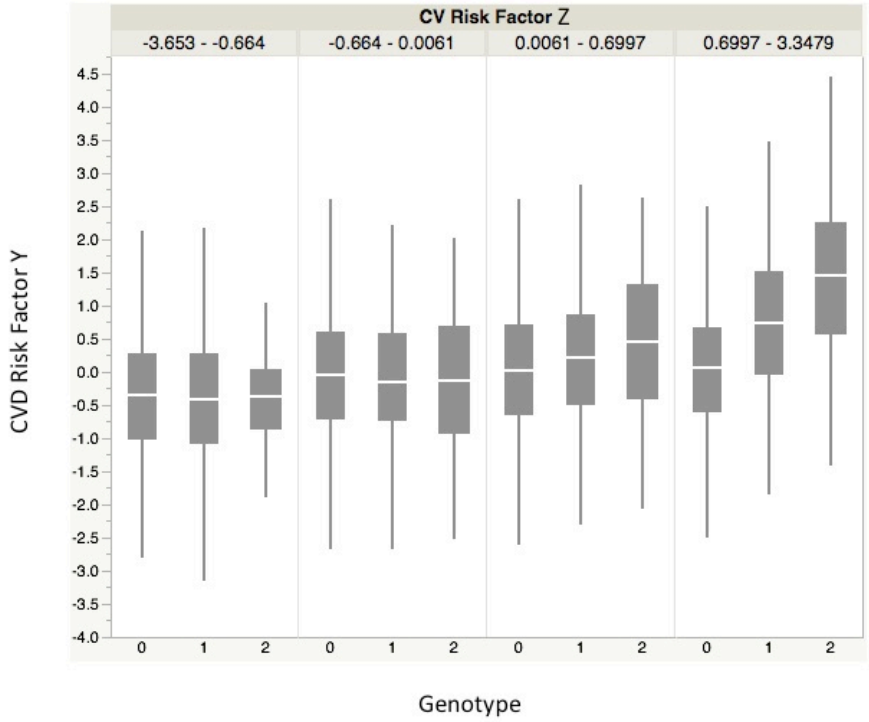


Figure 4-4 Data simulated using Model 1, illustrating the effect of a strong interaction effect with marginal effect, arranged by genotype and by covariate-quartile. The covariate *Z* is standardized, normally distributed, and correlated to *Y* at 0.10; N=5000. Note that because R^2_{XZ} was set to 1.5%, the proportion of variance explained by the marginal effect $\approx (3\%)/\pi$, as per the formula above.

Theoretical Model 2

In Theoretical Model 1, we arbitrarily set a threshold, $Z=0$, below which Z did not interact with SNP X. Moreover, Z had a multiplicative impact on the effect of X when Z>0. Here, we qualitatively change both those conditions: thresholds can be modified and the impact of Z on X is additive.

Consider the scenario where the effect of SNP X on quantitative trait Y depends entirely on the value of Z, but in a stepwise, not continuous manner. So, for example, a SNP may have no effect until $Z= -1$ (i.e. quantitative trait Z is one standard deviation below its mean); then $Z>-1$ might, for example, trigger a histone modification, such that SNP X has a small, constant effect until $Z>1$, when an enhancer is activated and the effect of X increases to its maximum.

For mathematical tractability we set 2 simplifying constraints: (1) all stepwise changes in the genetic effect of X occur at quantiles of the distribution of Z (there is no constraint on the number of quantiles), and (2) the genetic effect of X increases by a constant increment, c. See **Figure 4-5**.

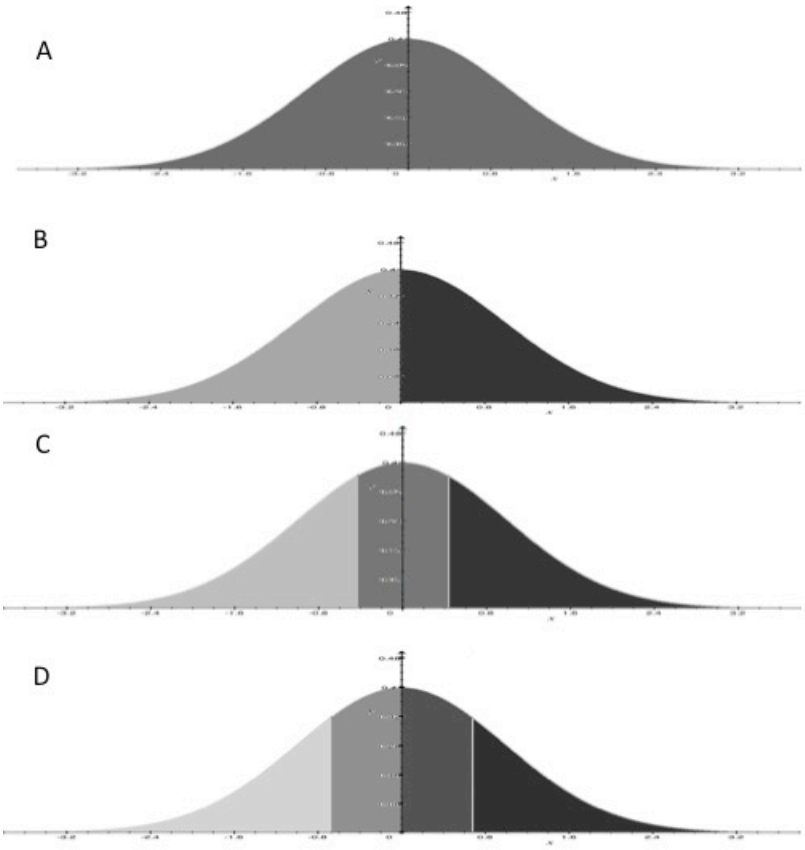


Figure 4-5 Schematic of the genetic architecture informing Model 2
Distributions are of covariate *Z~N(0,1)*. The effect of *X* on *Y* is represented by the darkness of the shaded regions: In panel (A), *Z* has no impact on the effect of *X;* in (B) the effect of *X* increases a fixed amount at *Z*=0, but afterwards does not vary with *Z* (an example might be a certain threshold of cigarettes that causes an epigenetic modification); (C) the effect of *X* increases by a fixed amount, *c,* after the first tertile, and by the same amount after the second tertile; (D) the effect of *X* increases by *c* at each quartile.

The outcome *Y* is modeled as a step function, divided into *q* quantiles of the covariate *Z*.

So, for the first quantile of Z,

$$Y = \beta_1 X_i + \beta_Z Z_i + \varepsilon_i$$

where $\beta_1 = c$ (the minimal effect of SNP *X*).

For the second quantile of *Z*,

$$Y = \beta_2 X_i + \beta_Z Z_i + \varepsilon_i$$

with $\beta_2 = 2c$

In general:

$$Y = \begin{cases} \beta_1 X_i + \beta_Z Z_i + \varepsilon_i, & 0 < \Phi\left(Z_i\right) < \frac{1}{q} \\ \beta_2 X_i + \beta_Z Z_i + \varepsilon_i, & \frac{1}{q} < \Phi\left(Z_i\right) < \frac{2}{q} \\ \vdots \quad \vdots \quad \vdots & \vdots \\ \beta_q X_i + \beta_Z Z_i + \varepsilon_i, & \frac{q-1}{q} < \Phi\left(Z_i\right) < 1 \end{cases}$$

where

$\beta_j = jc,$
$j \in \left[1,2,\ldots,q\right]$

As with Theoretical Model #1, the genetic effects here are also completely depdnet on the covariate *Z.* The fully parameterizable R code for generating data with this model can be found in Appendix D.

By calculating the variance of *Y*, we can define the value of *c* in terms of the heritability of *X:*

$$Var(Y) = E(Y^2) = \left(\frac{1}{q}\left(k\beta_1^2\right) + \frac{1}{q}\left(k\beta_2^2\right) + \ldots + \frac{1}{q}\left(k\beta_q^2\right)\right) + \beta_Z^2 + \sigma_\varepsilon^2 \qquad \text{Note}^1$$

$$= \frac{k}{q}\left(c^2 + \left(2c\right)^2 + \left(3c\right)^2 \ldots + \left(qc\right)^2\right) + \beta_Z^2 + \sigma_\varepsilon^2$$

$$= \frac{c^2 k}{q}\left(1 + 4 + 9 + \ldots + q^2\right) + \beta_Z^2 + \sigma_\varepsilon^2$$

$$= \left(\frac{c^2 k}{q}\right)\left(\frac{q(q+1)(2q+1)}{6}\right) + \beta_Z^2 + \sigma_\varepsilon^2$$

$$= \frac{(q+1)(2q+1)c^2 k}{6} + \beta_Z^2 + \sigma_\varepsilon^2$$

(Note that $\left(\frac{n(n+1)(2n+1)}{6}\right)$ is the formula for the sum of n consecutive squares.)

To simplify, we can adjust σ_ε^2 so that *Var(Y)* = 1 (see Statistical Overview).

Then, the heritability of X is

$$h_X = \frac{(q+1)(2q+1)c^2 k}{6}$$

and

$$c = \sqrt{\frac{6h_X}{k(q+1)(2q+1)}}$$

---

[1] It is permissible to divide the variance terms of *Y* into *q* parts because the quantiles are contiguous, and the integrals for variance "telescope"; e.g. in general:

$$\int_a^b f(x)dx + \int_b^c f(x)dx + \int_c^d f(x)dx$$
$$= F(x)\Big]_a^b + F(x)\Big]_b^c \ldots F(x)\Big]_c^d$$
$$= F(b) - F(a) + F(c) - F(b) + F(d) - F(c)$$
$$= \int_a^d f(x)dx$$

Thus, data can be simulated simply by specifying h_X, as well as the parameters q (number of quantiles) and β_Z (desired correlation between Z and Y).

As we did for Theoretical Model 1, we can derive the expected proportion of variance explained by marginal effects vs. interaction effects when a regression analysis is run on data generated by this model (see Appendix B-3 for derivations). The relative values will depend on the number of quantiles chosen for the model. **Figure 4-6** depicts the ratio $R_X^2 : R_{XZ}^2$ as the number of quantiles increases from 2 to 20. In contrast to Theoretical Model 1, where the expected proportion of variance explained by the marginal genetic effects was $2/\pi$ times the size of expected proportion of variance explained by the interaction, in Theoretical Model 2, R_x^2 is substantially larger than R_{xc}^2, converging quickly to π.

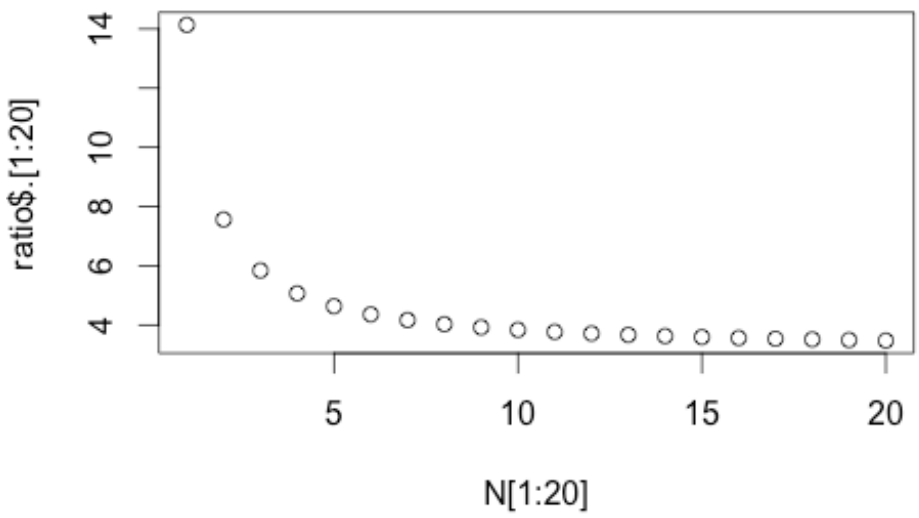


Figure 4-6 Expected ratio of R_X^2 : R_{XZ}^2 for Model 2, plotted against the number of quantiles. As the number of quantiles increases from 2 to ∞, the ratio between the expected proportion of variance explained by marginal genetic effects vs. the expected proportion of variance explained by the interaction falls from $\frac{9\pi}{2}$ to π.

Theoretical Model 3

As we increase the number of quantiles in Theoretical Model 2, the step function becomes more continuous.

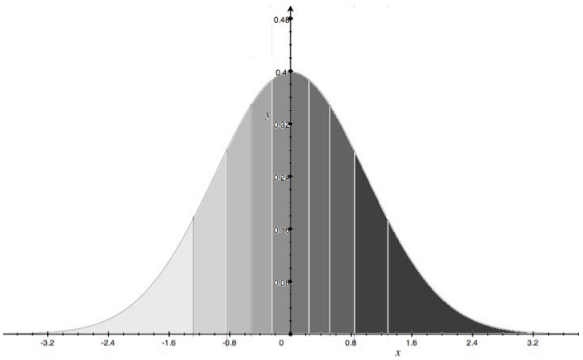


Figure 4-7 Increasing the number of quantiles. As the number of quantiles increases to ∞, the step function of Theoretical Model 2 transforms into the continuous function of Theoretical Model 3 without "main effects".

It may be helpful to imagine a biological scenario where this may be the case, such as *X* as a receptor variant sensitive to the concentration of a hormone *Z*. Increasing levels of Z may trigger more system-wide signal transductions by *X* in a manner that is continuous, not stepwise as in Theoretical Model 2; but the effect on *Y* may be proportional to the *percentile* of *Z* rather than its actual value, owing to rate-limiting factors. In other words, the interaction would not be captured by *XZ*, but rather by $X\Phi(Z)$:

$$Y = \beta_{X\Phi}\left(X_i \cdot \Phi(Z_i)\right) + \beta_Z Z_i + \varepsilon_i$$

to which we can add a "main effect" term

$$Y = \beta_X X_i + \beta_{X\Phi}\left(X_i \cdot \Phi(Z_i)\right) + \beta_Z Z_i + \varepsilon_i$$

creating a hybrid of Theoretical Models 1 and 2. A benefit of defining the cross-product interaction as $X\Phi(Z)$ versus *XZ* is that the former maintains the direction of effect of a SNP, as

in the other models. In other words, we can represent a deleterious (or beneficial) allele that is context-dependent, yet still more deleterious (or beneficial) than the other allele regardless of context.

We also see that

$$R_X^2 = k\left(\beta_X + \frac{\beta_{X\Phi}}{2}\right)^2$$

and

$$R_{ZX}^2 = \frac{\beta_{X\Phi}^2 k}{4\pi}$$

(See Appendix B-4 for derivations)

So that, when β_X =0, the ratio

$R_X^2 : R_{XZ}^2$

is equal to π, just as we would expect, since when $\beta_X = 0$, this model is the same as Theoretical Model 2 with an infinite number of quantiles (see **Figure 4-6, 4-7**).

Note that Theoretical Model 3 generated data for **Figure 4-3** above. A fully parameterizable R-code for generating data using this model can be found in Appendix D.

Discussion

We have modeled multiple types of SNP-by-covariate interactions that may plausibly occur at the biological level, and the statistical (i.e. populational) parameters to which they would give rise. We have already discussed how these models allow us to estimate the relative importance of the interaction coefficient versus the coefficient of marginal effect in standard regression analyses under various conditions. An additional strength is that data generated by these models will be more realistic, and structured differently than data generated using standard regression models, even when the same R^2 for marginal and/or interaction effects is specified. The joint distribution of the two phenotypes will be approximately bivariate normal, but not perfectly so; for instance, the effects of the interactions that shape the outcomes may be driven more by the tails of the two variables. Recently, the development of novel methods to detect interactions, especially GXE interactions, have become an active field of inquiry;[216] testing them on data generated by these models should thus better predict their performance with real data.

We have underscored that a statistical method will be more powered to identify context-dependent genetic effects in proportion to its sensitivity to genetically induced changes in covariance between "context" and "outcome." Here, two points must be addressed. First, changes in covariance can occur in multiple ways, such that statistical approaches should address differences in kind as well as degree (**Figure 4-8**). Second, the conceptualization of phenotypes as either "context" or "outcome" may not be apt. Physiological processes are complex, and the "feedback" mechanisms of trait interdependence are one likely reason. Thus, an ideal approach allows for ambiguity of directionality between putative "outcome" and "trait" variables.

As a final note: although we have framed the utility of detecting genetic modifications covariance as a way to discover biologically meaningful SNPs, such approaches can also be used

to discover the biologically significant contexts for SNPs already deemed significant. Moreover, although the focus of the next chapter will be CVD-related phenotypes, the statistical concepts presented in this chapter can in theory be extended to covariates of all kinds, including to a number of environmental factors, as well as quantitative measurements of genetic background, such as those obtained by principal components analysis (PCA). It is worth noting here that PCs, like the phenotypic covariates discussed above, are typically "adjusted for" in epidemiological studies, while potential gene-background interactions are ignored, despite strong evidence of such effects in model organisms (e.g. the strain-background differences discussed above), and some evidence that such effects exist in humans.[217]

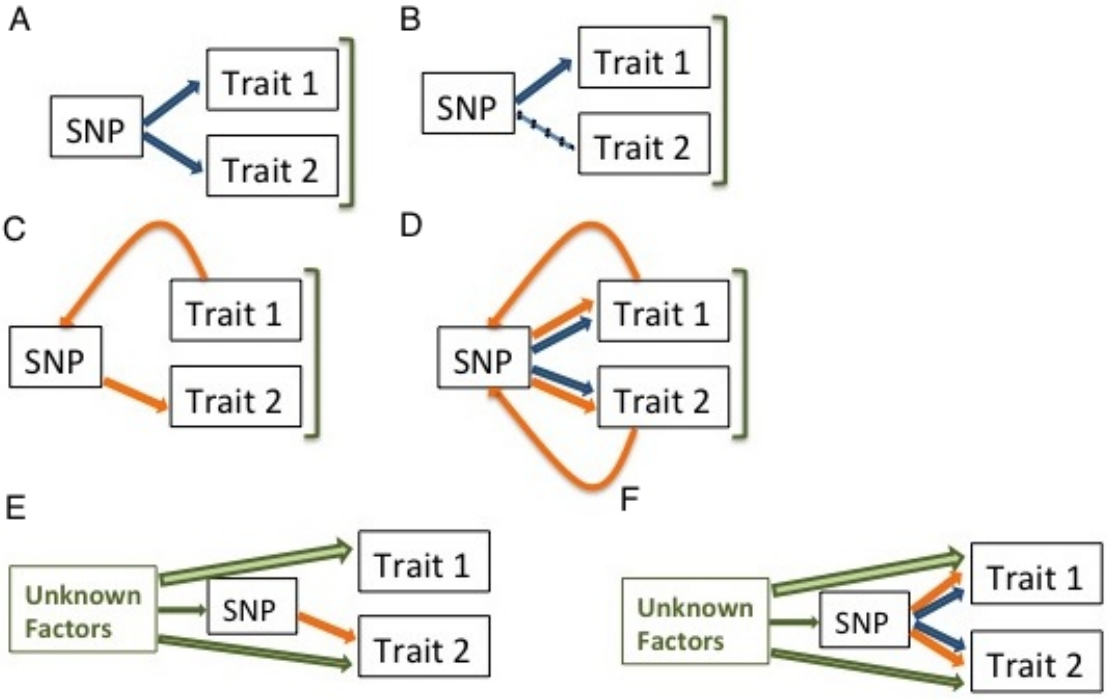


Figure 4-8: **Schematic of ways SNPs can influence covariance between traits.** The green brackets and arrows represent correlation between factors. Blue arrows denote independent effects. Orange arrows illustrate context-dependent genetic effects; the curved orange arrow represents the effect of a trait on the expression of a SNP in the direction of the straight orange arrow(s). Thus, (A) and (B) represent pleiotropy, with pleiotropic effects occurring in the (A) same and (B) opposite directions of the correlation. In (C), Trait 1 is influencing the expression of the SNP, which in turn affects Trait 2; while (D) shows that the situation can become far more complicated. Panel (E) depicts unknown factors that drive a correlation between two traits by influencing them to different degrees (denoted by the size of the arrow), including by changing the expression of a SNP. Finally, in (F) context-dependent pleiotropy is added to the picture.

CHAPTER V

GENETIC ANALYSES OF CVD RISK FACTORS INTERACTIONS IN A GHANAIAN POPULATION

A Multivariate Method to Identify Pleiotropic and Context-Dependent Genes

Introduction

Many of the genetic variants identified by genome-wide association studies (GWAS) are associated with multiple traits.[218] A conservative estimate of the number of cross-phenotypic associations in 2011 implicated 17% of genes, and the percentage for certain classes of phenotypes, such as autoimmune diseases, was even greater (44%).[219 220] Estimates of genetic correlations derived using quantitative genetic techniques have confirmed these observations. For example, a recent application of genome-wide complex trait analysis (GCTA)[221] found that the same genes accounted for ~39% of the heritability of several traits involved in the metabolic syndrome. [18] A similar analysis of case/control data concluded that the genetic correlation between type 2 diabetes and hypertension was 0.31.[222]

These observations have inspired a wave of new methods designed to analyze multiple traits simultaneously.[223 224 225,226 227] While these multivariate genome-wide association (mvGWA) methods share a common objective—improving SNP discovery by taking advantage of the additional information provided by multiple phenotypes—they have approached it from different angles, using diverse statistical tools, such as linear mixed models[224], canonical correlational analysis[223], and principal components regression.[225] Nevertheless, they generally yield remarkably similar results, with minor comparative advantages among them usually relating to particular aspects of study design.[225,227]

However, multivariate methods are not always more powered to detect pleiotropic SNPs than single-trait analyses run multiple times and adjusted for multiple testing. For example, they

perform worse than single-trait analyses when SNPs have equal effects on a number of equally correlated traits.[225,227] Moreover, although mvGWA methods are typically applied to correlated phenotypes, and their ability to "leverage" correlation is often emphasized,[225,227,228] the strength of correlation between phenotypes *per se* has no necessary bearing on their statistical power. Rather, multivariate tests have a comparative advantage over single-trait tests only when there is a contrast in size and direction between genetic effects on the one hand and residual trait correlations on the other.[226,227,229] Residual correlations are the correlations between traits minus the genetic effects.

Given that the comparative advantage of multivariate tests over univariate tests depends entirely on the strength of a gene's effect on phenotypic correlation, improving the sensitivity of multivariate methods to such effects should improve their power. Indeed, there are a number of ways a gene can influence correlation that are missed completely by existing multivariate methods. Genetic loci might influence covariance directly, for example, by controlling the synchronization of two traits in a homeostatic pathway. Variants at such loci would have no expected "marginal" effects, and thus be completely missed by existing multivariate and univariate models. Such "covariance genes" are discussed in more detail in the next section. Context-dependent SNPs, such as those that increase the expression of one trait only when another trait is past a certain threshold, also influence trait covariance (Chapter 4). Although existing mvGWA methods can detect such SNPs, they are not optimized to do so. These SNPs will be the focus of this section.

The ordinal joint interaction method (OJIM)

Perhaps the most straightforward of the recent multivariate methods, with regard to ease of applicability and interpretation of results, is MultiPhen[226], which inverts the standard regression model used in GWAS, such that the genotype X_{ig} (for individual *i* and SNP *g*) becomes the outcome variable and *K* phenotypes and/or covariates are modeled as predictor variables. Because genotype is an ordinal set, MultiPhen utilizes ordinal, not linear regression.

$$P(X_{ig} \leq m) = \frac{1}{1 + e^{(-\alpha_{gm} - \sum_{k=1}^{K} \beta_{gk} Y_{ik})}}$$

Rather than estimating the linear change in an outcome variable caused by a predictor, the beta coefficients here correspond to the change in the probability that an individual (X*i*) "moves up" to the next (genotype) class as the predictor variable increases by a unit. In the context of GWAS, *m* can equal 0, 1, or 2, with the simplifying constraint being that the odds of moving from 0 to 1 and from 1 to 2 are equal. The α-term is analogous to the intercept in linear regression. The null hypothesis tested to determine the significance of the overall model is that the beta coefficients of all predictor variables equal zero.

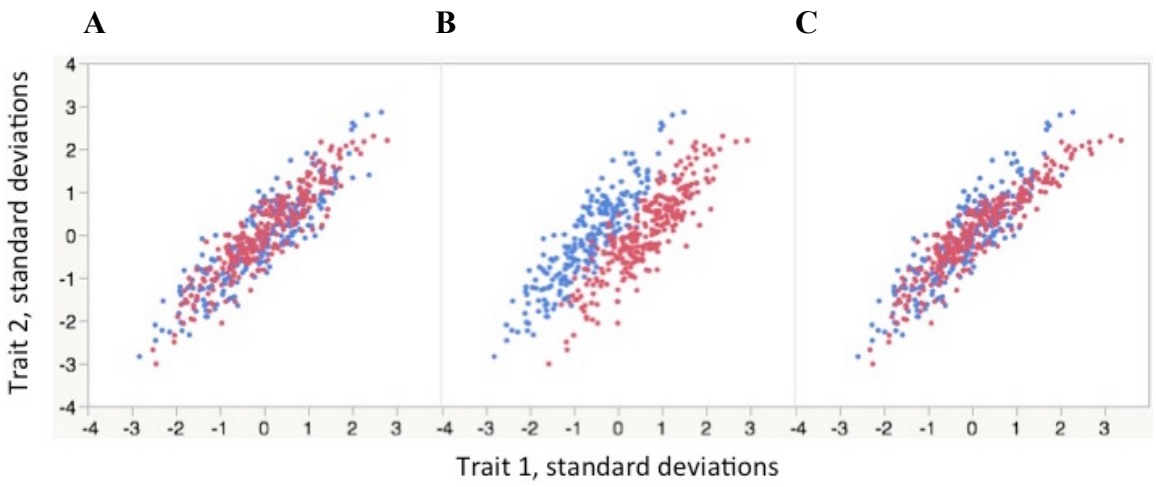
Here we propose that by adding an interaction term to this ordinal regression-based model, we can, under many conditions, increase its sensitivity to a wider spectrum of covariance-modifying genetic effects. While any number of phenotypes and interactions may be assessed, we limit our consideration below to two phenotypes and one interaction (plus any covariates to be adjusted for). In its most basic form, the model we propose and refer to as the ordinal joint interaction model (OJIM) is:

GENOTYPE ~ TRAIT1 + TRAIT2 + TRAIT1*TRAIT2

The interaction term above allows for the detection of heterogeneity of correlation by genotype. Existing mvGWA methods derive their power only by detecting genetically induced changes to the correlation between traits in the data as a whole, and not to differences in correlation among genotypic classes.

This distinction is best illustrated by visualizing a scatterplot of two highly correlated traits (**Figure 5-1**). If a SNP that only affects one of the traits is "inserted" into the population, thereby shifting a (random) set of the points to the right, the correlation between the two traits will clearly be weakened. Existing multivariate methods are better powered than single-trait GWAS to identify such SNPs because of the difference between the direction of such genetic effects and that of the residual correlation. Note that in this particular example, the inserted SNP is *not* pleiotropic; in fact, had it increased both traits equally, the induced change in correlation would have been less pronounced, to the relative disadvantage of mvGWA methods. Note also that the correlation *by genotype* does not change with the introduction of this new SNP; therefore, the interaction term in the ordinal regression model above would not have improved power. On the other hand, if an introduced SNP has an effect on one trait that depends on the value of another trait, then both residual correlation and correlation by genotype will generally be modified. In these cases, adding the interaction term to the multivariate model will increase power (**Figure 5-1**).

Figure 5.1: Existing multivariate genome-wide association methods are sensitive to the genetic effects depicted only by residual correlational changes.

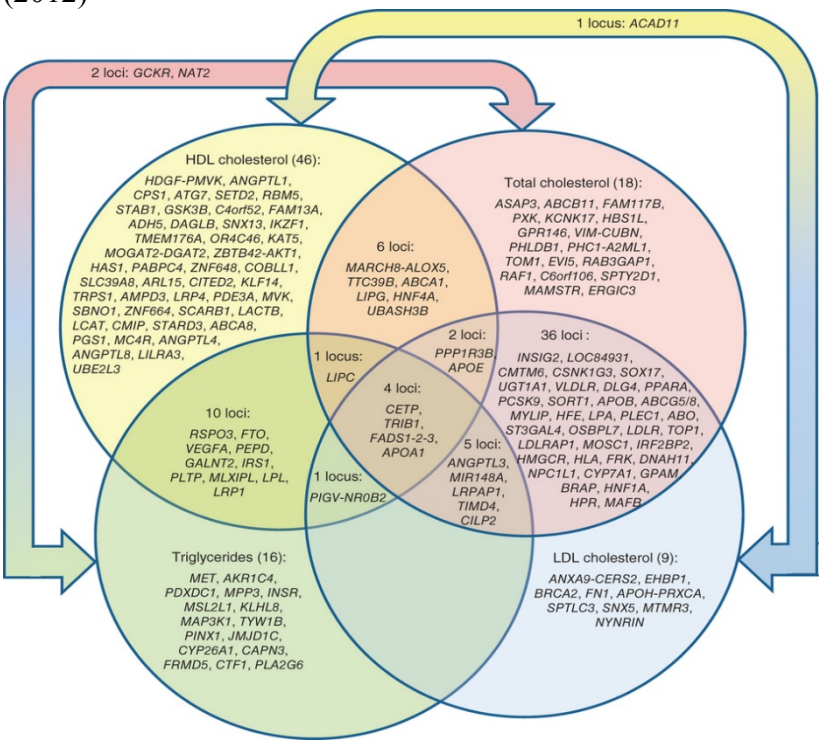


Trait 1 (horizontal axis) and Trait 2 (vertical axis) are highly correlated, but have no causal relationship with each other. **(A)** Red points denote individuals with at least one copy of an allele that is neutral with respect to Traits 1 and 2; **(B)**, the introduction of a new exposure has caused the allele to exert a strong, spontaneous effect on only Trait 1. Note that all points have been re-standardized, such that the increased dispersion of all points taken together reflects the weakened correlation between Traits 1 and 2 in the population as a whole. Note also that the correlations by genotype (red vs. blue) are not significantly different. In **(C)**, the introduction of a new exposure has similarly caused Trait 1 to increase dramatically in individuals with at least one copy of the allele, but in a way that depends on the value of Trait 2. Points have again been re-standardized, revealing a less pronounced change in total correlation (red and blue combined) than in **(B)**, but a significant difference in correlation by genotype (red vs. blue).

How common these context-dependent genetic effects are in nature is an open question, as the influence of genetic variants on phenotypic covariance has not been adequately explored. On the face of it, however, the idea that genes should have predominantly static effects on dynamic systems, such as lipid metabolism, would seem unlikely. It is well known, for example, that individuals with higher levels of plasma triglycerides (TG) typically have lower levels of high-density lipoprotein cholesterol (HDL), and vice versa. TG levels in individuals (as well as in populations) may rise and fall over time, but the negative relationship between TG and HDL persists. Thus, the genetic loci involved in shaping this negative relationship are unlikely to have the same effect on HDL regardless of whether TG is high or low (and vice versa).

We hypothesized that many of the cross-phenotypic associations previously reported for lipid-associated SNPs (**Figure 5-2)** reflect genetic effects that vary with the lipid trait values themselves. To the extent that this is true, the OJIM should outperform univariate and bivariate methods in detecting such loci. Using lipid measurements from1032 Ghanaian participants of the HeART study, we tested our hypothesis by assessing the ability of the OJIM to identify SNPs known to be associated with lipids and other complex phenotypes, and comparing the OJIM's performance to that of conventional analyses.

<u>Figure 5.2</u>: Genes associated with multiple lipid traits. In parentheses: the number of loci associated with only one trait. Figure borrowed from Global Lipids Genetics Consortium (2012)[230]



Methods

Power analyses

To evaluate the power of the ordinal joint interaction model to detect SNPs with context-dependent effects, and to compare its performance to that of other methods, we generated data for 5000 individuals using "Theoretical Model 1" from Chapter 4, setting β_X =0 and R^2_{XZ} to 0.5%. This model simulated phenotypic data for SNPs of moderately strong effect (expected $R^2=2/\pi$ *0.5=0.33 in univariate analysis) dependent on the value of an independent covariate (*Z*). Varying the correlation between *Y* and *Z* in increments of 0.1, we tested 1000 SNPs with the OJIM, with MultiPhen (i.e. the bivariate ordinal model without an interaction term), single-trait linear regression, and single-trait linear regression with a SNP-by-covariate interaction term. Because we knew beforehand that only the outcome variable (*Y*) associated with the SNP, we did not test the SNP for association with the covariate (*Z*) in any of the single-trait regression analyses (with or without interaction). Consequently, we did not correct results of the single-trait tests for multiple testing, making all comparisons with the multivariate methods conservative.

Assessment of Type I error

An advantage of a multivariate approach based on ordinal regression is that it does not assume normality of trait distributions or homoscedasticity. Ordinal regression is also far more robust to outliers than linear methods. To demonstrate this, we used a standard linear model to generate 10 datasets of phenotypic data for 5000 samples, but with the error term set to a *t*-distribution with 8 degrees of freedom to generate outliers (see Chapter 4). The simulated SNPs had no effect on the outcome and were not associated with the covariate. The correlation between the outcome and

covariate was set to 0.30. The distributions were controlled not to deviate too far from normality (Shapiro-Wilk, $p>.001$ for all cases).

Application of the OJIM to lipid traits

Study Cohort Description

Please see Section A of Chapter 3.

Anthropometric measurements and biochemical analysis

Please see Section A of Chapter 3.

Genotyping

A subset of 1105 urban participants from the Ghanaian HeART cohort was selected for genotyping. DNA was genotyped using the Illumina Infinium HumanExome BeadChip platform (Illumina Inc., San Diego, CA). This platform interrogates strictly exonic variants, covering ~240,000 markers.

Quality Control

Approximately 250,000 variants from 1105 participants were available prior to quality controls. We removed all SNPs with a genotyping call rate < 95%. Individuals for whom < 95% of variants were called were removed from analyses. Variants with a minor allele frequency < 20% were also removed, as were variants with a Hardy Weinberg p value < 0.001. Cryptic relatedness was assessed in the data, and one participant in each pair of related individuals (pi-hat >0.2) was randomly removed. Following quality control, 1032 participants and 15,890 variants remained All quality control procedures were performed in PLINK (version 1.07)[231].

Selection of SNPs

We used the Catalog of Published GWAS hosted by the National Human Genome Research Institute (NHGRI) to select SNPs. The regularly updated catalog lists ~15,000 annotated genetic associations from published GWAS for which p<10^-5. We culled all entries mapped to an rs-numbered SNP, of which a final total of 2669 (1) overlapped with the ExomeChip data after QC, and (2) had MAF>0.20 in our samples.

Statistical Analysis

A code was written in R (Appendix D) to identify associations between the 2669 NHGRI SNPs and total cholesterol (TC), triglycerides (TG), HDL, and LDL, using all of the single-trait and multivariate analyses discussed above. In the multivariate analyses, traits were assessed in pairs. All analyses were adjusted for age and sex.

Results

Using simulated data, we compared the power of single-trait and multivariate methods to detect SNPs with moderate effect sizes. The expected proportion of phenotypic variance explained by genotype was set to $2/\pi$ *0.5, or 0.33 (see Methods). The SNPs were entirely context-dependent, in that they had no expected effect when a normally distributed covariate was below its mean, and an effect that increased multiplicatively with the covariate above its mean (see Chapter 4). We used the genome-wide threshold of 5×10^{-8} to determine significance. **Figure 5-5** displays the number of successes out of 1000 simulations over a wide range of correlations. MultiPhen did not generally perform better than single-trait regression. The OJIM performed most consistently, with a success rate of 60% or higher across all correlations. The standard linear SNP-by-covariate interaction model caught up to the OJIM at higher correlations, but this improvement likely reflected the susceptibility of such models to Type 1 error at high correlations, driven in part by increased heteroscedasticity (see Discussion). The linear interaction model also displayed extreme levels of Type 1 error when outliers were present, in contrast to the OJIM (**Figure C-1,** Appendix).

Using exome data from 1032 Ghanaian men and women, we assessed 2669 SNPs from the NHGRI GWAS Catalog for association with TC, TG, LDL, and HDL. Lipid traits were tested individually (applying conventional single-trait regression analysis) and in pairs (applying a bivariate ordinal regression approach with and without an interaction term). Linear interaction models yielded highly inflated p-values (**Figure C-2**, Appendix) and are not presented below. The correlation between LDL and total cholesterol in our samples was 0.91. Thus, tests for the LDL-TC pairing suffered from high multicollinearity. Because TC had substantially less missing data than LDL, and because results for the pairings of LDL-TG and LDL-HDL were broadly

similar to those for TC-TG and TC-HDL, we focus our presentation and discussion below mainly on the results for TC-TG, TC-HDL, and TG-HDL. Results for tests with LDL are discussed where biologically interesting, and can be found in **Tables C-1** and **C-6**.

The TG-TC and TG-LDL tests provided strong support for the hypothesis that the genetic effects on lipid traits are mediated by other lipid traits. Superimposed QQ-plots of all p-values (**Figure 5-6**) and of p-values for only the 116 lipid SNPs that overlapped with our data (**Figure C-3A**) revealed that the OJIM yielded the most significant results, with Type I error comparable to that observed for the univariate tests. The most significant SNP was rs12740374 (p= 3.88 x 10-7 for TG-LDL; p=5.77 x 10-6 for TG-TC), a locus in perfect or nearly perfect linkage disequilibrium with rs7499892, rs629301, rs646776, and rs660240 (**Table C-1 and Table 5-1**). The minor haplotype group formed by these five SNPs has been associated with lipids and cardiovascular phenotypes in multiple studies across several populations (see Discussion), and moreover, has been shown to increase the expression of *SORT1* in the liver, which mediates LDL and very low-density lipoprotein (VLDL) production.[232] The OJIM's interaction p-values for the SNPs in the haplotype ranged from 0.001 to 0.0002 (TG-TC tests) and 0.006 to 0.002 (TG-LDL tests) (**Table C-1 and Table 5-1**). In accordance with the significant interaction, TG-TC and TG-LDL correlations differed by genotype. **Figure 5-7** shows that the minor allele of rs12740374 attenuated the positive association between LDL and TG. More specifically, the LDL-lowering effect of rs12740374 became evident only in individuals with above-average TG (**Figure 5-8**).

The OJIM also detected the lipoprotein lipase (*LPL*) gene, which converts VLDL to LDL. Although the interaction term was not significant for any of the four *LPL* SNPs, the joint interaction p-values were nonetheless most significant for two of them (**Table 5-1**). In the TG-

TC tests, only one lipid-associated gene identified by either univariate or bivariate analysis was not among the top ten associations for the OJIM (*BCHE*-rs1803274). However, *BCHE*-rs1803274 was the seventh most significant OJIM result in the TG-LDL tests (**Table 5-1**, **Table C-1**, and **Table C-2**). Six of the top ten results for the interaction term alone (including rs12740374 and rs646776) were associated with genes identified by GWAS of cardiovascular traits; in addition to rs12740374 and rs646776 described above: rs1829883, *CACNB2*-rs7076247*, IGF2AS*-rs1004446*, and THADA*-rs6732426 (**Table C-3**).[233-236]

The added interaction term provided less of an overall power advantage for the TG-HDL tests (**Figure C-2B** and **Figure C-3B**), but the OJIM still displayed the most power to detect lipid-associated loci (**Table 5-2**, **Table C-4**, and **Table C-5**). In particular, the OJIM identified four lipid-associated genes (*CETP*, *LIPC, APOA*, and *KLHL8*) among its top ten results, while the other methods detected only *CETP* and *LIPC.* The top HDL-TG result was rs7499892 in the *CETP* gene, which encodes the fundamental enzyme in TG and HDL metabolism, cholesteryl ester transfer protein. The interaction p-value for rs7499892 was only 0.06, but here, too, the joint interaction p-value was lowest of all methods tested (6.87 x 10-05) (**Table C-2**). Only the OJIM identified the fundamental *APOA*-cluster (*APOA1, APOA3, APOA4, APOA5, ZNF259,* and *BUD13*) among its top ten results (**Table C-2** and **Figure 5-4**). The significant interaction p-value (0.01) captured a recessive genetic effect on the correlation between TG and HDL (**Figure 5-9**). Further exploration revealed that the association between rs4938303 and HDL in the univariate analysis (p=0.002) was driven almost entirely by homozygote recessive individuals in the fourth quartile of TG (**Figure 5-10**).

The OJIM's second best association was rs1532085 in the hepatic lipase gene *LIPC*, another key factor of HDL and TG metabolism. Although the joint interaction p-value for

rs1532085 (0.0001) was more significant than the univariate p-values (0.003 for HDL, 0.011 for TG, unadjusted), the interaction term was not significant (p=0.74). MultiPhen therefore had the most power ($p=3.4 \times 10^{-5}$), indicating that the effect was likely independently pleiotropic (**Table 5-2**). The third best OJIM result was rs610604 in *TNFAIP3,* a gene previously associated with cardiac troponin-T levels and atherosclerosis in African Americans.[237] Three of the top ten interaction p-values were for loci related to cardiovascular traits: *NOS1AP*-rs2880058,[238] GHR-rs13188386,[239] and *KLHL8*-rs442177, a SNP previously associated with TG in multiple studies (discussed below).

The HDL-TC and HDL-LDL did not provide evidence for context-dependent genetic effects (**Figure C-3C**). *CETP* and *SORT1* were the two best associations for all methods, but no other lipid-associated genes were identified (**Table 5-3**). All methods performed comparably; the top ten MultiPhen and OJIM associations featured the same SNPs (in different orders), while only one of the top ten univariate results was unique, though unrelated to cardiovascular disease (**Table 5-3** and **Table C-7**).

Discussion

If a genetic variant (*X*) is introduced into a population, such that its effect on Trait *Y* depends on the value of Trait *Z*, then *X* will also modify the correlation between Traits *Y* and *Z*. Existing multivariate genome-wide association (mvGWA) methods derive their comparative advantage over single-trait analyses on account of their sensitivity to just such modifications. However, the correlation between Traits *Y* and *Z* will also differ by genotype, i.e. by the number of copies of *X*. We propose here that multivariate methods sensitive not only to the effects of *X* on residual correlation, but also to this correlational heterogeneity, will have increased power to detect context-dependent variants. The ordinal joint interaction model (OJIM) described here meets both criteria.

In Chapter 4, we developed biologically plausible models of context-dependent quantitative trait loci (QTL). Here, using the first of those models, we simulated data with which to assess the performance of the OJIM, bivariate ordinal regression (MultiPhen), univariate linear regression, and linear regression with a SNP-by-covariate interaction term. Because the effect of the QTL on the outcome phenotype was simulated to depend on a covariate, we expected the OJIM to outperform univariate and bivariate analyses. It did so, showing remarkable consistency over all correlations, and achieving genome-wide significance ~60% of the time (**Figure 5-5**). In contrast, the univariate analyses detected the simulated QTL at a genome-wide level of significance only about 10% of the time (as expected, given the QTL's marginal effect size of $2/\pi$ *0.5; see Methods) (**Figure 5-5**). Univariate results were approximately the same across all correlations, because there was no interacting covariate in these analyses. We ran only one univariate test per SNP, because we knew beforehand which the dependent variable was. Had we

not known this, adjustment for multiple testing would have been required. Thus, all comparisons with single-trait tests here are conservative.

We did not expect MultiPhen to perform as well as the OJIM, because estimates of marginal effects capture only a fraction of the total effect of a context-dependent QTL. Indeed, MultiPhen detected the QTL at a genome-wide level of significance less than 10% of the time over most correlational classes (**Figure 5-5**). However, we expected its performance to improve as the correlation between the outcome and covariate increased. For, when a QTL is associated with only one of two traits, its effect on the residual correlation is stronger when that correlation is stronger. This was also observed (**Figure 5-5)**.

The conventional way to identify genetic variants with context-dependent effects is to assess the significance of a gene-by-covariate interaction term in a regression model. However, we saw in Chapter 4 that even when a genetic effect is highly dependent on a covariate, reliance on the SNP-by-covariate interaction term alone is often not sufficient. In contrast, the OJIM assesses marginal effects, while simultaneously allowing for genetic interactions with any number of traits, making it a simple but comprehensive joint test of interaction. Joint tests based on single-outcome regression are not nearly as straightforward to design or implement, and are further complicated by inflated Type 1 error, as described below.[240,241] All single-outcome interaction models also require *a priori* selection of a dependent outcome variable. Such decisions may be straightforward enough with most gene-by-environment interactions, but with complex phenotypes, such as cardiovascular risk factors, the direction of interaction can often be counterintuitive. The very concept can also be misplaced, as when two phenotypes feed back into each other.

In our simulations, we knew beforehand that the modeled SNPs only interacted with one phenotype (*Z*), and, moreover, that they did not have pleiotropic effects (which single-outcome regression would not have had the power to detect). Despite these advantages, the interaction term achieved genome-wide significance less than 40% of the time when correlations between the outcome and covariate were 0.4 or below (**Figure 5-5**). P-values improved for the strongest correlations (>0.50), but inflation was likely a major factor. Attention has recently been drawn to the fact that conventional linear interaction models, such as gene-by-environment models, are highly susceptible to Type I error for structural, not empirical reasons.[242,243] We observed massive inflation when the linear interaction model was tested on simulated data that contained outliers (**Figure C1,** Appendix). Moreover, when we tested the linear interaction model on real data, Type I error for the interaction term p-values was likewise extreme (**Figure C2,** Appendix). We have therefore omitted the results for these tests from the main body of presentation.

The QTLs in the simulations above were not pleiotropic. In contrast, O'Reilly et al. tested MultiPhen by simulating QTLs that were pleiotropic (they affected two traits independently), but which did not have an interaction with either.[226] In such cases, the OJIM would still be expected to pick up the pleiotropic effects, but the extra degree of freedom for the superfluous interaction term would diminish its power. We found that this loss of power rarely exceeded half an order of magnitude when we used the same range of trait correlations and simulation parameters as O' Reilly et al (N=5000, MAF=0.2, expected variance explained per SNP $\leq$ 0.5%) (data not shown). For the reasons discussed above, MultiPhen performed worse than univariate regression (run separately for both traits and corrected for multiple testing) over some ranges of correlations between the traits (see O'Reilly et al.). For example, when the simulated pleiotropic SNP had equal, independent effects on two traits, univariate regression performed best for correlations

greater than 0.5. Expectedly, the OJIM performed worse than univariate linear regression over the same general ranges as MultiPhen in these simulations.

Clearly, with respect to statistical power, no model is ideally suited for every situation. Standard GWAS will generally have the strongest power to detect SNPs that affect only one of two weakly correlated traits, or SNPs that have independent effects on two traits in the same direction as their residual correlation. On the other hand, a SNP with independent effects on two weakly correlated traits will be most amenable to discovery by a standard mvGWA approach, including MultiPhen. Even in these cases, however, when the OJIM does not increase statistical power, it does nonetheless offer the advantage of providing information about genetic architecture that would otherwise be unavailable. For, significant or not, the p-values for the interaction term, for each predictor variable, and for the overall model fit, are all potentially meaningful. A significant interaction term indicates that the phenotypes have different relationships with each other by genotype, possibly on account of context-dependent genetic effects. If only the overall model p-value, but not the interaction p-value, is significant, the SNP may have independent pleiotropic effects on multiple traits, but its effect is unlikely to be dependent on their relative values. If only the interaction term is significant, and not the beta coefficients of the traits themselves, then we can further conclude that the SNP strengthens (or weakens) the relationship between the traits in a general way (i.e., not only in one direction). Thus, when the covariates are carefully chosen and the p-values carefully considered, the OJIM can be used not only to discover meaningful SNPs, but also to gain insight into genetic architecture.

On the other hand, all but a handful of the variants listed in the NHGRI Catalog have been discovered by univariate GWAS, and consequently provide very little information about

genetic architecture. Because marginal effects can account for a large proportion of genetic variance even when genetic effects depend entirely on other covariates, the degree to which the SNPs in the NHGRI Catalog are context-dependent remains an open and empirical question. To address it, we compared the relative power of the OJIM, MultiPhen, and univariate regression to detect the NHGRI Catalog SNPs previously associated with lipid traits and related cardiovascular phenotypes. To the extent that a gene's effect on lipid levels is mediated by other lipid levels (either directly or epiphenomenally), the OJIM's results should reflect it.

We chose lipid traits as our phenotypes for several reasons. First, SNPs associated with lipids are among the most numerous and best replicated in the NHGRI Catalog. Second, the complexity of lipid metabolism suggests that lipid traits are unlikely to be independent of each other at the physiological level. They may, for instance, interact in the context of metabolic pathways, or induce reciprocal changes at the level of gene expression. At the same time, lipid metabolism is not unmanageably complex. Many of its pathways and molecular mechanisms are well characterized, and much of the variation therein appears to be adequately (if crudely) captured by the four conventional lipid measurements[244]. It has also been suggested that pleiotropy is pervasive among SNPs associated with lipid traits. However, because they have generally been tested one at a time, some of these cross-trait associations may be merely artifactual. Alternately, some of these loci may truly have pleiotropic effects, but the pleiotropy may itself be context-dependent.[245] For example, a variant may be involved in a pathway that simultaneously raises HDL and lowers TG, but it may require HDL or TG to reach certain levels before its expression is activated. The OJIM, being sensitive to both pleiotropy and context-dependence, would correspondingly be most powered to identify them.

Lipid traits make up approximately 5% of the NHGRI Catalog of GWAS hits, while the proportion of traits related to cardiovascular disease (though a subjective calculation) make up approximately 20% when a wide set of criteria are applied. We did not limit our tests to only these SNPs, however, but considered all of the SNPs in the GWAS Catalog (that overlapped with our exome data), because functional variants tend to be pleiotropic, and pleiotropy can occur in unpredictable ways.[219] Moreover, by testing even SNPs associated with those phenotypes furthest removed from cardiovascular disease, we can gain additional insight into the OJIM's performance by assessing the overall distributions of p-values.

In our analyses, the strongest support for the hypothesis that the genetic effects on lipid traits are mediated by other lipid traits was provided by the TG-TC and TG-LDL tests. A comparative distribution of p-values for all TG-TC tests revealed that the OJIM yielded the most significant results, with minimal Type I error (**Figure 5-6**). Moreover, its top associations were enriched for lipid-associated SNPs (**Table 5-1 and Figure C-3A**). Among the top ten associations for the TG-TC and TG-LDL univariate and bivariate analyses, only one lipid-associated SNP (*BCHE*-rs1803274) was missed by the OJIM, and that only with the TG-TC pairing (**Table 5-1**, **Table C-1**, and **Table C-2**).

The most significant SNP was rs12740374, which had a joint interaction p-value of 3.88×10^{-7} for TG-LDL and 5.77×10^{-6} for TG-TC (**Table C-1** and **Table 5-1**). The locus was in perfect linkage disequilibrium (LD) with rs7499892, and in nearly perfect LD with three other SNPs (rs629301, rs646776, and rs660240). The minor haplotype group formed by these five SNPs, located on chromosome 1p13.3, has been shown to be associated with lower LDL levels in many populations,[246-253] as well as with reduced risk for coronary artery disease,[254-258] myocardial infarction,[258,259] coronary artery calcification,[260] and other cardiovascular

phenotypes. Although requiring Bonferroni multiple testing correction to the results of this study would be unreasonable in light of the widespread LD among the interrogated SNPs and the prior knowledge of function, the p-values for these OJIM associations were nonetheless well below that threshold ($p=1.9 \times 10^{-5}$). Importantly, the SNPs of 1p13.3 stood out in this study because of their highly significant *joint* interaction p-values. A study that assessed only the marginal effect of rs646776 on LDL (p=0.0002), for example, or only the interaction effect of rs646776*triglycerides on LDL (p=0.003), may have passed over the locus.

Whereas the causal genes and functional roles of most lipid-related GWAS SNPs are still unknown, the five SNPs in 1p13.3 are notable exceptions, having been shown by a recent study to increase the expression of *SORT1* in the liver.[232] Fine mapping revealed that rs12740374 was in fact the functional variant, its minor allele creating an enhancer for *SORT1*.[232] Studies assessing the connection between the gene product sortilin and LDL metabolism, however, have yielded conflicting results.[232,261-265] In some studies, including human studies, increased expression of hepatic *SORT1* resulted in lower VLDL production.[232,266] Because VLDL is converted to LDL in the blood, VLDL reduction could explain the reduced plasma LDL levels in individuals with the minor allele rs12740374-T (as also observed here). However, mechanisms by which *SORT1* overexpression leads to *increased* LDL, such as by facilitating VLDL secretion or by targeting LDL receptors for lysosomal destruction, have also been well characterized.[262,264] Moreover, studies in mice have found that sortilin deficiency, as by *SORT1* knockout, reduced VLDL production, and drastically so.[264,266] Thus, whereas all studies point to the 1p13 locus as a promising target for therapeutic intervention, the conditions under which *SORT1* and its product sortilin (termed a "many-headed hydra" in a recent editorial[267]) lower LDL are far from clear (**Figure 5-3**).

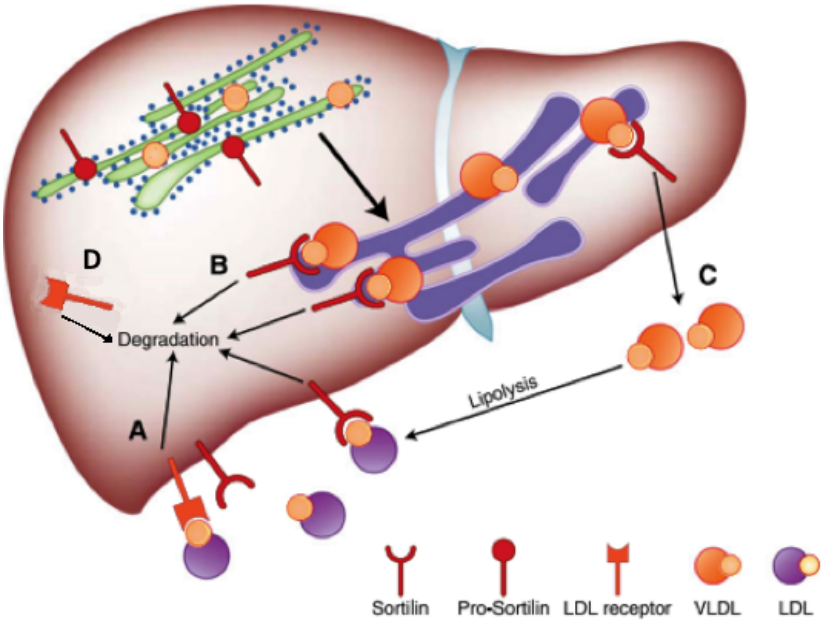


Figure 5-3. Mechanisms by which sortilin can influence LDL levels. Figure adapted from Strong et al. (2012).[263]

Very low-density lipoprotein (VLDL) synthesis begins in the rough endoplasmic reticulum (green) and is completed in the Golgi apparatus (blue). VLDL (orange) is either degraded or secreted. When secreted, VLDL is lipolyzed to generate LDL (purple) and taken up by the hepatic LDL receptor.

(A) *SORT1* overexpression decreases LDL by increasing hepatic uptake of LDL participles (Linsel-Nitschke et al.)[265];

(B) *SORT1* overexpression decreases LDL by reducing production and/or secretion of VLDL (Musunuru et al.)[232];

(C) *SORT1* overexpression increases LDL by facilitating secretion of VLDL (Kjolby et al.)[264];

(D) *SORT1* overexpression increases LDL by targeting LDL receptors for destruction (Gustafsen et al).[262]

In that regard, the OJIM is not only sufficiently powered to detect a context-dependent gene like *SORT1*, but can also can provide insight into its mechanism. That 1p13.3 was detected in tests of TG-LDL and TG-TC is noteworthy, because *SORT1* influences LDL (at least in part) by its more direct influence on VLDL (**Figure 5-3**), and almost all of the triglycerides in the blood are carried by VLDL. However, no study to date has found an association between rs12740374 and TG, despite the biological plausibility of such a connection.[268] In our univariate tests of TG, the p-value for rs12740374 was 0.97 (N=1032, **Table 5-1**). Yet, while rs12740372 may not influence TG, TG does appear to influence the effect of rs12740374 on LDL (interaction p-value= 0.0016) (**Table 5-1**).

We mentioned that the interaction term of the OJIM is sensitive to heterogeneity of correlation by genotype. Indeed, we see in **Figure 5-6** that the expected correlation between LDL and TG (r = 0.27 for major allele homozygotes at rs12740374), becomes statistically indistinguishable from zero when one or two copies of the minor allele are present. **Figure 5-7** provides additional insight into how the effect of rs12740374 on LDL may vary with TG**.** In individuals homozygous for the major allele, the increase in mean LDL by TG quartile is just as we would expect, given that increased TG implies increased VLDL, and VLDL is lipolyzed into LDL. In particular, individuals with the lowest TG levels (first quartile) had LDL levels almost one-half standard deviation below the mean, whereas those with above-average TG levels (third and fourth quartiles) had LDL levels approximately one-half standard deviation above the mean. However, the minor allele at rs12740374 appeared to disrupt this relationship. Interestingly, its LDL-lowering effect became evident only in the third and fourth quartiles of TG, when LDL "should" have been higher. These results suggest that the observed reduction in LDL may have more to do with accelerated LDL clearance (perhaps after VLDL reaches a certain level) than

with VLDL reduction or destruction per se (**Figure 5-3**). Future experiments assessing the directionality of sortilin's effects should therefore factor in triglyceride levels. The contradictory conclusions of previous studies may be reconcilable if cryptic variation in background lipid levels are accounted for, such as may be caused by differences in experimental method (e.g. knockout, knockdown and overexpression protocols) or noted differences in diet and mouse models.[263,267]

While the experimental evidence connecting *SORT1* and VLDL makes it tempting to speculate that TG mediates the LDL-lowering effects of rs12740374 directly, the correlational patterns observed here may very well be the epiphenomena of other unknown factors. It is, in fact, a merit of the OJIM that the phenotypes being assessed need not be at all causally related, as such a requirement would severely limit its utility. Rather, its power is enhanced whenever there are differences in phenotypic correlation by genotype, whatever the mechanism, and regardless of whether the phenotypes are distant proxies for an underlying set of unknown factors. Such indirect genetic modifications to correlations may very well be ubiquitous, given that arbitrarily small perturbations in complex systems, such as physiological systems, typically propagate across entire networks of interactions.

Interestingly, the other set of major lipid-associated SNPs identified by the TG-LDL and TG-TC tests were variants in the lipoprotein lipase (*LPL)* gene, which converts VLDL into LDL. Although the OJIM's interaction term was not significant for any of the four associations (p=0.14, 0.08, 0.09, and 0.44 in the TG-TC tests), its full model p-value was nevertheless the most significant for two of them (**Table 5-1**). Somewhat surprisingly, MultiPhen did not significantly outperform even univariate analysis in detecting lipid-associated SNPs in the TG-

TC and TG-LDL tests (**Table 5-1**, **Table C-1**, and **Table C-2)**. It is possible, therefore, that genetic effects on TG, TC, and LDL are more context-dependent than pleiotropic.

With the TG-HDL tests, MultiPhen's performance improved, while the added interaction term provided less of a power advantage overall (**Figure C-2B** and **Figure C-3B**). Although this may provide nominal evidence that genetic effects on TG and HDL are not particularly dependent on relative TG and HDL levels, the small sample sizes for tests with HDL (N=869) preclude any strong conclusions. More importantly, when only the top results for each of the methods were considered, the OJIM still displayed the most power to detect lipid-associated loci (**Table 5-2**, **Table C-4**, and **Table C-5**). In particular, of the four lipid-associated genes (*CETP, LIPC, APOA,* and *KLHL8*) identified by at least one of the methods, only the OJIM identified all four; the others missed *APOA* and *KLHL8*.

The top HDL-TG result, rs7499892, is a SNP in *CETP*, the gene that encodes the key HDL remodeling factor, cholesteryl ester transfer protein. A large number of studies have shown that *CETP* is strongly associated with lipid traits as well as with cardiovascular risk factors.[50,269,270] Although the OJIM's interaction p-value for rs7499892 was only 0.06, its full model p-value was nonetheless the lowest of all methods tested (6.87×10^{-05}) (**Table C-2**). The enhanced power of the OJIM to detect *CETP* makes sense in light of the fact that cholesteryl ester transfer protein (CETP) facilitates transfer of cholesteryl ester (CE) from HDL to triglyceride-rich lipoprotein in exchange for triglycerides.[269] Moreover, *in vitro* evidence indicates that CE transfer from HDL is increased in plasma from hypertriglyceridemic individuals. Accordingly, plasma TG levels have been shown to correlate with the rate of cholesterol esterification, net CE transfer, and HDL remodeling.[271,272]

The OJIM was also the only method to detect the important *APOA*-cluster (*APOA1, APOA3, APOA4, APOA5, ZNF259,* and *BUD13*) among its top ten results (**Table C-2** and **Figure 5-4**). Apolipoprotein A1 (APOA1) is the major protein component of HDL in plasma (its deficiency is one of only three known Mendelian disorders of HDL metabolism),[269] while APOA5 is also a major component of HDL as well as VLDL, which transports TG. Interestingly, the SNP detected here (rs4938303), which was approximately 35,000 base pairs upstream of the closest gene in the region (**Figure 5-4**), was only significantly associated with HDL in our univariate analyses (p=0.002), whereas it has previously been shown to associate with TG.[273] The OJIM's TG-HDL interaction p-value was not especially significant (p=0.01), but it did point to a possible context-dependent effect, which we explored further.

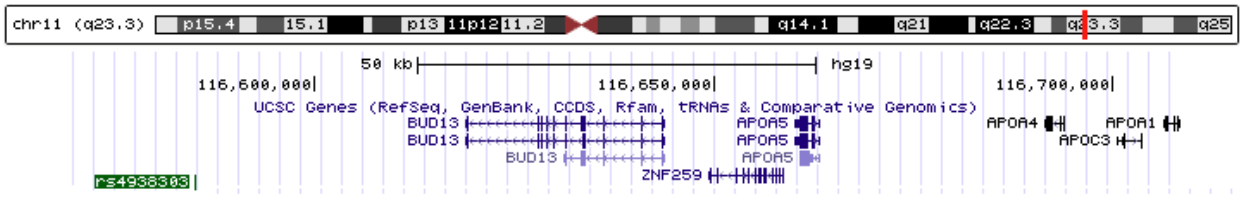


Figure 5-4. The ZNF259/BUD13 region of Chromosome 11 (q23.3), which includes the apolipoprotein genes APOA5, APOA4, APOC3, and APOA1.

Stratifying participants by genotype revealed a recessive genetic effect on the correlation between TG and HDL (**Figure 5-9**). Specifically, the TG-HDL correlation was -0.20 among both major allele homozygotes and heterozygotes, but strengthened to -0.38 among minor allele heterozygotes. In contrast to the similar analysis of rs12740374 above, the sample sizes by genotype were more equitably distributed here, owing to the larger minor allele frequency of rs4938303 (48%). This allowed a meaningful pattern to emerge in the homozygote recessive group; namely, the stronger correlation therein appeared to be driven, at least in part, by the ~5% of individuals with very low HDL levels. In fact, an assessment of mean HDL by TG quartile for each genotype showed that the association between rs4938303 and HDL in the univariate analysis (**Figure 5-10**, top panel) was driven almost entirely by homozygote recessive individuals in the fourth quartile of TG (**Figure 5-10**, bottom panel).

Although inferences drawn from statistical results that are disproportionately influenced by a relatively small number of samples should be treated cautiously, with the understanding that they may not be generalizable, the overwhelming abundance of epidemiologic and biological evidence linking the *APOA* cluster to lipid metabolism allows for some speculation. For, one may almost *expect* a SNP that tags this locus to associate with HDL. Thus, if the association observed here is completely artifactual, then it must be considered a remarkable coincidence. On the other hand, from a Bayesian perspective, if the association is not artifactual, then the context-dependent effect we have observed is likely to be real. For, if TG actually has no influence on the genetic effect of rs4938303, then we have observed a true association driven almost entirely by an artifactual effect.

The HDL-TC and HDL-LDL tests provided the weakest evidence for context-dependent genetic effects (**Table 5-3** and **Figure C-3C**). Indeed, the small sample sizes (N=869)

notwithstanding, the distribution of p-values for the OJIM interactions were so far below the expected null distribution (**Figure C-3C**), that the results may be taken as evidence of a *lack* of such effects. While it is well established that changes in plasma HDL rarely occur without any concomitant changes in triglycerides,[274] perhaps HDL and LDL are more independent, at least with respect to their genetic architecture. However, it is important to keep in mind that the NHGRI GWAS SNPs were discovered by virtue of their marginal effects. Thus, although strong marginal effects by no means rule out the possibility of context-dependence, SNPs with truly independent effects (to the extent they exist) will nonetheless be overrepresented in the NHGRI Catalog.

For that reason, we particularly did not expect lipid-associated SNPs to have significant interaction effects in the complete absence of marginal effects. However, we did observe one such SNP (rs442177) in our HDL-TG tests, which multiple studies have shown to be associated with TG[50,230,273] (**Table C-5**). The SNP would have been difficult to detect with traditional methods, as illustrated by **Figure 5-11** (which may be compared to **Figure 4-2** in the previous Chapter). We see that the minor allele of rs442177 was associated with lower TG when HDL was below average, and higher TG when HDL was above average, canceling out any net effect. If this context-dependent effect is in fact real, then the strong marginal associations with TG in previous studies would need to be explained. One possibility may relate to the fact that HDL levels were exceptionally low in our study population; in fact, 44% of urban men and 59% of urban women had clinically low levels (**Table S2, Appendix A**). If, in those studies that did detect a marginal association with TG, the HDL distributions were shifted to the right, then the TG-raising properties of rs442177 would have been more pronounced. While such an

explanation is highly speculative, it would at least be consistent with the direction of the minor allele's effect in all studies.

It is also worth noting that rs442177 was one of the "triglyceride-increasing alleles" that (collectively) associated with protection against Type 2 Diabetes (T2D) in a recent study[275]. Because low HDL may be considered a risk factor for T2D, a SNP that reduces TG—itself a risk factor for T2D—when HDL is low would indeed be expected to confer some protection to T2D. Thus, perhaps what we have observed for rs442177 is a more general phenomenon. Interestingly, the study also reported a significant *interaction* between the TG-increasing alleles and TG levels. Since higher TG levels imply lower HDL levels, this result would also be in line with our finding for rs442177. Future studies of this phenomenon should therefore take HDL concentrations into account.

Conclusion

Our lipid trait analyses suggest that many of the cross-phenotypic associations identified by GWAS in recent years may reflect context-dependence as well as pleiotropy. Although the association of dyslipidemias with cardiovascular disease is well established, the mechanistic links between them are far from clear, and the SNPs that have strong marginal associations with cardiovascular disease may associate with lipids only via interactions. We have introduced a flexible tool that can detect such interactions, as well as provide insight into the degree to which they exist. Indeed, in addition to increasing power to detect genetic variants likely to be context-dependent, the OJIM can be adapted to test specific hypotheses and answer different questions. Although in our analyses we have focused on identifying meaningful variants, perhaps a more pressing goal in genetic epidemiology is the elucidation of function, particularly of SNPs already known to be significant. Varying trait combinations rather than genetic loci could thus be highly informative. Such an approach may be extended into a phenome-wide screen, as a complement to phenome-wide association testing (PheWAS).[276]

Figure 5-5. Power comparison of single-trait and multivariate methods. The phenotypic data (N=5000) was simulated based on a locus with effects contingent on a normally distributed, independent variable correlated with the phenotype at levels denoted in legend. The expected proportions of variance explained by the locus and interaction were 0.33 and o0.5 respectively. The vertical axis depicts the number of times out of 1000 simulations results were significant at a genome-wide level.

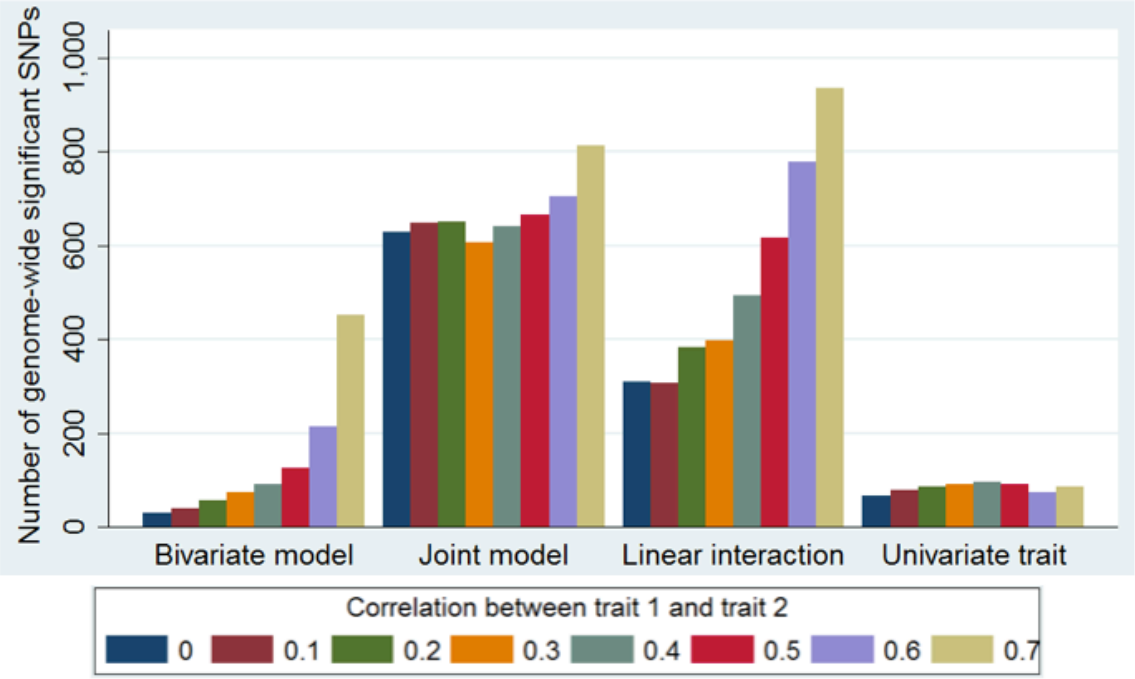


Table 5-1. Triglycerides and total cholesterol: top ten associations for the ordinal joint interaction model. 2669 exonic SNPs from the NHGRI GWAS Catalog were assessed for 1032 Ghanaian participants.

				p-value					
SNP	Chr.	Minor Allele	MAF	Single Trait TG	Single Trait TC	Bivariate	OJIM	Interaction	Gene
rs12740374	1	T	0.26	0.9658	0.0002	0.0003	**5.77E-06**	0.0010	*SORT1**
rs646776	1	C	0.37	0.7855	0.0018	0.0041	**1.36E-05**	0.0002	*SORT1**
rs204993	6	G	0.29	0.0002	0.9769	0.0004	9.92E-05	0.0199	*PBX2*
rs9990343	3	G	0.49	0.0094	0.0011	0.0008	0.0003	0.0324	*CCR3*
rs301	8	C	0.35	0.0001	0.6917	0.0003	0.0003	0.1436	*LPL**
rs3803064	12	A	0.28	0.4633	0.8875	0.6758	0.0004	2.70E-05	*RPH3A*
rs326	8	A	0.39	0.0029	0.0150	0.0023	0.0017	0.0847	*LPL**
rs229527	22	A	0.34	0.2929	0.0036	0.0007	0.0021	0.7133	*C1QTNF6*
rs331	8	A	0.42	0.0007	0.5991	0.0009	0.0021	0.4428	*LPL**
rs10096633	8	C	0.50	0.0068	0.0082	0.0035	0.0027	0.0920	*LPL**

Note: SNPs in perfect linkage disequilibrium were not listed: rs7528419 (with rs12740374); rs660240 and rs629301 (with rs646776); rs176095 (with rs204993).

Genes marked with an asterisk have been previously associated with lipids by GWAS; p-values in bold were significant after Bonferroni correction; all tests were adjusted for age and sex; single trait tests were not adjusted for multiple testing (i.e. 2 tests).

Columns**:** Chr. = chromosome; MA = minor allele; MAF = minor allele frequency; TG = triglycerides; TC = total cholesterol; Bivariate = MultiPhen, which models genotype as a function of TG and TC; OJIM adds an interaction term; Interaction = interaction term for the OJIM.

Table 5-2. HDL and triglycerides: top ten associations for the ordinal joint interaction model. 2669 exonic SNPs from the NHGRI GWAS Catalog were assessed for 869 Ghanaian participants.

				p-value					
SNP	Chr.	Minor Allele	MAF	Single Trait HDL	Single Trait TG	Bivariate	OJIM	Interaction	Gene
rs7499892	16	T	0.44	8.83E-05	0.0146	9.83E-05	6.87E-05	0.0637	*CETP**
rs1532085	15	G	0.40	0.0027	0.0113	3.37E-05	0.0001	0.7394	*LIPC**
rs610604	6	T	0.29	0.1143	0.0025	0.0013	0.0003	0.0160	*TNFAIP3*
rs247616	16	T	0.26	**1.49E-05**	0.0285	0.0002	0.0003	0.2219	*CETP**
rs4938303	11	T	0.48	0.0022	0.5235	0.0039	0.0007	0.0142	*BUD13**
rs1335532	1	A	0.40	0.0002	0.0336	0.0003	0.0009	0.4595	*CD58*
rs2548145	5	G	0.29	0.0040	0.2453	0.0011	0.0012	0.1392	*Loc285634*
rs204993	6	G	0.29	0.0416	0.0001	0.0005	0.0014	0.6017	*PBX2*
rs7769051	6	A	0.38	0.0214	0.0071	0.0075	0.0014	0.0163	*SNORA33*
rs9268877	6	A	0.35	0.3279	0.0012	0.0010	0.0014	0.1957	*HLA-DRB9*

Genes marked with an asterisk have been previously associated with lipids by GWAS; p-values in bold were significant after Bonferroni correction; all tests were adjusted for age and sex; single trait tests were not adjusted for multiple testing (i.e. 2 tests).

Columns**:** Chr. = chromosome; MA = minor allele; MAF = minor allele frequency; HDL = high-density lipoprotein cholesterol; TG = triglycerides; Bivariate= MultiPhen, which models genotype as a function of HDL and TG; OJIM adds an interaction term; Interaction = interaction term for the OJIM.

Table 5-3. Total cholesterol and HDL: top ten associations for the ordinal joint interaction model; 2669 exonic SNPs from the NHGRI GWAS Catalog were assessed for 869 Ghanaian participants.

				p-value					
SNP	Chr.	Minor Allele	MAF	Single Trait TC	Single Trait HDL	Bivariate	OJIM	Interaction	Gene
rs12740374	1	T	0.26	**4.29E-06**	0.7565	**7.53E-06**	**1.99E-05**	0.3485	*SORT1**
rs7499892	16	T	0.44	0.6321	6.59E-05	3.60E-05	0.0001	0.4816	*CETP**
rs646776	1	C	0.37	0.0003	0.5222	0.0001	0.0004	0.5391	*SORT1**
rs247616	16	T	0.26	0.1110	4.51E-05	0.0002	0.0006	0.3961	*CETP**
rs261360	20	A	0.41	0.8527	0.0004	0.0005	0.0013	0.4687	*RPS21P7*
rs1335532	1	A	0.40	0.2243	0.0002	0.0008	0.0023	0.7354	*CD58*
rs13361189	5	T	0.50	0.1378	0.0107	0.0011	0.0017	0.2126	*IRGM*
rs1024020	4	T	0.44	0.0025	0.4817	0.0013	0.0033	0.5166	*intergenic*
rs9990343	3	G	0.49	0.0003	0.0698	0.0013	0.0027	0.3296	*CCR3*
rs1219648	10	G	0.44	0.0003	0.1994	0.0016	0.0048	0.7366	*FGFR2*

Note: SNPs in perfect linkage disequilibrium were not listed: rs7528419 (with rs12740374); rs660240 and rs629301 (with rs646776).

Genes marked with an asterisk have been previously associated with lipids in GWAS; p-values in bold were significant after Bonferroni correction; all tests were adjusted for age and sex; single trait tests were not adjusted for multiple testing (i.e. 2 tests).

Columns**:** Chr. = chromosome; MA = minor allele; MAF = minor allele frequency; TC = total cholesterol; HDL = high-density lipoprotein cholesterol; Bivariate= MultiPhen, which models genotype as a function of TC and HDL; OJIM adds an interaction term; Interaction = interaction term for the OJIM.

Figure 5-6. QQ-plots of p-values for tests assessing 2269 NHGRI SNPs for association with triglycerides and/or total cholesterol in 1032 Ghanaian participants. Univariate tests of triglycerides (purple); univariate tests of total cholesterol (brown); joint tests of triglycerides and total cholesterol with MultiPhen (green) and the ordinal joint interaction model (OJIM) (black triangle); tests of the OJIM interaction term only (grey cross).

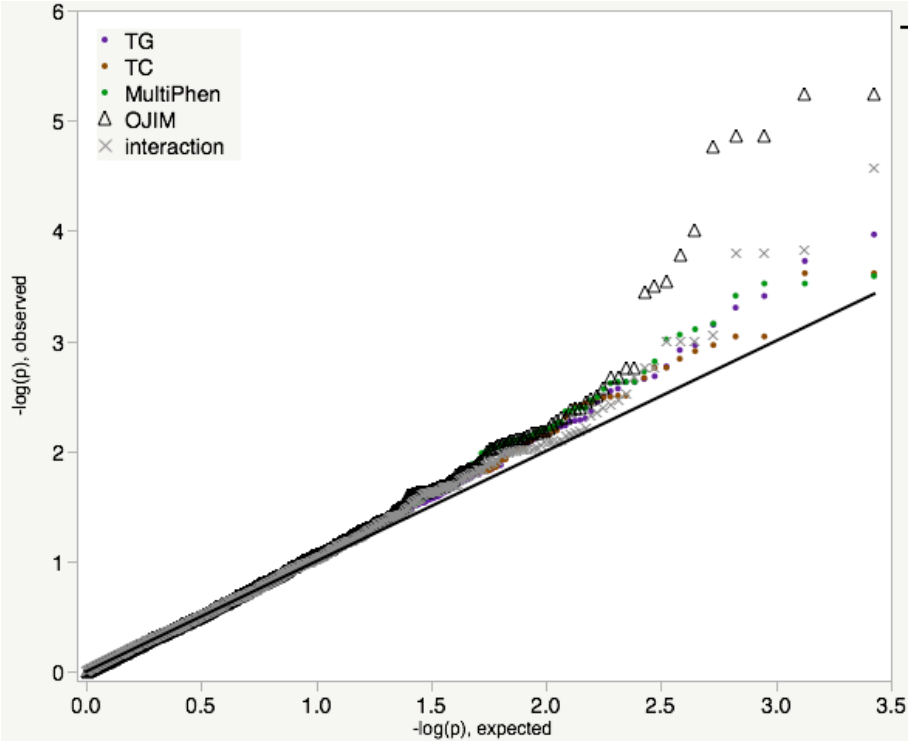


Figure 5-7. Correlation between LDL and TG by genotype at rs12740374; (GG=0, GT=1, TT=2). TG and LDL were standardized before stratification by genotype.

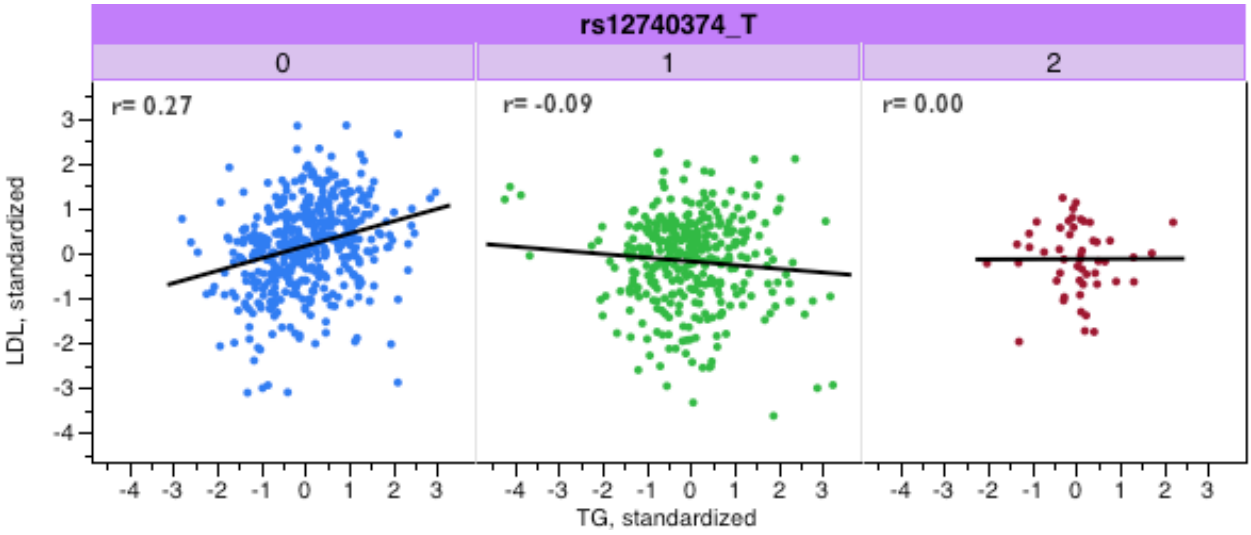


Figure 5-8. LDL measurements by rs12740374 genotype and mean LDL levels by rs12740374 genotype and triglycerides quartile; (GG=0, GT=1, TT=2). Error bars denote 95% confidence intervals. Note that triglycerides quartiles are based on the population distribution (not genotypic class). Points in top panel are randomly jittered.

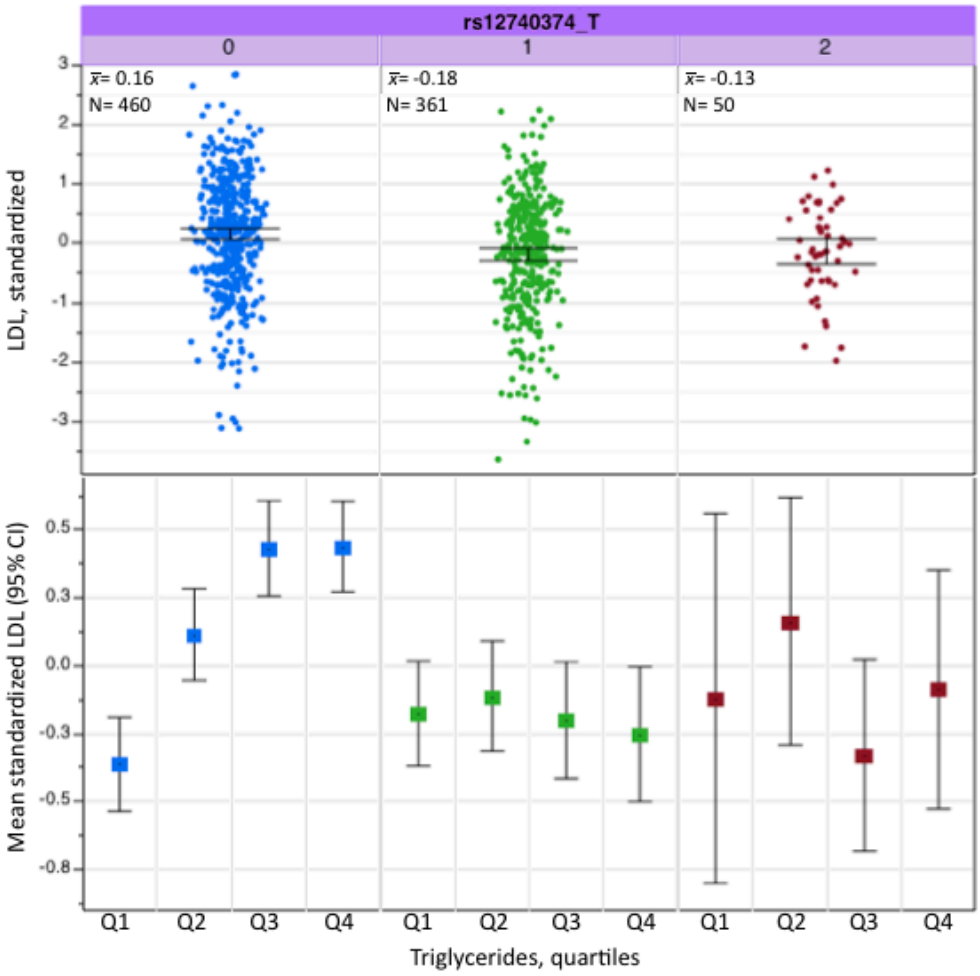


Figure 5-9. Correlation between HDL and TG by genotype at rs4938303; (CC=0, CT=1, TT=2). TG and HDL were standardized before stratification by genotype.

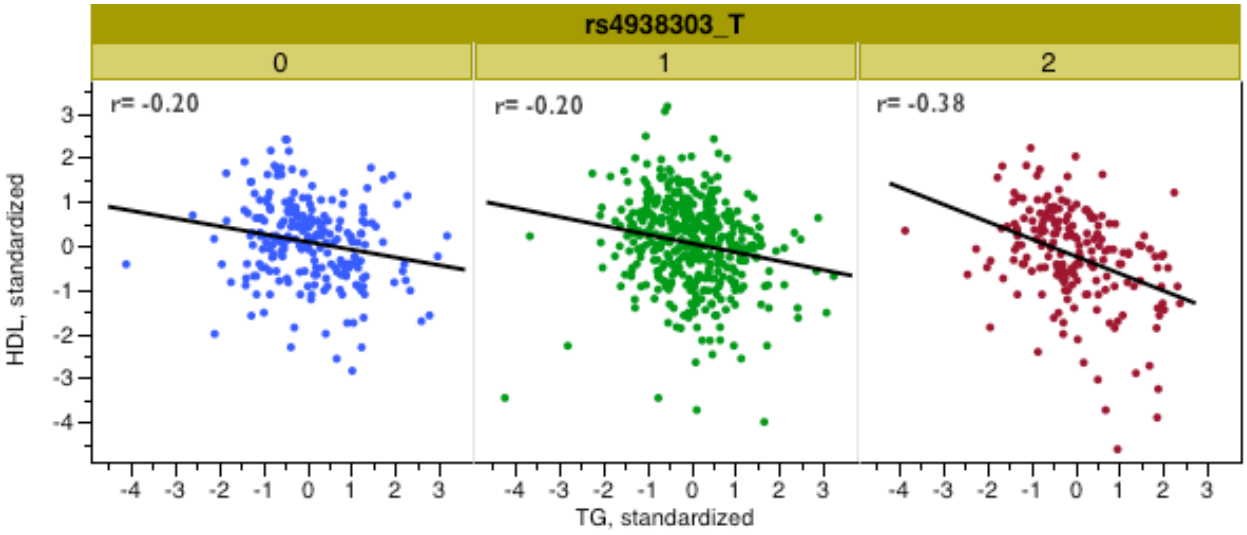


Figure 5-10. HDL measurements by rs4938303 genotype (top panel), and mean HDL levels by rs4938303 genotype and triglycerides quartile (bottom panel); (CC=0, CT=1, TT=2). Error bars denote 95% confidence intervals. Note that triglycerides quartiles are based on the population distribution (not genotypic class). Points in top panel are randomly jittered.

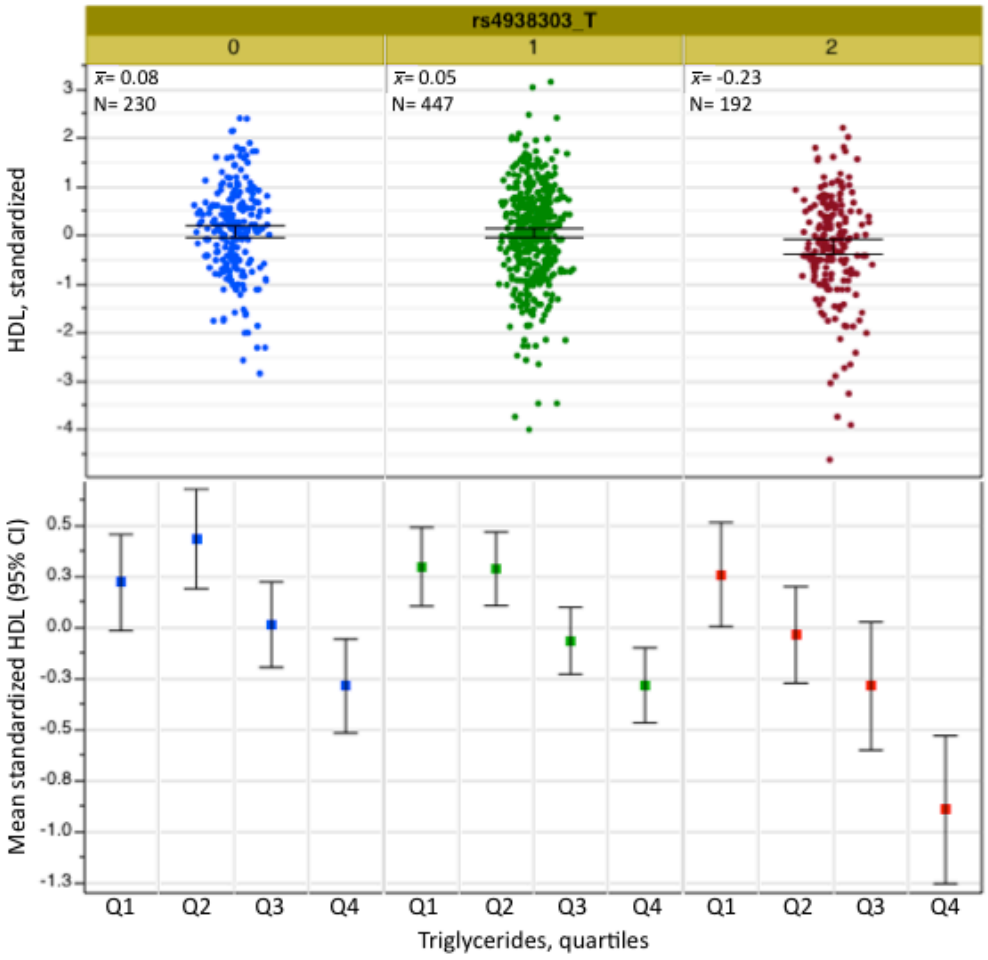
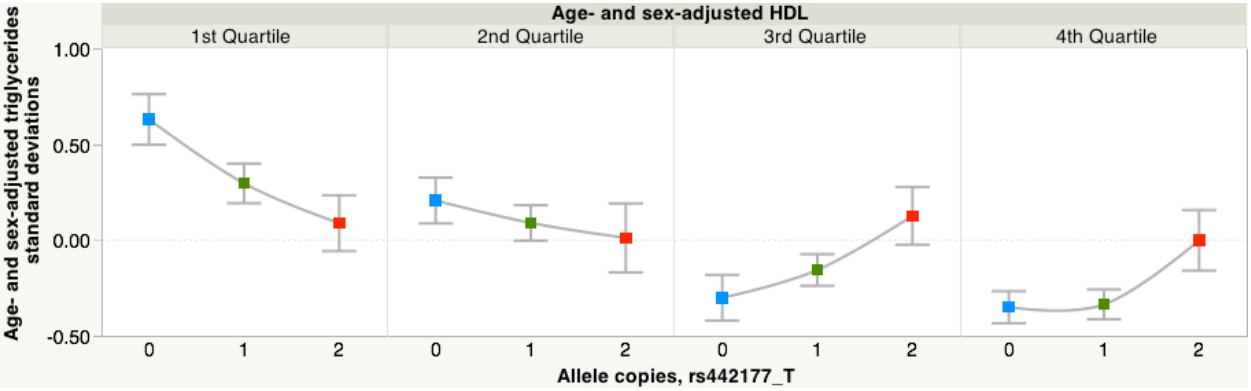


Figure 5-11. The effect of rs442177 on triglycerides changes direction when HDL increases beyond its median value. Points denote mean TG by HDL quartile for genotypes CC (blue), CT (green), and TT (red), +/- standard error; N=869.



Genetic Variants with Conditional Effects on PAI-1 and CVD Risk Factors

Introduction

Endophenotypes and missing heritability

Cardiovascular disease (CVD) is responsible for almost one-half of all non-communicable disease-related deaths worldwide.[277] It comprises multiple disorders of the circulatory system, among which venous and arterial thrombotic disorders are the most common. The enzyme plasminogen activator inhibitor-1 (PAI-1) plays a major role in the etiology of thrombosis by impeding fibrinolysis, or clot breakdown.[278] Elevated plasma PAI-1 is accordingly a major risk factor for thrombotic events, such as deep vein thrombosis, myocardial infarction, and stroke.

In genetic epidemiologic studies, plasma PAI-1 concentration has emerged as a promising endophenotype for CVD, because it provides a single heritable and quantitative measurement that is biochemically linked to heterogeneous clinical endpoints. By separating complex diseases into more precisely definable components with simpler genetic architectures, endophenotypes can improve the power of genetic association studies to find biologically and clinically meaningful variants. These advantages were recently demonstrated by a GWAS on serum-transferrin (a biomarker for iron deficiency), which identified two loci that accounted for 40% of the genetic variation.[279] Similar attempts to characterize the genetic architecture of PAI-1, however, have not been nearly as successful. A recent meta-analysis identified only three genome-wide significant loci, which together explained less than 3% of the genetic variance.[57]

A fundamental criterion for choosing endophenotypes for GWAS has been high heritability. For, if genes do not explain much of the endophenotypic variation, they will ultimately explain even less of the phenotypic variation. However, the heritability of PAI-1 has been estimated to be as high as 0.83,[56] making the inability to identify any major genetic factors (beyond the well documented 4G/5G variant in the PAI-1 gene) rather puzzling.[280] Moreover, the small number of variants that are associated with PAI-1 do not appear to be associated with CVD-related outcomes,[281] although PAI-1 itself is.

Thus, on the one hand, we see that PAI-1 satisfies all of the conventional criteria of an endophenotype: it is precisely measurable, heritable, and a biochemically integral component of an etiological pathway, and its genetic architecture (in terms of loci that influence its production directly) is undoubtedly simpler than that of the clinical endpoints with which it is associated. On the other hand, the variants discovered thus far that appear to affect PAI-1 levels, such as the 4G/5G promoter polymorphism, explain only a small fraction of its total heritability and have no clinically meaningful effect on CVD.

One explanation to this paradox is that the heritability of PAI-1 levels may mostly be due indirect genetic effects. In other words, it is possible that monozygotic twins have more highly correlated PAI-1 levels than dizygotic twins because the monozygotic pairs also share many of the heritable traits that increase PAI-1. We know that PAI-1 levels increase steadily with cardiometabolic risk factors such as BMI and triglycerides (Chapter 3), and that these risk factors are themselves highly heritable, as are most anthropometric traits.[282,283] Even dietary habits, across a wide range of categories, are consistently heritable at 0.3-0.5.[284,285] This being so, the thousands of loci that influence cardiovascular risk factors, anthropometric traits, and behavior may explain much of the heritability of PAI-1.

Put another way, even if the genes involved in the manufacture and release of PAI-1 were perfectly conserved and devoid of any variation, PAI-1 would still be a heritable trait, because conditions such as obesity, hypertriglyceridemia, hypertension, and nicotine addiction are heritable. It is worth mentioning that if these risk factor conditions were Mendelian traits, their causative loci could easily be detected by a GWAS of PAI-1 (despite having no direct influence on PAI-1). Yet, because their architecture is far from Mendelian, and, moreover, because any combination of such conditions can increase PAI-1 levels, their causative loci typically do not stand out among PAI-1 GWAS results. Vastly increasing sample size might change that, but the point of such an exercise would not be clear.

Genetic association studies of PAI-1 typically adjust for triglycerides (TG) and BMI, and justifiably so, because every variant that increases TG and BMI should, on average, also increase PAI-1, and the point of a GWAS on PAI-1 is not to find such indirect and artifactual associations. Yet, *direct* associations that are mediated by TG and BMI likely also exist. For example, we saw in Chapter 3 that median PAI-1 does not increase at all with BMI until standardized BMI exceeds -1σ. If, correspondingly, the pathways by which adipocytes accelerate PAI-1 production are (to some degree) under genetic control, then variants that disrupt or enhance those pathways must exist. Similarly, a polymorphism in the gene that encodes VLDL-inducible factor may increase its binding power to the PAI-1 promoter, such that PAI-1 production rises faster than usual as triglyceride levels rise. In association studies, the signals of these direct but dependent genetic effects will (1) be relatively weak, since they only pertain to a fraction of the total population, (2) will vary with the underlying distributions of TG and BMI of each study, and (3) will not be improved (and may be made worse) by adjusting for TG and BMI. Thus, another explanation for the missing heritability may be that, although many variants

do indeed influence PAI-1 directly, their effects are highly context-dependent.

We should expect to see small effect sizes, unpredictable replications, and wide ranges of heritability estimates whether the "missing heritability" of PAI-1 stems from indirect or context-dependent genetic effects (two possibilities that are not mutually exclusive). A further implication of both explanations is that the loci with the largest independent effects on PAI-1 have likely already been identified; only the loci with small independent effects and the loci with (potentially) strong, but context-dependent effects remain to be found. Because both types of variants can be expected to display weak marginal effects, finding them will require increased statistical power. The usual recommendations apply here: increased sample sizes and genomic coverage, better phenotypic measurements (PAI-1 assays being particularly variable)[286], and more studies on diverse populations (particularly genetically homogenous populations). However, as mentioned above, if direct and independent genetic effects are truly rare, increasing power will also generate scores of indirect associations that are technically legitimate but often uninterpretable, ungeneralizable, and unreplicable. Thus, the continued improvement in sequencing technology and the trend towards meta-analyses, far from a cure-all, will require that particular care be taken to define the phenotype precisely and to adjust for many confounding covariates, lest artifactual associations of small effect size (including those deriving from cryptic population structure[287]) crowd out meaningful results.

With regard to these issues, which beset GWAS in general, endophenotypes offer several particular advantages, such as their ability to be defined and quantified precisely, that are well known, but others which, in our view, have not yet been fully articulated or appreciated. First, because an endophenotype typically links risk factor "inputs" with multifactorial CVD "outputs," many of the possible confounding factors are the risk factors themselves. Moreover, rather than

just adjusting for these risk factors, they can be exploited by assessing them jointly with the endophenotype in a multivariate model. Although independent pleiotropic effects on both the endophenotype and the risk factors may seem unlikely, the strengths of multivariate analysis are not confined to detecting pleiotropic variants.[227] As discussed in the previous section, when variables are highly correlated (as endophenotypes and associated risk factors are, almost by definition), any gene with strong effects on only one of them will modify their correlational structure, and hence provide a stronger signal in a multivariate test. It is also straightforward to interpret which phenotype (or phenotypes) a SNP is associated with (e.g., using the R-based platform MultiPhen, one can simply compare the beta coefficients of each).[226] Thus, multivariate analyses can simultaneously adjust for likely confounders, distinguish true from indirect associations with the endophenotype, and increase the power to detect loci with small, independent effects on *any* of the phenotypes.

Knowledge of the risk factors upstream of an endophenotype can be even more beneficial for discovering context-dependent genetic effects. For, although such effects have frequently been found where sought (Chapter 4), identifying the germane "contexts" can be difficult. Genetic background, for example, though clearly a fundamental modifier of genetic effects, is difficult to define. However, the risk factors that precede an endophenotype on the etiological chain are natural candidates to test for SNP-by-phenotype interactions. The loci involved in such interactions may also be especially clinically relevant and biologically meaningful. In the case of PAI-1, for example, a hypothetical variant that causes its expression to spike abnormally when insulin secretion exceeds a certain threshold could be an important risk allele for ischemic events and provide insight into to the connection between diabetes and cardiovascular endpoints. Yet, such a variant could easily be missed by case-control studies, or buried among the false positives

and indirect associations, because any number of controls with lower insulin levels may carry it. It is therefore surprising that association studies of endophenotypes have not typically explored such interactions. This may partly be because using tests of the SNP-by-covariate interaction term alone to assess significance are highly underpowered, while joint interaction tests of single-outcome models are not straightforward to implement (see Chapter 4 and previous section of this chapter).

Here, using exome-wide data obtained from a cohort 1032 Ghanaians, we apply the ordinal joint interaction model (OJIM) to PAI-1 and the four cardiometabolic risk factors that are independently associated with it (as described in Chapter 3). We note that, with regard to the issues discussed above, the OJIM is particularly well suited to detect both context-dependent effects and small, independent effects, while distinguishing between likely indirect and direct associations, making it a powerful tool for the study of endophenotypes. Because it simultaneously assesses marginal genetic effects and conditional genetic effects, it does not require that the interaction term be highly significant for the context-dependent locus to be identified—only that an interaction *exist*. This is important, because the statistical significance of the interaction term is not a good proxy for the strength of a context dependent effect (Chapter 4). Moreover, the OJIM does not force us to choose *a priori* which of the phenotypes the SNP interacts with, and which it influences. This may be especially important for studies of PAI-1, because the correlations between PAI-1 and cardiometabolic risk factors cannot be assumed to be causal, or unidirectionally causal. There is evidence, for example, that PAI-1 is not only released by adipocytes, but can promote adipogenesis itself.[161] Similarly, although VLDL (which transports triglycerides in the blood) can induce PAI-1 expression, PAI-1 is also a ligand for the VLDL-receptor.[288]

In our lipid trait analysis of the previous section of this chapter, we did not expect to see many significant interactions in the absence of significant marginal effects, since the NHGRI SNPs were discovered on the basis of their marginal effects. By extending our analysis to the whole exome here, we hope to gain some insight into the significance and prevalence of such SNPs, regarding which there is a great deal of uncertainty.[208] The brief discussion of the nature and expected statistical properties of such SNPs follows below.

Heterogeneity of correlation by genotype as a complement to regression-based analyses

In theory, a "purely" context-dependent gene would have an effect on one quantitative trait that always and entirely depended on the value of another. One example would be a gene involved in coordinating the values of two quantitative traits in relation to each other, functioning in the manner of a molecular thermostat. We could imagine this gene participating in an endocrine feedback circuit, for example, guiding one trait to rise or fall in response to another. There have been surprisingly few studies exploring genetic variants that directly influence changes in covariance in this way. Two decades ago, Reilly et al. found that the correlation structure between various apolipoproteins varied with apoliporotein E (*ApoE)* genotype, and in a gender-specific manner.[289] This ability of *ApoE* to modulate lipid trait relationships was again demonstrated in a 2013 study, which concluded that the *ApoE* isoform genotype not only influenced the correlation between triglycerides and total cholesterol, but changed the relationship between both those traits and incident coronary heart disease as well, in a population-specific manner.[290] No high-throughput study of genes influencing the covariance among traits has been performed.

A somewhat larger number of studies have looked for genes that affect trait variance. A recent study, for example, reported that the FTO gene, known to associate with mean BMI, also increases BMI variance, in a way that is possibly mediated by DNA methylation.[291] The study did not find any other such variance genes for BMI despite a large sample size, but the authors were especially conservative in adjusting for multiple testing as well as in controlling for mean effects, potentially attenuating the signal of factors that increase both a trait's mean and its variance. In contrast, a recent study on *Arabidopsis* concluded that genetic variance heterogeneity appeared to be as common as normal additive effects on a genome-wide scale.[292] A few other noteworthy examples of variance genes have been reported in human studies. One found that polymorphisms in the Apo E gene affected total cholesterol variance to such an extent that the genotype group with the lowest mean total cholesterol actually had among the largest fraction of its members above a "high risk" threshold.[293] Variance in allele-specific expression has also been found to associate with colorectal cancer.[294]

An important concept in evolutionary biology is canalization, which refers to the robustness of a phenotype to developmental and environmental conditions.[295] It is hard to imagine that genetic factors do not play a major role in the phenotypic "buffering" that defines canalization. Indeed, there is growing evidence that the genetic basis of complex human disease centers on decanalization, its dissolution.[296] Phenotypic buffering can be achieved by managing the phenotypic variance of a single trait, but it can also be achieved by controlling the covariance of multiple traits, as demonstrated by a seminal study on yeast. The expression profiles of 276 single nucleotide deletion mutants were shown to induce expression changes in hundreds of genes–in effect compensating at the phenotypic level for the effect of the deletion.[297] If changing a single factor has a "propagation effect" across a network that generates a compensatory

response, biological insight will not be gained by looking for mean trait changes, but rather for the changes in connectivity that ensue. In that regard, a gene that modulates covariance may be the hub of a gene network, and variants of that gene may tighten or loosen the connections of both epistatic networks.[298] From another angle, we can think of variants of such a gene as influencing the stochastic noise around the regression line of two standardized quantitative traits. Interestingly, a recent study on *Arabidopsis* noted that stochastic noise is a heritable trait, and identified genes altering its variation, many with no mean effects.[299]

Although the interaction term of the OJIM is well suited to detect heterogeneity of correlation by genotype, as described in the previous section, it loses power with increasing dominance deviation, and completely misses instances of overdominance. Thus, we complemented the OJIM analysis here with a test for homogeneity of correlation by genotype (see Methods), which is *only* sensitive to the changes in covariance by genotype (and not, as the OJIM is, to changes in residual correlation that are caused by marginal effects).

Methods

Study Population

Please see Section A of Chapter 3.

Anthropometric measurements and biochemical analysis

Please see Section A of Chapter 3.

Genotyping

Please see previous section of this chapter.

Quality Control

Approximately 250,000 variants from 1105 participants were available prior to quality controls. We removed all SNPs with a genotyping call rate < 95%. Individuals for whom < 95% of variants were called were removed from analyses. Variants with a minor allele frequency < 20% were also removed, as were variants with a Hardy Weinberg p value < 0.001. Cryptic relatedness was assessed in the data, and one participant in each pair of related individuals (pi-hat >0.2) was randomly removed. Following quality control, 1032 participants and 15,890 variants remained for analyses. All quality control procedures were performed in PLINK (version 1.07)[231].

Statistical Analyses

All statistical models and analyses are as described in the previous section of this chapter, except for the addition of tests for homogeneity of correlation by genotype (below). In the multivariate analyses, PAI-1 was paired with each of four cardiometabolic risk factors, namely, BMI, triglycerides (TG), fasting glucose (GLUC), and mean arterial pressure (MAP), that were chosen based on the partial correlational analysis of Chapter 3. All models were adjusted for age and sex. Associations with p-values below the $1x10^{-4}$ level in any model were annotated using SNPinfo[300] and are presented in the results. Gene functions were ascertained using a literature search.

Test for homogeneity of correlation by genotype

The high-throughput screen for covariance-modifying genes that is proposed here is statistically straightforward. Individuals are grouped by genotype, the correlations between two traits are calculated for each group, and a test of homogeneity of correlation among the three groups (0,1,2) is applied to assess whether the three sample correlation coefficients could have been drawn from the same population.

The variance of ρ, the population parameter of correlation between two traits, decreases as its absolute value approaches 1. Fisher's *r*-to-*z* transformation $z = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$ stabilizes the variance at $\sigma_z^2 = \frac{1}{n-3}$, and makes the distribution approximately normal, enabling conventional statistical tests. If we estimate correlation, r_i, for $k=3$ genotypic groups, and transform each to z_i, the weighted sum of squares is then distributed approximately as χ^2 with *k-1*, or 2 degrees of freedom:

$$\chi^2 = \sum_{i=1}^{k} \left(n_i - 3\right)\left(z_i - \mu_z\right)$$

The R-code for this approach allows for the adjustment for any number of covariates, making it essentially a test for partial correlation as well. The code also allows Spearman's rank sum correlations to be used when deviation form normality is an issue. However, the variance term for the Z-transforms has to be adjusted by a factor of 1.06 (Appendix D).[301] Although not implemented in this study, the code also allows for tests of dominant and recessive effects on correlation.

Results and Discussion

Endophenotypes such as PAI-1 have been considered promising targets for GWAS, because they exhibit less phenotypic heterogeneity and higher heritability than the complex disease-related endpoints with which they are associated. The theoretical rationale for their utility has been that (1) the genotype-endophenotype map should be substantially simpler than the genotype-phenotype map, allowing for the efficient detection of variants of relatively large effect size, and that (2) such variants should be especially likely to provide insight into complex disease. In the case of PAI-1, however, very few associations have been found, and those that have been found have not been clinically relevant. In our analyses, we demonstrate a novel way to study endophenotypes such as PAI-1 that do not fit the above model.

Our guiding premise is that the intensity of association between PAI-1 and cardiovascular risk factors must, to some extent, be under genetic control. Since the nature of that control can be considered a phenotype in itself, it is likely characterized by heritable variation. We propose that the loci responsible for this variation are more likely to be biologically meaningful than those that influence PAI-1 independently. For, given the sensitivity of plasma PAI-1 concentration to many cardiovascular risk factors, a variant that raises it independently of those risk factors would need to have a very strong effect indeed before it had any bearing on CVD-related endpoints. On the other hand, because PAI-1 (in its capacity as an endophenotype) links CVD risk factors to CVD endpoints, variants that modulate how PAI-1 concentration responds to those risk factors may be of particular clinical and biological interest.

We have already discussed the theoretical strengths of the ordinal joint interaction method (OJIM) to discover context-dependent variants, and used it to demonstrate that such variants are likely abundant. In our preliminary study on lipid traits, the OJIM displayed the most

power to detect SNPs with well-known functional roles in lipid metabolism, and its top results were consistently lipid-associated. Yet, none of its associations achieved genome-wide significance at the conventional threshold of 5×10^{-8}. We may attribute that failure to the rather small sample sizes as well as to the unreasonable strictness of the conventional threshold. Additionally, the SNPs of the Exome chip were chosen based on European genetic data, and consequently may not tag functional loci adequately in African populations. The signal of true associations may be further attenuated by the weaker linkage disequilibrium in African populations in general.[302] Regardless of the reason, because the top lipid associations (which were almost certainly true associations) had p-values between 10^{-3} and 10^{-7} in our preliminary analyses (2,669 SNPs), we chose $p=10^{-4}$ as our nominal threshold for significance in the exome-wide tests (15,890 SNPs).

The most significant regression-based association in this study was the OJIM's top result for the TG-PAI-1 tests, rs29234, a SNP within the myelin oligodendrocyte glycoprotein gene (*MOG*) ($p=1.06 \times 10^{-5}$) (**Table 5-4**). The univariate association with TG was nearly as significant ($p=1.16 \times 10^{-5}$, unadjusted for the two univariate tests), but here the additional insight afforded by the OJIM, even in cases when it provides only a marginal improvement in power, was well illustrated. The rs29234 association with PAI-1 (p=0.08) was not significant; therefore, running only univariate or bivariate tests (without an interaction) would have connected *MOG* only to TG. Yet, even a nominally significant interaction term (p=0.02) can provide qualitative information indicating that a connection between PAI-1 and *MOG* exists. And, whereas TG has no connection with *MOG* in the literature, PAI-1 has been shown to interact with it *in vivo*. Specifically, autoimmune encephalomyelitis (EAE) was experimentally induced by *MOG* in urokinase PA (u-PA) knockout and knockdown mice models, and rescued by PAI-1.[303] We

cannot know, of course, whether even the marginal association between *MOG* and TG ($p=1.16x10^{-5}$) is real. But, the fact that the interaction with PAI-1 was significant enough to give the OJIM the most power overall to detect rs29234, combined with prior evidence for a biological connection between PAI and *MOG*, suggests that if the association is true, PAI-1 may indeed be involved in the relationship between TG and *MOG*, and in a way that would have been missed by univariate tests.

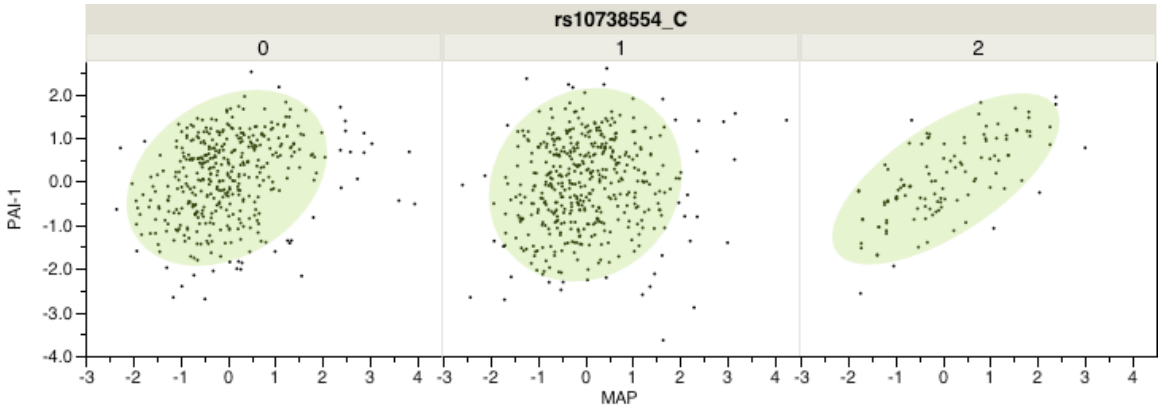
In the MAP-PAI tests, the top p-value (7.60×10^{-6} for rs10738554) was again yielded by a multivariate approach (the test for homogeneity of correlation by genotype). The highly significant p-value is compelling given (1) the high MAF (0.34) at the locus, which implies relatively stable estimates of correlation for each genotype, and (2) the fact that Spearman's rank correlation was used for all tests, making it highly conservative. We chose Spearman's rho to keep Type 1 error to the low levels observed for the OJIM's interaction term, with which we aimed to compare its performance. In addition to the strong p-value, rs10738554 is located near *SLC24A2* (also known as *NCKX2*), a gene previously associated with high blood pressure in African Americans.[304] *SLC24A2* belongs to a family of proteins that transport sodium, potassium and calcium ions to regulate homeostasis, and can thus be plausibly implicated in the improper regulation of blood solutes that characterizes hypertension. Additionally, in the context of the renin-angiotensin system, high blood pressure also promotes overexpression of PAI-1 levels.[305]

While the biological relationship between PAI-1 and *SLC24A2* has not been previously explored, a recent study found that the disruption of *SLC24A2* (*NCKX2*) renders neurons more susceptible to ischemic insult. In particular, primary cortical neurons in *SLC24A2* knockout models displayed a higher vulnerability and greater tendency to release Ca^{2+} ions under hypoxic conditions.[306] Because hypoxia also stimulates PAI-1 expression,[307] it is possible that in our

study, PAI-1-level is serving as a proxy for hypoxic conditions (or some other correlate thereof), such that its increase corresponds with abnormal ion exchange in individuals with poorly functioning *SLC24A2*. If so, this context-dependence would explain why rs10738554 had only a weak marginal effect on MAP in this study (p=0.24). This interpretation is also consistent with the apparent recessive effect of rs10738554-C on the MAP-PAI-1 correlation (**Figure 5-12**).

It is worth noting that the OJIM performed poorly with rs10738554; its interaction p-value was 0.87 despite significant heterogeneity of correlation by genotype. However, as mentioned above, the OJIM interaction term detects only additive effects on correlation, whereas here, the correlation for the heterozygote genotype was lower than that for both homozygotes (**Figure 5-12**). Perhaps the deviation from additivity we observed was merely due to sampling error, and with larger sample size, the correlation among the major allele homozygotes would have equaled or exceeded that of the heterozygotes. But, because the observed genetic effect appeared to be overdominant, the OJIM had no power to detect it. Consistent with true overdominance is the fact that the minor allele frequency (MAF) of rs10738554 is close to 50% in all HapMap populations (in fact, its MAF of 36% in Yorubans is the lowest among continental populations), suggesting the possibility of balancing selection. Regardless of the true dominance deviation, we see here that complementing the OJIM with tests for homogeneity of correlation when sample sizes are not particularly large can be valuable. With the largest sample sizes, however, the OJIM will generally outperform the test for homogeneity of correlation in all cases except true overdominance.

Figure 5-12. Correlation between PAI-1 and MAP by genotype at rs10738554; (TT=0, CT=1, CC=2). The Spearman's correlations for genotypes TT, CT, and CC are 0.33, 0.13, and 0.57, respectively. PAI-1 and MAP were adjusted for age and sex and standardized before stratification by genotype.



Of the seven p-values less than 10^{-4} in the BMI-PAI-1 tests, five had no significant marginal effects on either BMI or PAI-1 (**Table 5-6**). Two of them, rs199818197 ($p=3.37 \times 10^{-05}$) and rs2233391 ($p=7.87 \times 10^{-05}$), are in genes that have been directly implicated in lysosomal storage disorders (LSD) (*SUMF1* and *NEU2,* respectively). Lysosomal dysfunction is biologically associated with BMI, because it is known to affect energy balance and interfere with normal adipose storage.[308] Mutations in *SUMF1* cause the severe LSD multiple sulfatase deficiency[309], while *NEU2* belongs to a family of mammalian sialidases that are involved in the LSD sialidosis, as well as other conditions such as diabetes and arteriosclerosis.[310,311 312] Interestingly, a recent study found that NEU1, closely related to NEU2, was much more active in the epididymal fat of obese and diabetic mice, and concluded that fluctuations in NEU1 activity might be associated with the pathological states of excessive visceral fat.[312] Importantly, PAI-1 is secreted by adipocytes and has a terminal sialic acid residue, the sialylation status of which should affect how much of it is released or activated.[312,313] _ENREF_312 Thus, a relationship wherein *SUMF1* and/or *NEU2* modulate the covariance of BMI and PAI-1 directly, i.e. such that low adiposity leads to reduced PAI-1 expression and high adiposity leads increased PAI-1 expression, is both biologically plausible and consistent with the lack of main effects observed in our analyses.

The most significant joint interaction p-value for the BMI-PAI-1 tests was for rs1420101 ($p=8.66 \times 10^{-5}$), an intronic variant of *IL1RL1*, a gene previously associated with inflammatory responses (**Table 5-6**). IL1RL1 is selectively expressed on Th2 cells and mast cells, and binding of its ligand, IL33, produces an IL4 mediated response in allergic airway inflammation of extrinsic asthma.[314] IL4 stimulates isotype switching to IgE production, which in turn leads to mast cell degranulation and the release of histamine and other

mediators. Resultant bronchoconstriction, mucous production and pronounced leukocyte response in the airway leads to symptoms of expiratory wheezing and cough.[315] Importantly, both increased BMI and elevated PAI-1 levels have been associated with asthma severity[316]. PAI-1 is believed to play a role in the cell adhesion, chemotactic signaling for leukocytes and tissue remodeling.[316] The BMI-dependent increased risk of asthma has been reported for patients in the overweight and obese categories (i.e. BMI>30 and >25, respectively), and the level of PAI-1 present in sputum of asthmatics was an order of magnitude larger than that observed in healthy controls. In light of these clinical findings, the interaction p-value that links rs1420101 to both BMI and PAI-1 is biologically plausible, even though it was barely significant (p=0.044). Yet, as discussed above, it would have been missed by univariate analyses alone. It is also worth noting that, although the univariate association with BMI was nominally better ($p=7.71 \times 10^{-5}$) than the joint interaction p-value ($p=8.66 \times 10^{-5}$), it was not adjusted for the extra univariate test with PAI-1.

Variant rs404890 upstream of *NOTCH4* on chromosome 6 was significant using the heterogeneity of correlation model ($p=6.88 \times 10^{-5}$) in the analysis of the glucose-PAI-1 pairing. Murine knockouts of *NOTCH4* display severe angiogenic vascular remodeling defects, consistent with the well-known functional role of Notch4 in in promoting arterial endothelial cell specification.[317] PAI-1 is also known to promote angiogenesis, although the exact mechanism has not been well described.[318] Plasma glucose has been shown to have an inverse relationship with vascular endothelial growth factor (VEGF) expression.[319] Furthermore, severe, chronic hyperglycemia as observed in cases of poorly controlled type 2 diabetes damages vessels by non-enzymatic glycosylation, thereby increasing vessel permeability, atherogenesis and hyaline

arteriolosclerosis.[315] Proliferative diabetic retinopathy is another endpoint of poorly managed diabetes, the hallmark of which is aberrant angiogenesis leading to abnormal, fragile vessels in the eye. In the early, non-proliferative phase of diabetic retinopathy, microaneurysms in retinal vessels occur, eventually leading to blockage, hemorrhages and, in some individuals, the proliferative, angiogenic stage of the disease. While threshold specific effects of Notch4 and PAI-1 are not well established, making it difficult to speculate on their physiological effects with respect to angiogenesis, the role of glucose is well known. The effects of clinical hyperglycemia, both through direct action on the vessels and indirect modulation of VEGF, fit the context-dependent model, where they become active and pathogenic beyond a certain level of vessel injury.

Overall, the multivariate methods that assessed or allowed for modifications to correlation (i.e., the OJIM, the OJIM interaction, and the test for heterogeneity of correlation by genotype) yielded by far the most results significant at the 10^{-4} level. Type I error for these multivariate tests was incredibly low, as illustrated by the QQ-plot in **Figure 5-13** for the MAP-PAI-1 tests, below. In fact, those analyses (as is also clear from **Table 5-5**) only the homogeneity of correlation tests appeared to have sufficient power. The strong performance of the homogeneity of correlation approach in general was one of the surprises of this study. For, we expected it only to complement the interaction term of the OJIM, providing more power in exceptional cases of strong dominance deviation. Perhaps such deviations from additivity are more common than generally appreciated. Alternately, the OJIM's interaction term may require larger sample sizes to generate more consistent results (we note that the simulations in the previous section of this chapter were performed with N=5000).

Surprisingly, MultiPhen performed especially poorly. In fact, there was only one result in the entire study for which it provided a (minimal) advantage over the OJIM, namely rs1048347 in the TG-PAI-1 analyses (p=2.03 x 10-05 vs. 3.54 x 10-05), a locus with no connection to either phenotype in the literature. Linear univariate tests did not fare much better; tests of glucose generated no p-values significant at the 10^{-4} level, and tests of BMI and PAI-1 generated one for each. Those were *IL1RL1* for BMI (discussed above) and *MGAM* for PAI-1. *MGAM* has not previously been linked to PAI-1 in the literature, but its role in starch digestion may be relevant n.[320]

Conclusion

While it is impossible to generalize from one study, in these analyses we noted essentially no disadvantages or tradeoffs to assessing traits in pairs and adding an interaction. There were, however, substantial advantages. Moreover, we performed a genome-wide scan for genetic effects on correlation by genotype, which to our knowledge had never been done before, and the results were encouraging. Allowing for interactions or seeking them directly generated the majority of significant associations in this study, with no evidence of inflation for Type 1 error, further supporting our conclusions drawn from the lipid trait analyses. Context-dependent effects very well may be ubiquitous, and our methods are singularly suited to detect them. This is promising for future studies of CVD, because risk factors are generally not well understood from an etiological perspective, and their genetic architecture even less so.

Figure 5-13. QQ-plots of p-values for tests assessing 15,890 exonic SNPs for association with mean arterial pressure and PAI-1 in 1032 Ghanaian participants. Results of tests using the ordinal joint interaction model (OJIM) (black triangle), its interaction term alone (grey cross), and tests for homogeneity of (Spearman's) correlation by genotype (purple) are depicted.

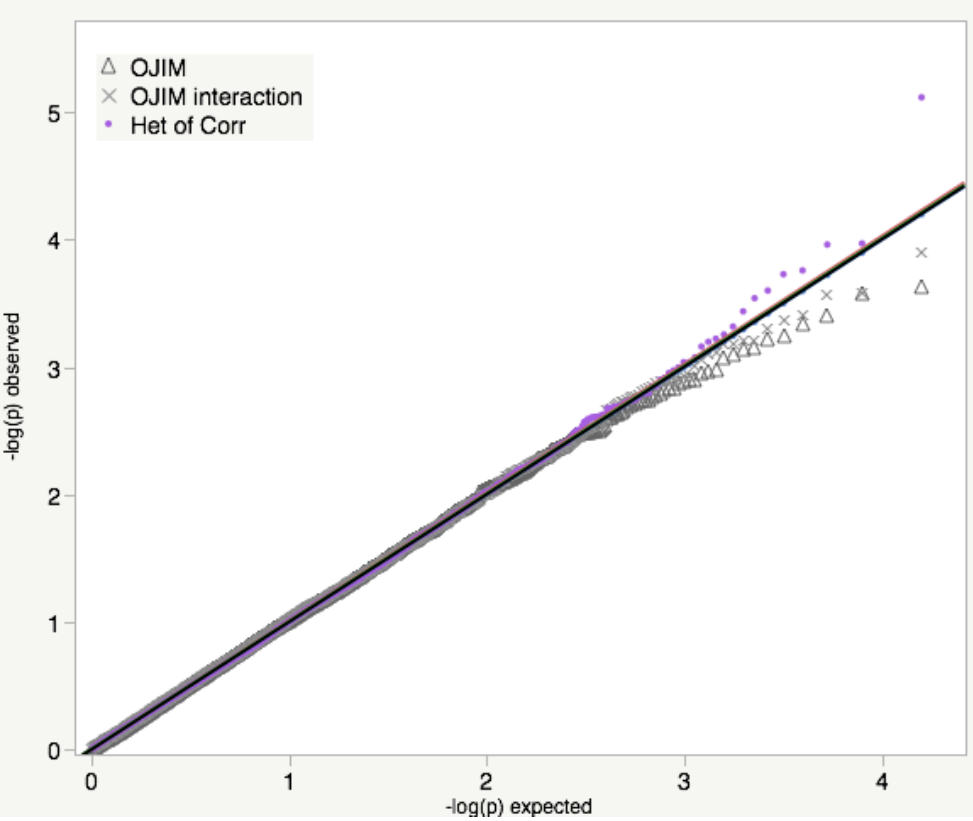


Table 5-4. Associations ($p<10^{-4}$) with triglycerides and PAI-1 in 1032 Ghanaian participants; 15,890 exonic SNPs (MAF≥0.20) were tested for association using univariate and multivariate methods; all models were adjusted for age and sex.

				p value						
SNP	Chr.	Minor Allele	MAF	Single Trait TG.	Single Trait PAI-1	Bivariate	OJIM	Interaction term	Homogeneity of Correlation	Gene
rs29234	6	G	0.20	**1.16E-05**	0.078	**3.47E-05**	**1.06E-05**	0.022	0.343	*MOG**
rs29272	6	A	0.20	**1.35E-05**	0.090	**3.99E-05**	**1.17E-05**	0.021	0.323	*MOG**
rs1048347	10	C	0.33	1.98E-04	0.371	**2.03E-05**	**3.54E-05**	0.196	0.056	*BTBD16*
rs9997165	4	G	0.33	0.628	0.753	0.682	0.073	0.013	**8.95E-05**	*Loc100131135*
rs896999	15	A	0.27	**1.59E-05**	0.020	1.66E-04	3.68E-04	0.324	0.682	*ANP32A**
rs13165786	5	T	0.43	**5.97E-05**	0.653	1.31E-04	4.59E-04	0.867	0.919	*EDIL3**
rs3131875	6	G	0.29	**9.81E-05**	0.281	3.77E-04	0.001	0.880	0.627	*ZFP57*
rs10266732	7	T	0.37	0.105	**5.40E-05**	2.63E-04	4.74E-04	0.245	0.400	*MGAM**

*previously associated with cardiovascular disease

Column abbreviations: Chr.=chromosome; MAF=minor allele frequency; TG=triglycerides; Bivariate= MultiPhen, which models genotype as a function of TG and PAI-1; OJIM adds an interaction term.

Table 5-5. Associations ($p<10^{-4}$) with mean arterial pressure and PAI-1 in 1032 Ghanaian participants; 15,890 exonic SNPs (MAF≥0.20) were tested for association using univariate and multivariate methods; all models were adjusted for age and sex.

				p value						
SNP	Chr.	Minor Allele	MAF	Single Trait MAP	Single Trait PAI-1	Bivariate	OJIM	Interaction term	Homogeneity of Correlation	Gene
rs10738554	9	C	0.34	0.246	0.217	0.116	0.225	0.819	**7.60E-06**	*SLC24A2**
rs3736582	10	G	0.38	**6.04E-05**	0.481	3.78E-04	0.001	0.911	0.966	*PSTK*
rs16907312	11	T	0.31	**8.54E-05**	0.075	4.11E-04	0.001	0.446	0.857	*OR51G2*
rs10266732	7	T	0.37	0.472	**5.40E-05**	2.60E-04	3.93E-04	0.189	0.518	*MGAM**

*previously associated with cardiovascular disease or hypertension

Column abbreviations: Chr.=chromosome; MAF=minor allele frequency; MAP=mean arterial pressure; Bivariate= MultiPhen, which models genotype as a function of MAP and PAI-1; OJIM adds an interaction term.

Table 5-6. Associations ($p<10^{-4}$) with body mass index and PAI-1 in 1032 Ghanaian participants; 15,890 exonic SNPs (MAF≥0.20) were tested for association using univariate and multivariate methods; all models were adjusted for age and sex.

				p value						
SNP	Chr.	Minor Allele	MAF	Single Trait BMI	Single Trait PAI-1	Bivariate	OJIM	Interaction term	Homogeneity of Correlation	Gene
rs1420101	2	A	0.32	**7.71E-05**	0.004	1.69E-04	**8.66E-05**	0.044	0.416	*IL1RL1**
rs28550932[△]	9	A	0.29	0.932	0.479	0.858	1.17E-04	**6.05E-06**	**9.21E-05**	*Loc286238*
rs7835830	8	T	0.34	0.319	0.716	0.569	1.82E-04	**1.51E-05**	4.62E-04	*FAM135B*
rs199818197	3	T	0.33	0.879	0.345	0.643	4.24E-04	**3.37E-05**	0.035	*SUMF1**
rs2233391	2	A	0.18	0.462	0.790	0.544	7.74E-04	**7.87E-05**	7.45E-04	*NEU2*
rs9880989	3	G	0.45	0.076	0.058	0.102	0.003	0.002	**2.51E-05**	*IQCG*
rs10266732	7	T	0.37	0.048	**5.40E-05**	2.61E-04	5.57E-04	0.316	0.565	*MGAM**

*previously associated with cardiovascular disease or obesity

[△]rs28429833, not listed, was in almost perfect linkage disequilibrium with rs28550932

Column abbreviations: Chr.=chromosome; MAF=minor allele frequency; BMI=body mass index; Bivariate= MultiPhen, which models genotype as a function of BMI and PAI-1; OJIM adds an interaction term.

Table 5-7. Associations ($p<10^{-4}$) with glucose and PAI-1 in 1032 Ghanaian participants; 15,890 exonic SNPs (MAF≥0.20) were tested for association using univariate and multivariate methods; all models were adjusted for age and sex.

				p value						
SNP	Chr.	Minor Allele	MAF	Single Trait Glucose	Single Trait PAI-1	Bivariate	OJIM	Interaction term	Homogeneity of Correlation	Gene
rs1649292	2	A	0.29	0.443	0.608	0.717	0.720	0.413	**2.04E-05**	*Loc129293*
rs63111160	18	T	0.49	0.409	0.039	0.040	0.055	0.289	**3.59E-05**	*SETBP1**
rs404890	6	T	0.29	0.615	0.386	0.465	0.003	4.54E-04	**6.88E-05**	*NOTCH4**
rs10266732	7	T	0.37	0.122	**5.40E-05**	2.31E-04	7.03E-04	0.608	0.307	*MGAM**

*previously associated with cardiovascular disease, type 1 or type 2 diabetes mellitus

Column abbreviations: Chr.=chromosome; MAF=minor allele frequency; Bivariate= MultiPhen, which models genotype as a function of glucose and PAI-1; OJIM adds an interaction term.

CHAPTER VI

CONCLUSIONS AND FUTURE DIRECTIONS

It is well known that cardiovascular disease is caused by risk factors that tend to co-occur, such as obesity, high blood pressure, dyslipidemia, and diabetes, and that these risk factors can be prevented or controlled by behavioral and dietary changes to a large extent. Yet, there is also a great deal of variation (among individuals and populations) in how these risk factors respond to lifestyle modifications, how they associate with each other, and how they contribute to clinical endpoints. Although we know that some of these differences are rooted in genetics, attempts to identify and characterize the genetic variants responsible for them have not been successful. The central premise of this work is that much insight into the genetics of cardiovascular disease can be gained by shifting focus away from genes that may influence risk factors and endpoints in isolation, to genes that may modify how they relate and interact with each other at different points in the etiological sequence. In our opinion, finding a genetic variant that increases the risk of myocardial infarction when cholesterol levels rise (to give an example) can offer more clinical insight and better guidance for future studies than finding a variant that is associated with either high cholesterol or myocardial infarction alone.

This is the first study to address in a systematic way the genetics of cardiovascular risk factor correlations. We did not consider endpoints of cardiovascular disease in this study, but rather an endophenotype of cardiovascular disease, PAI-1. Because PAI-1 plays a direct role in thrombosis and subsequent ischemic events, identifying genetic variants that strengthen its correlation with other cardiovascular risk factors can improve assessment of risk, provide

etiological insight, and suggest targets for intervention. The strategy we have used here can be adapted to answer many different kinds of questions in genetic epidemiology.

The phenotypes of interest in this study were not only cardiovascular risk factors, but the correlations among them. Cardiovascular risk factors such as systolic blood pressure and visceral adiposity are known to be, to some extent, under genetic control; if, as we propose here, the relationship between them is also under genetic control, then that relationship can be considered a phenotype in itself, which, like virtually all quantifiable phenotypes, is likely characterized by heritable variation. Accordingly, our first goal in this study was to understand these phenotypes in our Ghanaian study population as completely as possible. We aimed not only characterize correlational networks of cardiovascular risk factors, but (before introducing a whole new layer of genetic complexity into the picture) also to identify how non-genetic factors, particularly urban lifestyles, perturbed these networks.

Although we found that urban residence and, to a lesser extent, sex had dramatic effects on the mean values of cardiovascular risk factors (Chapter 3A), our partial correlational analyses revealed that the *relationships* among the risk factors remained remarkably robust (Chapter 3B). To our knowledge, this had not been shown before. The relationships between risk factors and PAI-1, however, were far more sensitive to differences in sex and environment. We found that triglycerides and BMI had the strongest independent relationships with PAI-1, followed by glucose and mean arterial pressure (MAP). Although the relationships with glucose and MAP were substantially weaker overall, we noted that over certain parts of their range, their relationship with PAI-1 intensified. We hypothesized that these non-linear and non-continuous relationships might be under genetic control and have particular etiological (and potentially clinical) significance. We therefore paired PAI-1 with these four traits in our subsequent genetic

analyses. We believe that our findings in Chapter 3 represent a major step forward in studies of PAI-1 and its role in cardiovascular disease (and the metabolic syndrome in particular). We explored the relationship from multiple angles, and used novel methods to address open questions, such as whether co-occurring risk factors have an effect on ischemic risk greater than the sum of their individual contributions. It will be interesting to see if the patterns we have identified here are generalizable to other populations.

Only a few studies, none of them recent, have reported genetic variants that modify correlations between traits, and none has sought to find them on a genome-wide basis. Because the genetics of correlations is, to a great extent, uncharted territory, we first explored the matter from a theoretical perspective in Chapter 4. The insight that a change in covariance between two traits by genotype is mathematically equivalent to a genotype-by-covariate interaction effect on an outcome allowed us to focus our attention provisionally on the more tractable class of linear regression equations. We modeled multiple types of biological SNP-by-covariate interactions and derived the statistical parameters to which they should give rise. In doing so, we demonstrated why it is a major error to assume that the significance of a statistical interaction term can capture the significance of a biological interaction. We demonstrated that even the strongest gene-by-covariate interactions at the biological level could have weak interaction effects when general linear models are used. Moreover, we quantified how strong we can expect the interaction effect to be relative to the marginal effect, depending on the nature of the biological interaction.

The analyses of Chapter 4 laid the groundwork for the development of the ordinal joint interaction model (OJIM), which can identify both marginal effects and SNP-by-covariate interactions where they exist (i.e. changes in correlation by genotype), while leveraging the

change in residual correlation induced by marginal effects (i.e. changes to the total population correlation) into increased power. The OJIM had more power than univariate or bivariate analysis to detect lipid SNPs of known biological significance, indicating that context-dependent genetic effects are probably quite common, and that the OJIM can identify them where they exist.

Although one of the strengths of the OJIM is that no a priori decision needs to be made regarding which is the interacting variable and which is the outcome, it can be used even when one of the covariates is (e.g.) an environmental exposure, and still outperform traditional tests of interaction, with minimal inflation for Type 1 error (as demonstrated in Chapter 5). We therefore recommend that future studies assessing gene-by-covariate interactions of any kind consider using the OJIM over the gene-by-covariate interaction term of the conventional single-outcome model.

APPENDIX

Appendix A: Supplemental Figures and Tables, Chapter III

Table S1. Age-standardized prevalence rates and 95% confidence intervals of dichotomous risk factors in the Ghanaian cohort.

	Males			Females			Urban	Rural
	Urban	Rural	p-value	Urban	Rural	p-value	p-value by sex	p-vaalue by sex
N	972	469		1293	583			
Hypertension	0.34 (0.31, 0.37)	0.20 (0.16, 0.24)	<.001	0.32 (0.30, 0.35)	0.21 (0.18, 0.24)	<.001	0.316	0.690
IFG (>100 mg/dL)	0.23 (0.20, 0.26)	0.19 (0.16, 0.23)	0.084	0.29 (0.27, 0.32)	0.31 (0.28, 0.35)	0.380	0.001	<.001
IFG (>110 mg/dL)	0.10 (0.09, 0.12)	0.05 (0.03, 0.07)	0.001	0.12 (0.10, 0.14)	0.10 (0.08, 0.13)	0.207	0.133	0.003
Diabetes	0.06 (0.04, 0.07)	0.02 (0.01, 0.03)	<.001	0.07 (0.05, 0.08)	0.03 (0.02, 0.05)	<.001	0.342	0.307
Overweight or Obese	0.35 (0.32, 0.38)	0.11 (0.08, 0.14)	<.001	0.60 (0.58, 0.63)	0.26 (0.22, 0.29)	<.001	<.001	<.001
Obese	0.07 (0.05, 0.09)	0.00 (0.00, 0.01)	<.001	0.26 (0.24, 0.28)	0.05 (0.04, 0.07)	<.001	<.001	<.001
Hypercholesterolemia	0.22 (0.20, 0.25)	0.07 (0.05, 0.09)	<.001	0.31 (0.28, 0.33)	0.10 (0.08, 0.13)	<.001	<.001	0.086
High TG	0.28 (0.25, 0.31)	0.27 (0.23, 0.31)	0.691	0.21 (0.19, 0.24)	0.26 (0.23, 0.30)	0.017	0.001	0.715
Low HDL-C	0.40 (0.36, 0.44)*	0.38 (0.32, 0.45)^	0.592	0.26 (0.23, 0.29)~	0.38 (0.32, 0.43)+	<.001	<.001	0.946
High LDL-C	0.22 (0.19, 0.25)*	0.05 (0.03, 0.08)^	<.001	0.30 (0.27, 0.33)~	0.11 (0.08, 0.15)+	<.001	<.001	0.014
Smoker	0.03 (0.02, 0.04))	0.16 (0.13, 0.20)	<.001	0.00 (0.00, 0.00)	0.02 (0.01, 0.03)	<.001	<.001	<.001
Any schooling	0.96 (0.95, 0.98)	0.64 (0.60, 0.68)	<.001	0.88 (0.86, 0.90)	0.44 (0.40, 0.48)	<.001	<.001	<.001
Schooling >JSS	0.48 (0.45, 0.52)	0.05 (0.04, 0.08)	<.001	0.30 (0.27, 0.33)	0.02 (0.01, 0.03)	<.001	<.001	0.007

*n=722 ^n=225 ~n=955 +n=317.

Hypertension = SBP ≥140 or DBP ≥90 or self-reported diagnosis with current use of medication; IFG = impaired fasting glucose; Diabetes =glucose ≥126 mg/dL or self-reported diagnosis with current use of medication; Overweight = BMI ≥25; Obese = BMI ≥30; Hypercholesterolemia = TC ≥200; High TG = triglycerides ≥ 110; Low HDL-C = ≤40 mg/dL; High LDL-C = ≥130; Schooling >JSS = education beyond junior secondary school (usually attended through age 15).
Prevalences age-standardized to WHO standard population.

Figure S1. **Education by age group among urban and rural men and women in Brong Ahafo, Ghana**. Left panels (A) and (C): estimates by age group are for urban females (purple circles) and rural females (green circles). Right panels (B) and (D): estimates by age group are for urban males (purple triangles) and rural males (green triangles). Error bars denote 95% confidence intervals. JSS = Junior Secondary School (usually attended through age 15).

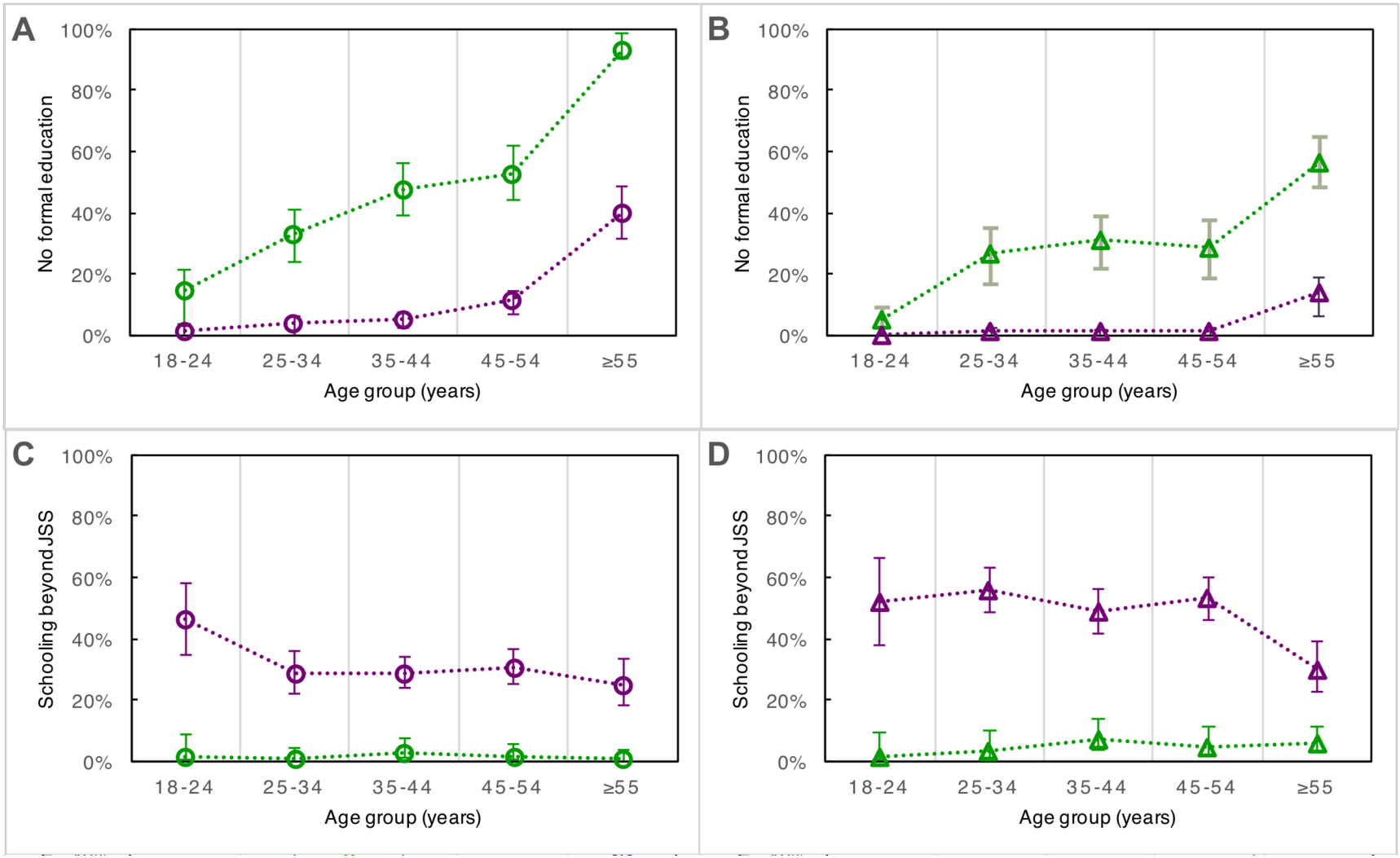


Figure S2. **Mean systolic and diastolic blood pressure by age group in urban and rural men and women in Brong Ahafo, Ghana**. Left panels (A) and (C): mean estimates by age group for urban females (purple circles) and rural females (green circles). Right panels (B) and (D): mean estimates by age group for urban males (purple triangles) and rural males (green triangles). Error bars denote 95% confidence intervals.

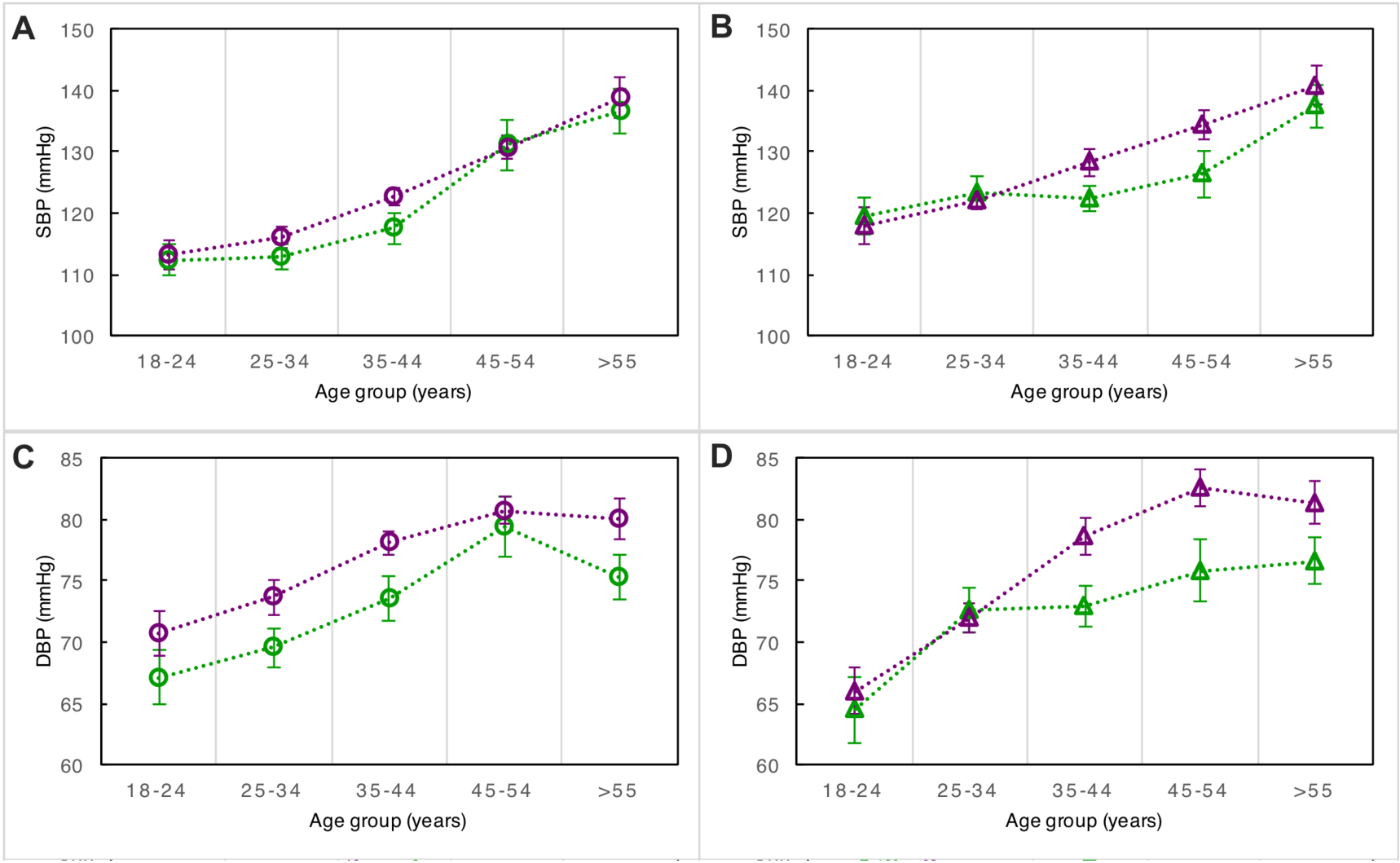


Figure S3. **Mean BMI and overweight prevalence by age group in urban and rural men and women in Brong Ahafo, Ghana**. Left panels (A) and (C): estimates by age group for urban
females (purple circles) and rural females (green circles). In the right panels (B) and (D), estimates by age group are depicted for urban males (purple triangles) and rural males (green triangles). Error bars denote 95% confidence intervals. Overweight is defined as BMI ≥ 25 23 kg/m2.

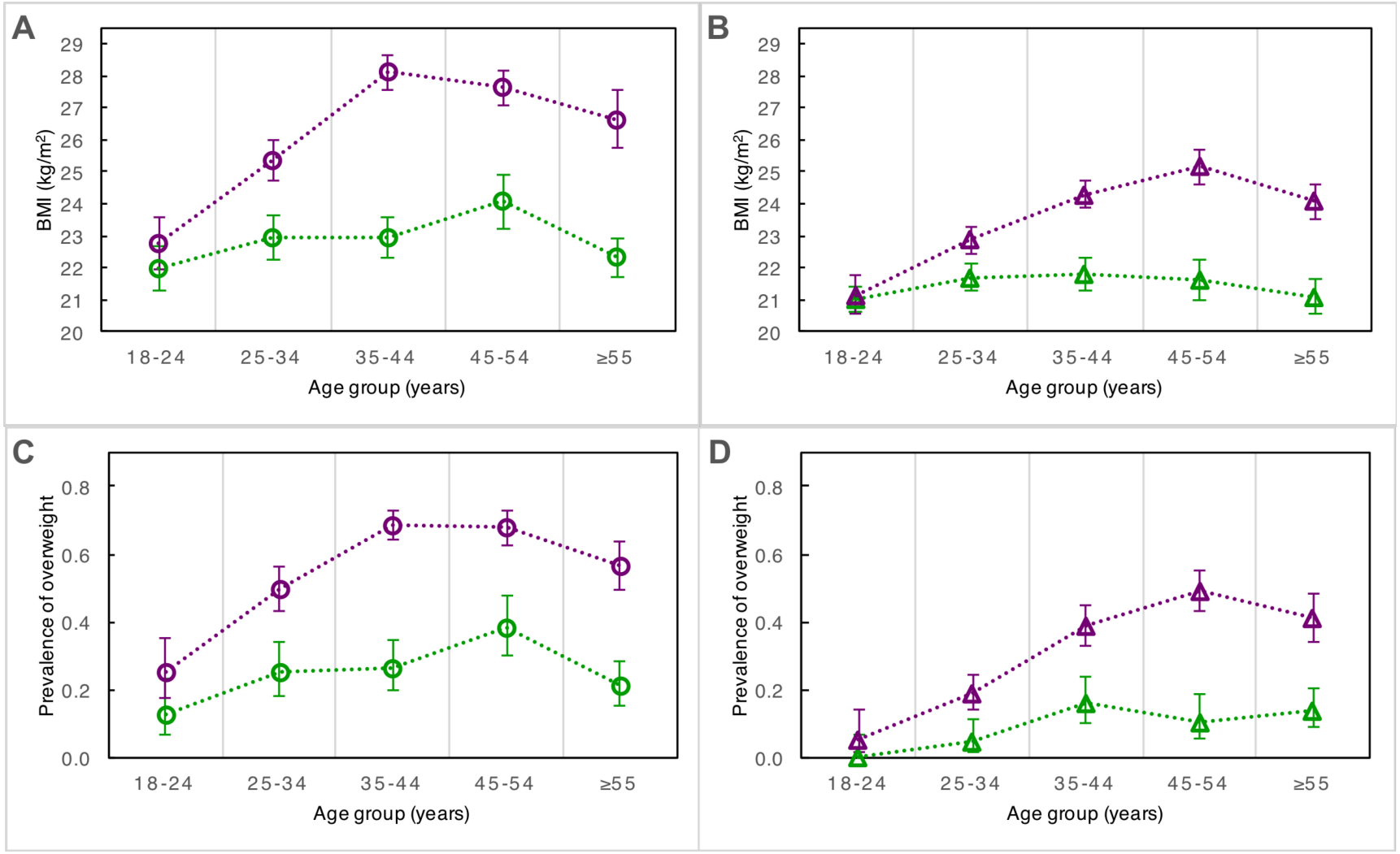


Figure S4. Mean fasting glucose and prevalence of impaired fasting glucose by age group.

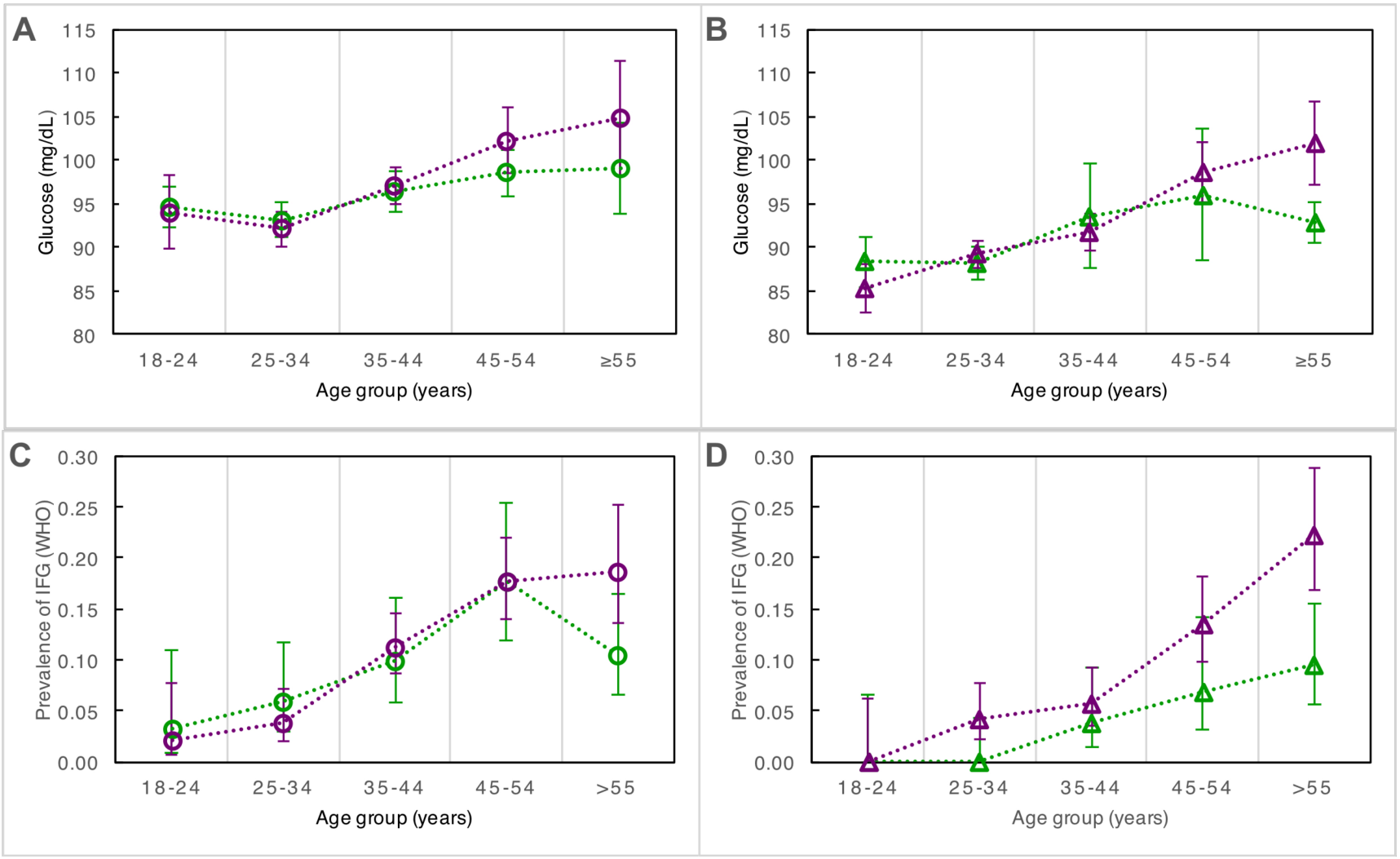


Figure S5. Mean high-density lipoprotein cholesterol by age group.

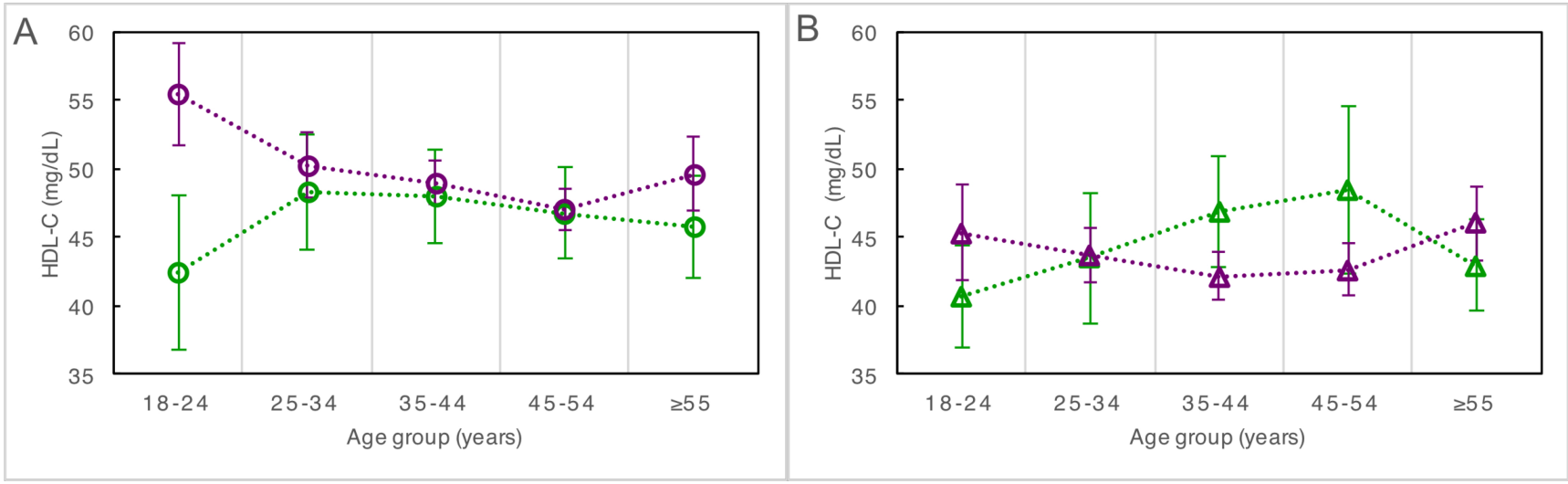


Figure S6. Mean t-PA and PAI-1 levels by age group

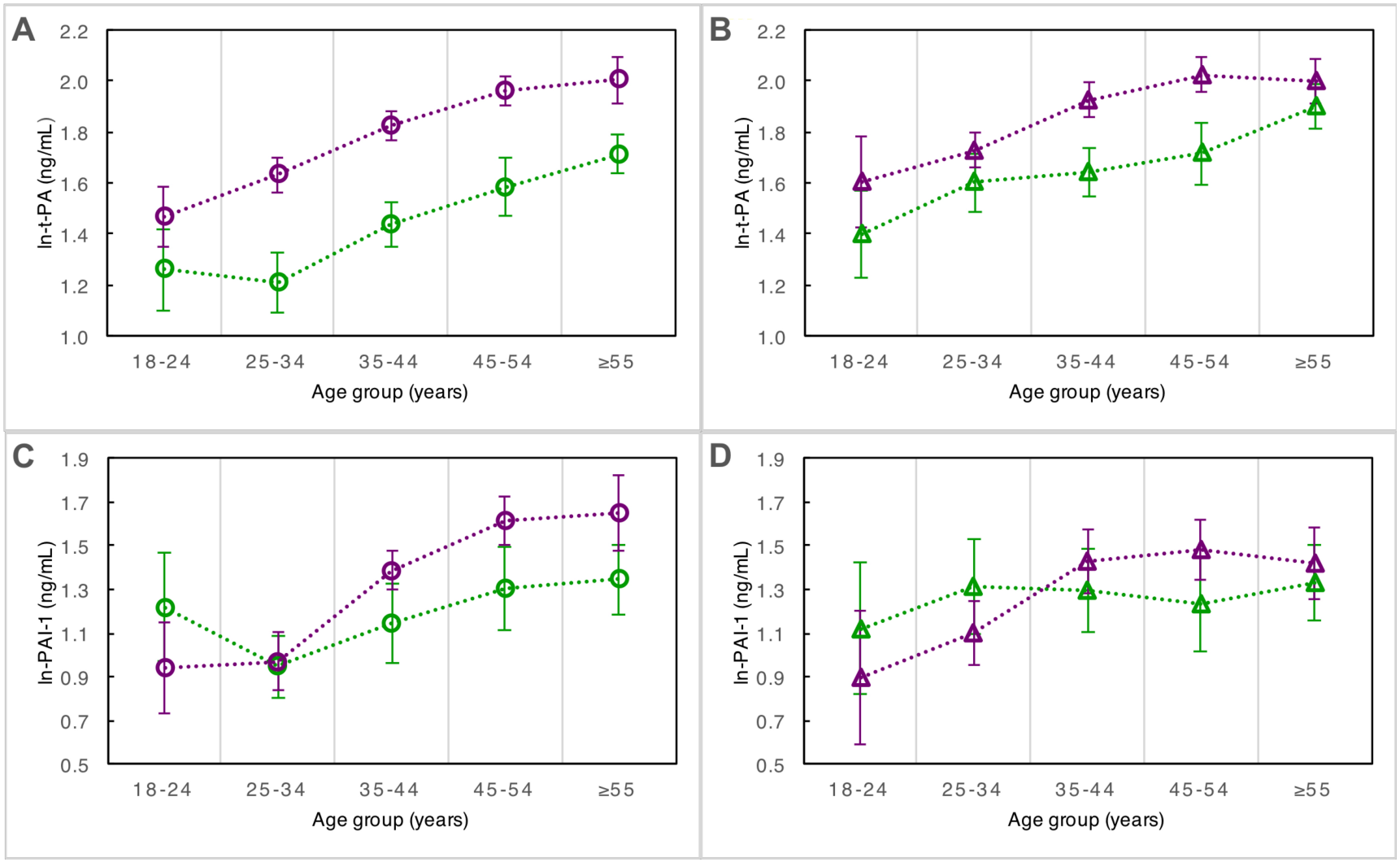


Figure S7. **The BMI-adjusted effect of urban/rural environment on cardiovascular risk**. Absolute differences between urban and rural standardized means (with 95% confidence intervals) are depicted for each risk factor, with colors representing the group with the higher mean (purple: urban; green: rural). Data were adjusted for age, sex, and BMI.

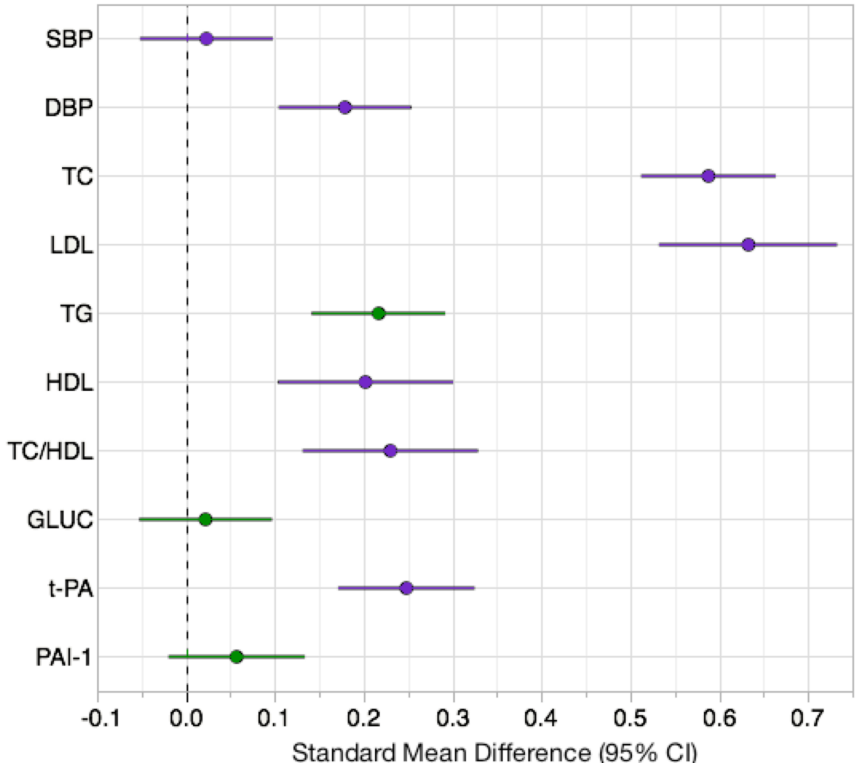


Table S2. Prevalence rates (and 95% confidence intervals) of the metabolic syndrome and its component risk factors among 2220 Ghanaian men and women from urban and rural settings.

	Males		Females	
	Urban	**Rural**	**Urban**	**Rural**
N	721	225	957	317
MetS*	0.126 (0.102, 0.151)	0.078 (0.043, 0.113)	0.214 (0.188, 0.240)	0.112 (0.077, 0.147)
Obesity	0.065 (0.049, 0.086)	0.00 (0.00, 0.017)	0.265 (0.238, 0.294)	0.050 (0.031, 0.080)
Hypertension	0.130 (0.108, 0.157)	0.111 (0.076, 0.159)	0.097 (0.080, 0.118)	0.088 (0.062, 0.125)
Hyperglycemia	0.413 (0.378, 0.450)	0.267 (0.213, 0.328)	0.389 (0.358, 0.420)	0.293 (0.246, 0.346)
High TG	0.239 (0.209, 0.271)	0.173 (0.129, 0.228)	0.315 (0.286, 0.345)	0.293 (0.246, 0.346)
Low HDL-C	0.437 (0.401, 0.473)	0.391 (0.330, 0.456)	0.588 (0.557, 0.619)	0.612 (0.557, 0.664)

*MetS= the metabolic syndrome; prevalence rates age-standardized to the WHO standard population
N= sample size of participants for whom no data was missing

Table S3. Pairwise correlations between cardiovascular risk factors, by sex.

		Females				Males				Homogeneity of Correlation, p-value
Trait 1	**Trait 2**	***r***	**CI**	**N**	**p-value**	***r***	**CI**	**N**	**p-value**	
BMI	**PAI-1**	0.44	(0.41, 0.48)	1881	<.0001	0.41	(0.36, 0.45)	1450	<.0001	0.2008
MAP	**PAI-1**	0.20	(0.15, 0.24)	1881	<.0001	0.27	(0.22, 0.32)	1450	<.0001	0.0239
MAP	**BMI**	0.27	(0.23, 0.32)	1881	<.0001	0.35	(0.30, 0.39)	1450	<.0001	0.0186
HDL	**PAI-1**	-0.20	(-0.25, -0.15)	1275	<.0001	-0.14	(-0.20, -0.08)	950	<.0001	0.1569
HDL	**BMI**	-0.15	(-0.20, -0.10)	1275	<.0001	-0.17	(-0.23, -0.11)	950	<.0001	0.6037
HDL	**MAP**	0.05	(0.00, 0.11)	1275	0.0604	0.01	(-0.06, 0.07)	950	0.8036	0.2993
TG	**PAI-1**	0.34	(0.30, 0.38)	1878	<.0001	0.36	(0.31, 0.40)	1443	<.0001	0.5472
TG	**BMI**	0.24	(0.19, 0.28)	1878	<.0001	0.27	(0.23, 0.32)	1443	<.0001	0.2610
TG	**MAP**	0.13	(0.09, 0.18)	1878	<.0001	0.19	(0.14, 0.24)	1443	<.0001	0.0830
TG	**HDL**	-0.26	(-0.31, -0.20)	1274	<.0001	-0.28	(-0.34, -0.23)	946	<.0001	0.4758
GLUC	**PAI-1**	0.23	(0.19, 0.28)	1881	<.0001	0.15	(0.10, 0.20)	1450	<.0001	0.0116
GLUC	**BMI**	0.15	(0.10, 0.19)	1881	<.0001	0.17	(0.12, 0.22)	1450	<.0001	0.5756
GLUC	**MAP**	0.11	(0.07, 0.16)	1881	<.0001	0.16	(0.11, 0.21)	1450	<.0001	0.1371
GLUC	**HDL**	-0.13	(-0.18, -0.07)	1275	<.0001	-0.12	(-0.18, -0.05)	950	0.0003	0.7984
GLUC	**TG**	0.19	(0.14, 0.23)	1878	<.0001	0.15	(0.10, 0.20)	1443	<.0001	0.3103

r = Pearson correlation coefficient, calculated using residuals after adjustment for age and residence, by sex;
CI = 95% confidence interval;
p-value = probability of *r* if true correlation is zero;
Homogeneity of Correlation, p-value = probability of these data if true correlation is equal for men & women;
Note: p-values > 0.05 have been grayed out.

Table S4. Partial correlations between components of the metabolic syndrome, including PAI-1, by sex.

		Females				Males				Homogeneity of Correlation p-value
Trait 1	**Trait 2**	***r***	**CI**	**N**	**p-value**	***r***	**CI**	**N**	**p-value**	
BMI	**PAI-1**	0.35	(0.30, 0.40)	1274	<.0001	0.29	(0.23, 0.34)	946	<.0001	0.0814
MAP	**PAI-1**	0.07	(0.02, 0.13)	1274	0.0094	0.12	(0.06, 0.19)	946	0.0001	0.2363
MAP	**BMI**	0.21	(0.16, 0.26)	1274	<.0001	0.26	(0.20, 0.32)	946	<.0001	0.2487
HDL	**PAI-1**	-0.09	(-0.14, -0.03)	1274	0.002	-0.01	(-0.08, 0.05)	946	0.6987	0.0842
HDL	**BMI**	-0.08	(-0.13, -0.02)	1274	0.0052	-0.11	(-0.17, -0.05)	946	0.0008	0.4682
HDL	**MAP**	0.13	(0.07, 0.18)	1274	<.0001	0.12	(0.05, 0.18)	946	0.0004	0.7427
TG	**PAI-1**	0.22	(0.17, 0.27)	1274	<.0001	0.25	(0.19, 0.31)	946	<.0001	0.4270
TG	**BMI**	0.07	(0.01, 0.12)	1274	0.0142	0.09	(0.02, 0.15)	946	0.0067	0.6497
TG	**MAP**	0.07	(0.01, 0.12)	1274	0.0134	0.09	(0.02, 0.15)	946	0.0078	0.6872
TG	**HDL**	-0.20	(-0.25, -0.14)	1274	<.0001	-0.25	(-0.31, -0.19)	946	<.0001	0.2210
GLUC	**PAI-1**	0.14	(0.09, 0.20)	1274	<.0001	0.05	(-0.02, 0.11)	946	0.1410	0.0233
GLUC	**BMI**	0.02	(-0.03, 0.08)	1274	0.4626	0.07	(0.00, 0.13)	946	0.043	0.2920
GLUC	**MAP**	0.06	(0.01, 0.12)	1274	0.0241	0.10	(0.04, 0.17)	946	0.0014	0.3414
GLUC	**HDL**	-0.07	(-0.12, -0.01)	1274	0.0128	-0.08	(-0.14, -0.01)	946	0.0202	0.8921
GLUC	**TG**	0.10	(0.04, 0.15)	1274	0.0006	0.07	(0.00, 0.13)	946	0.0447	0.4765

r = Pearson partial correlation coefficient, calculated using residuals after adjustment for age and residence, by sex;
CI = 95% confidence interval;
p-value = probability of *r* if true partial correlation is zero;
Homogeneity of Correlation, p-value = probability of these data if true partial correlation is equal for men & women;
Note: p-values > 0.05 have been grayed out.

Table S5. Partial correlations between the five components of the metabolic syndrome.

Trait	Trait	*r*	CI	N	p-value
MAP	**BMI**	0.27	(0.23, 0.31)	2220	<.0001
HDL	**BMI**	-0.11	(-0.15, -0.07)	2220	<.0001
HDL	**MAP**	0.12	(0.07, 0.16)	2220	<.0001
TG	**BMI**	0.16	(0.12, 0.20)	2220	<.0001
TG	**MAP**	0.11	(0.07, 0.15)	2220	<.0001
TG	**HDL**	-0.24	(-0.28, -0.20)	2220	<.0001
GLUC	**BMI**	0.08	(0.04, 0.12)	2220	0.0002
GLUC	**MAP**	0.09	(0.05, 0.13)	2220	<.0001
GLUC	**HDL**	-0.08	(-0.12, -0.04)	2220	<.0001
GLUC	**TRIG**	0.11	(0.07, 0.15)	2220	<.0001

r = Pearson partial correlation coefficient, calculated using residuals after adjustment for age, sex, and residence;
CI = 95% confidence interval;
p-value = probability of *r* if true partial correlation is zero.

Table S6. Partial correlations between the five components of the metabolic syndrome, by urban or rural residence.

		Rural				Urban				Homogeneity of Correlation p-value
Trait 1	**Trait 2**	***r***	**CI**	**N**	**p-value**	***r***	**CI**	**N**	**p-value**	
MAP	**BMI**	0.24	(0.16, 0.32)	542	<.0001	0.28	(0.24, 0.33)	1678	<.0001	0.3335
HDL	**BMI**	-0.06	(-0.14, 0.02)	542	0.1631	-0.14	(-0.19, -0.09)	1678	<.0001	0.1083
HDL	**MAP**	0.12	(0.04, 0.20)	542	0.0042	0.12	(0.07, 0.17)	1678	<.0001	0.9314
TG	**BMI**	0.09	(0.01, 0.18)	542	0.0297	0.19	(0.14, 0.24)	1678	<.0001	0.0430
TG	**MAP**	0.16	(0.07, 0.24)	542	0.0002	0.09	(0.04, 0.13)	1678	0.0004	0.1449
TG	**HDL**	-0.28	(-0.36, -0.20)	542	<.0001	-0.23	(-0.27, -0.18)	1678	<.0001	0.2250
GLUC	**BMI**	0.13	(0.05, 0.21)	542	0.0021	0.05	(0.01, 0.10)	1678	0.0245	0.1180
GLUC	**MAP**	0.11	(0.02, 0.19)	542	0.0116	0.08	(0.03, 0.13)	1678	0.0011	0.5601
GLUC	**HDL**	-0.09	(-0.17, -0.01)	542	0.0356	-0.08	(-0.13, -0.03)	1678	0.0011	0.8230
GLUC	**TRIG**	0.12	(0.04, 0.20)	542	0.0044	0.11	(0.06, 0.15)	1678	<.0001	0.7341

r = Pearson partial correlation coefficient, calculated using residuals after adjustment for age and sex, by residence;
CI = 95% confidence interval;
p-value = probability that the true partial correlation is zero;
Homogeneity of Correlation, p-value = probability of these data if true partial correlation is equal for urban & rural;
Note: p-values > 0.05 have been grayed out.

Table S7. Partial correlations between the five components of the metabolic syndrome, by sex.

		Females				Males				Homogeneity of Correlation p-value
Trait 1	**Trait 2**	***r***	**CI**	**N**	**p-value**	***r***	**CI**	**N**	**p-value**	
MAP	**BMI**	0.26	(0.20, 0.31)	1274	<.0001	0.31	(0.25, 0.37)	946	<.0001	0.1710
HDL	**BMI**	-0.12	(-0.17, -0.07)	1274	<.0001	-0.12	(-0.18, -0.05)	946	0.0003	0.9642
HDL	**MAP**	0.13	(0.07, 0.18)	1274	<.0001	0.11	(0.05, 0.18)	946	0.0004	0.7842
TG	**BMI**	0.16	(0.11, 0.21)	1274	<.0001	0.17	(0.11, 0.23)	946	<.0001	0.7575
TG	**MAP**	0.09	(0.03, 0.14)	1274	0.0016	0.12	(0.06, 0.18)	946	0.0002	0.4210
TG	**HDL**	-0.22	(-0.27, -0.17)	1274	<.0001	-0.26	(-0.32, -0.20)	946	<.0001	0.3481
GLUC	**BMI**	0.08	(0.02, 0.13)	1274	0.0060	0.08	(0.02, 0.15)	946	0.0106	0.8846
GLUC	**MAP**	0.08	(0.02, 0.13)	1274	0.0074	0.11	(0.05, 0.17)	946	0.0006	0.4021
GLUC	**HDL**	-0.09	(-0.14, -0.03)	1274	0.0023	-0.08	(-0.14, -0.01)	946	0.0189	0.8331
GLUC	**TRIG**	0.13	(0.08, 0.19)	1274	<.0001	0.08	(0.02, 0.14)	946	0.0141	0.2184

r = Pearson partial correlation coefficient, calculated using residuals after adjustment for age and residence, by sex;
CI = 95% confidence interval;
p-value = probability of *r* if true partial correlation is zero;
Homogeneity of Correlation, p-value = probability of these data if true partial correlation is equal for men & women;
Note: p-values > 0.05 have been grayed out.

Figure S8. Proportions of participants with $N \in [0,5]$ component risk factors of the metabolic syndrome, by sex and environment. Length of rectangles represents the percentage of participants with *N* risk factors in each labeled group (UM = urban males; UF= urban females; RM= rural males; RF= rural females; Total= all 2220 participants for whom no data were missing). Areas of rectangles for UM, UF, RM, RF represent proportions with respect to all 2220 participants.

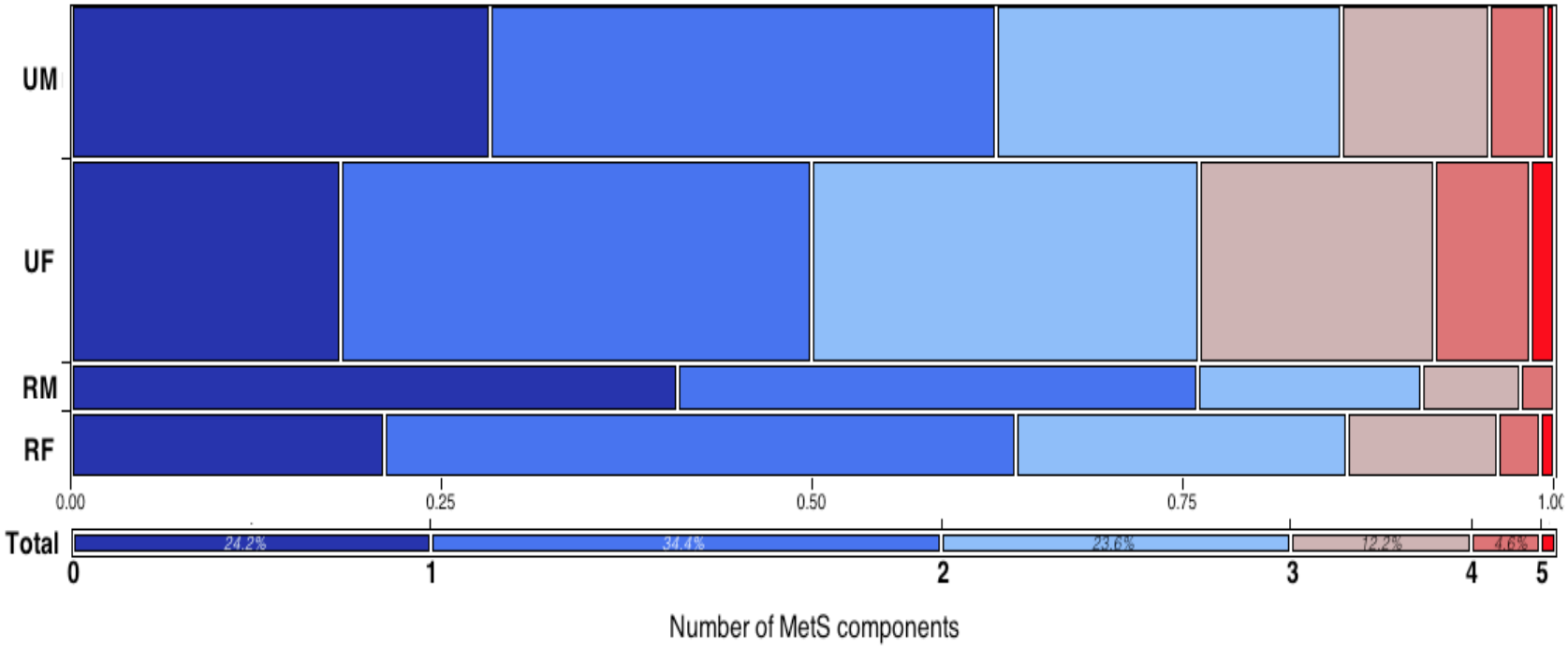


Figure S9: Isolated cases of risk factors associated with the metabolic syndrome. Among 764 subjects who had exactly one risk factor, the proportion for whom the isolated case was hypertension (green), impaired fasting glucose (pink), low high-density lipoprotein cholesterol (orange), or other (blue), by sex and urban/rural residence. Only samples with complete data for all five risk factors were considered for this analysis (N=2220).

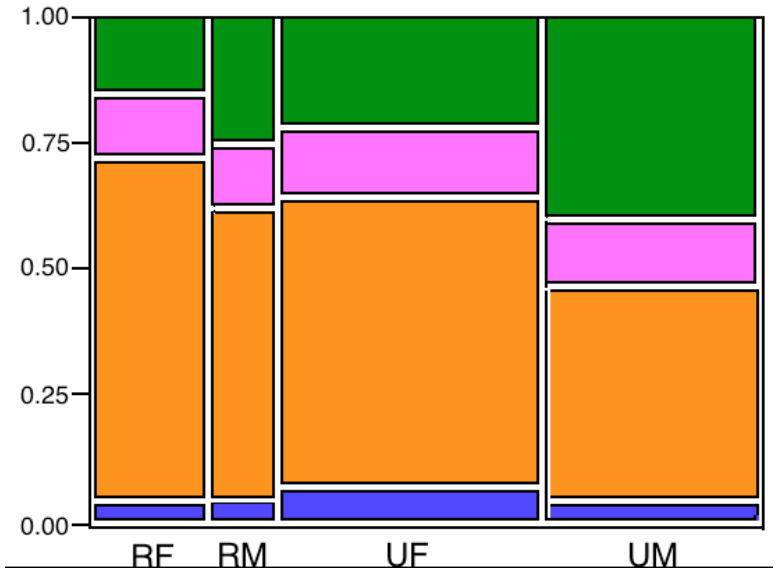
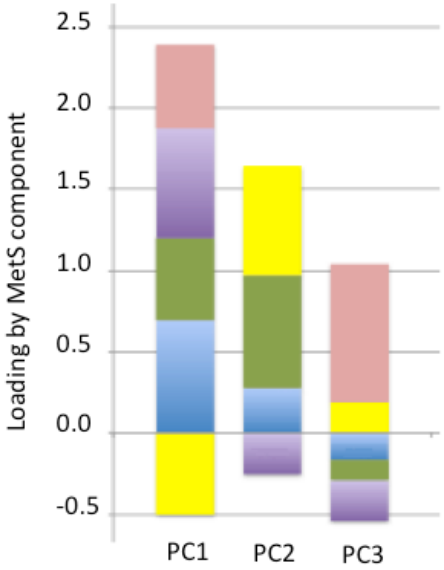


Figure S10. Loadings of the first three principal components of the five risk factors that define the metabolic syndrome. Pink=glucose, yellow=HDL, purple=triglycerides, green=MAP, blue=BMI. Negative loadings are below zero on the vertical axis. The size of a "stack" of loadings is related to the variance explained by the particular principal component. The data were adjusted for age, sex, and residence.



Note that because low values of HDL are associated with increased risk, HDL *as a risk factor* clusters with the risk factors of opposite sign.

Figure S11. Moving medians and 1st and 3rd quartiles of standardized PAI-1 values as a function of the first three standardized principal components of MetS risk factors, for men (blue) and women (red). PAI-1 and MetS risk factors were adjusted for age, sex, and residence. Period for quartiles = 100. Data smoothed using cubic spline (see Methods).

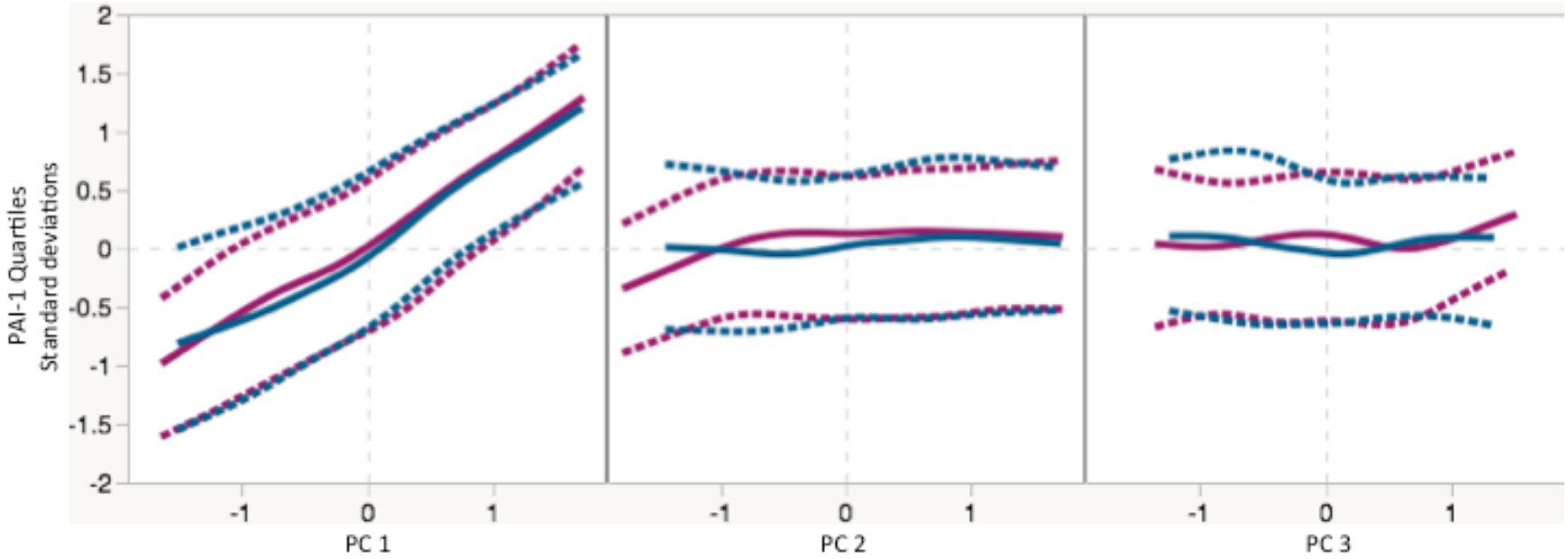
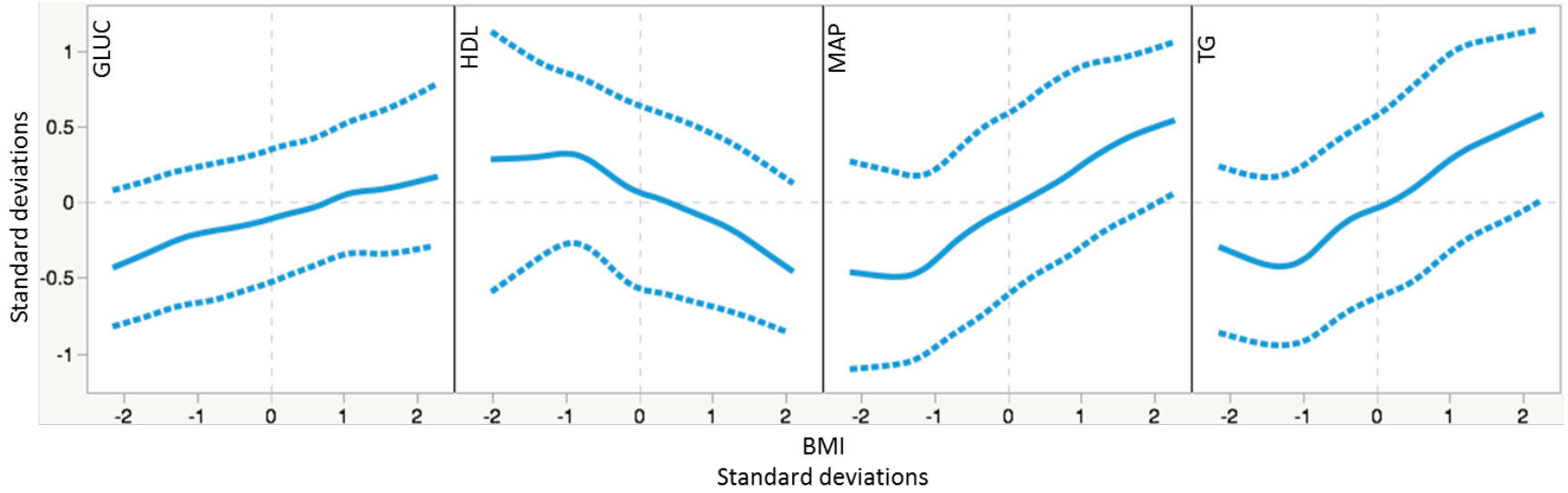


Figure S12 Moving medians and 1st and 3rd quartiles of standardized GLUC, HDL, MAP, and TG as a function of standardized BMI. Data adjusted for age, sex, and residence. Period for quartiles = 100. Data smoothed using cubic spline (see Methods).



Appendix B: Supplemental Material, Chapter IV

B-1

$$Y_i = \begin{cases} \beta_X X_i + \beta_Z Z_i + \varepsilon_i, & Z_i \le 0 \\ \beta_X X_i + \beta_Z Z_i + \beta_{XZ} X_i Z_i + \varepsilon_i, & Z_i > 0 \end{cases}$$

$$Cov(X,Y) = E(XY) - E(X)\cdot E(Y)$$
$$= \frac{1}{2}\left(E\left(X\cdot\left(\beta_X X + \beta_Z Z + \varepsilon \mid Z \le 0\right)\right)\right) + \frac{1}{2}\left(E\left(X\cdot\left(\beta_X X + \beta_Z Z + \beta_{XZ} XZ + \varepsilon \mid Z > 0\right)\right)\right)$$
$$= \frac{1}{2}\left(\beta_X \cdot E(X^2) + \beta_Z \cdot E(XZ \mid Z \le 0) + E(X\varepsilon)\right) + \frac{1}{2}\left(\beta_X \cdot E(X^2) + \beta_Z \cdot E(XZ \mid Z > 0) + \beta_{XZ} \cdot E(X\cdot XZ \mid Z > 0) + E(X\varepsilon)\right)$$
$$= \beta_X \cdot E(X^2) + \frac{\beta_{XZ}}{2} E(X^2)\cdot E(Z \mid Z > 0)$$
$$= \beta X k + \sqrt{\frac{2}{\pi}}\cdot\frac{\beta_{XZ} k}{2}$$
$$= k\left(\beta_X + \frac{\beta_{XZ}}{\sqrt{2\pi}}\right)$$
$$R_X^2 = \frac{\left[Cov(X,Y)\right]^2}{Var(X)\cdot Var(Y)} = k\left(\beta_X + \frac{\beta_{XZ}}{\sqrt{2\pi}}\right)^2$$

Recall from the Statistical Overview that the variance of *Y* can be set to 1 by adjusting the variance of the error term; note also that $\sqrt{\frac{2}{\pi}}$ is the expected value of a truncated standard normal distribution, Z>0, as per the formula $\frac{\phi(\alpha) - \phi(\beta)}{\Phi(\beta) - \Phi(\alpha)}$ where the interval $[\alpha, \beta)$ denotes support for Z.

B-2

$$Y_i = \begin{cases} \beta_X X_i + \beta_Z Z_i + \varepsilon_i, & Z_i \le 0 \\ \beta_X X_i + \beta_Z Z_i + \beta_{XZ} X_i Z_i + \varepsilon_i, & Z_i > 0 \end{cases}$$

$$Cov(XZ,Y) = E(XZY) - E(XZ)\cdot E(Y)$$
$$= \frac{1}{2}\Big(E\Big(XZ\cdot\big(\beta_X X + \beta_Z Z + \varepsilon \mid Z \le 0\big)\Big)\Big) + \frac{1}{2}\Big(E\Big(XZ\cdot\big(\beta_X X + \beta_Z Z + \beta_{XZ} XZ + \varepsilon \mid Z > 0\big)\Big)\Big)$$
$$= \frac{1}{2}\Big(\beta_X \cdot E(X^2)\cdot E(Z \mid Z \le 0) + \beta_Z \cdot E(X)\cdot E(Z^2 \mid Z \le 0) + E(\varepsilon)\cdot E(XZ \mid Z \le 0)\Big)$$
$$+ \frac{1}{2}\Big(\beta_X \cdot E(X^2)\cdot E(Z \mid Z > 0) + \beta_Z \cdot E(X)\cdot E(Z^2 \mid Z > 0) + \beta_{XZ} \cdot E(X^2)\cdot E(Z^2 \mid Z > 0) + E(\varepsilon)\cdot E(XZ \mid Z > 0)\Big)$$
$$= \frac{\beta_{XZ} \cdot E(X^2)\cdot E(Z^2 \mid Z > 0)}{2} = \frac{\beta_{XZ} k}{2}$$

$$R^2_{XZ} = \frac{\Big[Cov(XZ,Y)\Big]^2}{Var(XZ)\cdot Var(Y)} = \frac{\beta^2_{XZ} k}{4}$$

Note that

$$Var(XZ) = E(X^2Z^2) - \Big[E(XZ)\Big]^2 = E(X^2)E(Z^2) = k$$

B-3

$$Y=\begin{cases}\beta_1 X_i+\beta_Z Z_i+\varepsilon_i, & 0<\Phi\left(Z_i\right)<\frac{1}{q}\\ \beta_2 X_i+\beta_Z Z_i+\varepsilon_i, & \frac{1}{q}<\Phi\left(Z_i\right)<\frac{2}{q}\\ \vdots \quad \vdots \quad \vdots & \vdots\\ \beta_q X_i+\beta_Z Z_i+\varepsilon_i, & \frac{q-1}{q}<\Phi\left(Z_i\right)<1\end{cases}$$

where

$$\beta_j = jc,$$
$$j\in\left[1,2,\ldots,q\right]$$

$$Cov(X,Y)=E(XY)-E(X)E(Y)=$$
$$=\frac{E(X^2)}{q}\left(\beta_1+\beta_2+\ldots+\beta_n\right)$$
$$=\frac{k}{q}\left(c+2c+\ldots+qc\right)=\left(\frac{kc}{q}\right)\cdot\left(\frac{q(q+1)}{2}\right)$$
$$=\frac{kc(q+1)}{2}$$
$$R_x^2=\frac{\left[Cov(X,Y)\right]^2}{Var(X)\cdot Var(Y)}=\frac{\frac{k^2c^2(q+1)^2}{4}}{k\cdot 1}=\frac{kc^2(q+1)^2}{4}$$

Deriving $Cov(XZ,Y)$ =E(XZY) - E(XZ)E(Y) is complicated by the fact that, after multiplying XZ

with Y , each linear equation in the step function contains the term E(X^2)E(Z), which varies based

on the quantile of Z. However, note that whatever the value E(Z), it has the same magnitude but

different sign for the first and last quantile, the second and second-to-last quantile, etc.

Thus

$$Cov(XZ,Y)=\frac{1}{q}\left(ck\cdot E\left(Z\,|\,\Phi\left(Z_i\right)\in(0,\tfrac{1}{q})\right)\right)+\frac{1}{q}\left(2ck\cdot E\left(Z\,|\,\Phi\left(Z_i\right)\in(\tfrac{1}{q},\tfrac{2}{q})\right)\right)+\ldots$$
$$+\frac{1}{q}\left((q-1)ck\cdot E\left(Z\,|\,\Phi\left(Z_i\right)\in(\tfrac{q-2}{q},\tfrac{q-1}{q})\right)\right)+\frac{1}{q}\left(qck\cdot E\left(Z\,|\,\Phi\left(Z_i\right)\in(\tfrac{q-1}{q},1)\right)\right)$$
$$=\frac{ck}{q}\left((q-1)\cdot E\left(Z\,|\,\Phi\left(Z_i\right)\in(\tfrac{q-1}{q},1)\right)+(q-3)\cdot E\left(Z\,|\,\Phi\left(Z_i\right)\in(\tfrac{q-2}{q},\tfrac{q-1}{q})\right)+\ldots\right)$$

Which can be expressed in closed form as,

$$Cov(XZ,Y)=\frac{ck}{\sqrt{2\pi}}\sum_{j=1}^{\left\lfloor\frac{q}{2}\right\rfloor}\left(q-\left(2j-1\right)\right)\left(\phi\left(\Phi^{-1}\left(1-\tfrac{j}{q}\right)\right)-\phi\left(\Phi^{-1}\left(1-\tfrac{j-1}{q}\right)\right)\right)$$

where $\lfloor x \rfloor$ represents the floor function

B-4

$$Y = \beta_X X_i + \beta_{X\Phi}\left(X_i \cdot \Phi(Z_i)\right) + \beta_Z Z_i + \varepsilon_i$$

$$Var(Y) = E(Y^2) - \left[E(Y)\right]^2$$
$$= \beta_X^2 E(X^2) + \beta_{X\Phi}^2 \cdot E(X^2) \cdot E\left(\left(\Phi(Z)\right)^2\right) + \beta_Z^2 E(Z^2) + \sigma_\varepsilon^2$$
$$= \beta_X^2 k + \frac{\beta_{X\Phi}^2 k}{3} + \beta_Z^2 + \sigma_\varepsilon^2$$

Using the general formula: $Cov(aX, bX + cY) = ab \cdot Var(X) + ac \cdot Cov(X, Y)$

$$= \beta_X Var(X) + \beta_{X\Phi} Cov\left(X, X \cdot \Phi(Z)\right) + \beta_Z Cov(X, Z) + Cov(X, \varepsilon)$$
$$= \beta_X k + \frac{\beta_{X\Phi} k}{2}$$
$$= k\left(\beta_X + \frac{\beta_{X\Phi}}{2}\right)$$
$$R_X^2 = \frac{k^2\left(\beta_X + \frac{\beta_{X\Phi}}{2}\right)^2}{k \cdot Var(Y)} = \frac{k\left(\beta_X + \frac{\beta_{X\Phi}}{2}\right)^2}{Var(Y)}$$
$$Cov(XZ, Y) = E(XZ \cdot Y) - E(XZ)E(Y)$$
$$E\left[XZ \cdot \left(\beta_X X + \beta_{X\Phi}\left(X \cdot \Phi(Z)\right) + \beta_Z Z + \varepsilon\right)\right]$$
$$= \beta_{X\Phi} E(X^2) \cdot E\left(Z \cdot \Phi(Z)\right)$$
$$= \frac{\beta_{X\Phi} k}{2\sqrt{\pi}}$$
$$Var(XZ) = E(X^2 Z^2) - \left[E(XZ)\right]^2 = E(X^2)E(Z^2) = k$$
$$R_{EX}^2 = \frac{\left(\frac{\beta_{X\Phi} k}{2\sqrt{\pi}}\right)^2}{k \cdot Var(Y)} = \frac{\beta_{X\Phi}^2 k}{4\pi \cdot Var(Y)}$$

Note: $E\left(Z \cdot \Phi(Z)\right)$
derived using integration by parts:

$$E\left[Z\cdot\Phi\left(Z\right)\right]$$

$$=\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\cdot x\left[\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy\right]dx$$

$$=\left[\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy\right]_{-\infty}^{\infty}\left[\frac{-1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\right]_{-\infty}^{\infty}-\int_{-\infty}^{\infty}\left[\frac{-1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\right]\left[\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\right]dx$$

$$=0+\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-x^2}dx$$

$$=\frac{\sqrt{\pi}}{2\pi}=\frac{1}{2\sqrt{\pi}}$$

Appendix C: Supplemental Figures and Tables, Chapter V

Table C1. LDL and triglycerides: top ten associations for the ordinal joint interaction model; 2669 exonic SNPs from the NHGRI GWAS Catalog were assessed for 871 Ghanaian participants.

				p-value					
SNP	Chr.	Minor Allele	MAF	Single Trait LDL	Single Trait TG	Bivariate	OJIM	Interaction	Gene
rs12740374	1	T	0.26	**6.09E-06**	0.7453	**1.22E-05**	**3.88E-07**	0.0016	*SORT1**
rs646776	1	C	0.37	0.0002	0.4420	0.0005	4.19E-05	0.0060	*SORT1**
rs9990343	3	G	0.49	0.0049	0.0196	0.0018	0.0005	0.0262	*intergenic*
rs204993	6	G	0.29	0.8785	0.0001	0.0006	0.0006	0.1221	*PBX2*
rs8005962	14	C	0.37	0.0694	0.7659	0.1894	0.0006	0.0002	*intergenic*
rs176095	6	G	0.29	0.9678	0.0002	0.0012	0.0008	0.0664	*PBX2*
rs1803274	3	T	0.21	0.0002	0.8707	0.0003	0.0009	0.5529	*BCHE**
rs301	8	C	0.35	0.3252	0.0003	0.0004	0.0014	0.9986	*LPL**
rs2660753	3	C	0.41	0.1188	0.9027	0.2614	0.0017	0.0004	*intergenic*
rs1845344	4	T	0.27	0.0142	0.4472	0.0310	0.0021	0.0055	*MAD2L1*

Note: SNPs in perfect linkage disequilibrium were not listed: rs7528419 (with rs12740374); rs660240 and rs629301 (with rs646776); rs6445035 (with rs1803274).

Genes marked with an asterisk have been previously associated with lipids in GWAS; p-values in bold were significant after Bonferroni correction; all tests were adjusted for age and sex; single trait tests were not adjusted for multiple testing (i.e. 2 tests).

Columns**:** Chr. = chromosome; MA = minor allele; MAF = minor allele frequency; LDL = low-density lipoprotein cholesterol; TG = triglycerides; Bivariate= MultiPhen, which models genotype as a function of LDL and TG; OJIM adds an interaction term; Interaction = interaction term for the OJIM.

Table C2. Triglycerides and total cholesterol: top ten associations for the univariate (blue) and bivariate (green) tests not featured in Table 5-1; 2669 exonic SNPs from the NHGRI GWAS Catalog were assessed for 1032 Ghanaian participants.

				p-value					
SNP	Chr.	Minor Allele	MAF	Single Trait TG	Single Trait TC	Bivariate	OJIM	Interaction	Gene
rs1803274	3	T	0.21	0.8121	0.0009	0.0024	0.0067	0.7528	*BCHE**
rs3884558	15	A	0.21	0.0011	0.8346	0.0015	0.0034	0.3960	*RORα*
rs7769051	6	A	0.38	0.0012	0.0865	0.0044	0.0114	0.6733	*RPS12*
rs11884476	2	G	0.25	0.3445	0.0012	0.0032	0.0049	0.2349	*PARD3B*
rs229527	22	A	0.34	0.2929	0.0036	0.0007	0.0021	0.7133	*C1QTNF6*

Note: rs180327, rs6445035, and rs3884558 were also in top 10 for bivariate model; not listed: rs6445035, which was in perfect linkage disequilibrium with rs1803274.

Genes marked with an asterisk have been previously associated with lipids or atherosclerotic cardiovascular disease; all tests were adjusted for age and sex; single trait tests were not adjusted for multiple testing (i.e. 2 tests), and only top 10

Columns**:** Chr. = chromosome; MA = minor allele; MAF = minor allele frequency; TG = triglycerides; TC = total cholesterol; Bivariate = MultiPhen, which models genotype as a function of TG and TC; OJIM adds an interaction term; Interaction = interaction term for the OJIM.

Table C3. Triglycerides and total cholesterol: top ten p-values for the interaction term of the ordinal joint interaction model; 2669 exonic SNPs from the NHGRI GWAS Catalog were assessed for 1032 Ghanaian participants.

SNP	Chr.	Minor Allele	MAF	p-value					Gene
				Single Trait TG	Single Trait TC	Bivariate	OJIM	Interaction	
rs3803064	12	A	0.28	0.4633	0.8875	0.6758	0.0004	2.70E-05	*RPH3A*
rs646776	1	C	0.37	0.7855	0.0018	0.0041	**1.36E-05**	0.0002	*SORT1**
rs7076247	10	T	0.42	0.2001	0.2449	0.3339	0.0042	0.0010	*CACNB2*
rs12740374	1	T	0.26	0.9658	0.0002	0.0003	**5.77E-06**	0.0010	*SORT1**
rs8041863	15	T	0.43	0.1649	0.3551	0.2859	0.0040	0.0010	*ACAN*
rs1004446	11	A	0.45	0.6074	0.8572	0.8748	0.0181	0.0018	*IGF2AS**
rs1829883	5	T	0.29	0.2119	0.6276	0.4696	0.0102	0.0018	*intergenic*
rs743777	22	A	0.41	0.2821	0.3814	0.289	0.0077	0.0021	*IL2RB*
rs17039212	2	A	0.21	0.8952	0.3965	0.6716	0.0223	0.0030	*intergenic*
rs6732426	2	C	0.30	0.0843	0.4082	0.1948	0.0080	0.0034	*THADA*

Note: SNPs in perfect linkage disequilibrium were not listed: rs7528419 (with rs12740374); rs660240 and rs629301 (with rs646776).

Genes marked with an asterisk have been previously associated with lipids in GWAS; p-values in bold were significant after Bonferroni correction; all tests were adjusted for age and sex; single trait tests were not adjusted for multiple testing (i.e. 2 tests).

Columns**:** Chr. = chromosome; MA = minor allele; MAF = minor allele frequency; TG = triglycerides; TC = total cholesterol; Bivariate = MultiPhen, which models genotype as a function of TG and TC; OJIM adds an interaction term; Interaction = interaction term for the OJIM.

Table C4. HDL and triglycerides: top ten associations for the univariate (blue) and bivariate (green) tests not featured in Table 5-2; 2669 exonic SNPs from the NHGRI GWAS Catalog were assessed for 869 Ghanaian participants.

				p-value					
SNP	Chr.	Minor Allele	MAF	Single Trait HDL	Single Trait TG	Bivariate	OJIM	Interaction	Gene
rs176095	6	G	0.29	0.0410	0.0002	0.0010	0.0021	0.3673	*intergenic*
rs3129055	6	G	0.20	0.1922	0.0002	0.0009	0.0030	0.9927	*HLA-F*
rs739401	11	T	0.25	0.0005	0.5889	0.0020	0.0045	0.4205	*CARS*
rs261360	20	A	0.41	0.0006	0.1820	0.0016	0.0050	0.8750	*intergenic*
rs12999542	2	C	0.25	0.2765	0.0007	0.0055	0.0049	0.1171	*ILRL1*
rs2043085	15	C	0.28	0.0196	0.0354	0.0008	0.0018	0.3859	*LIPC**

Note: rs176095 and rs3129055 were also in top 10 for bivariate model.

Genes marked with an asterisk have been previously associated with lipids in GWAS; all tests were adjusted for age and sex; single trait tests were not adjusted for multiple testing (i.e. 2 tests), and only top 10

Columns**:** Chr. = chromosome; MA = minor allele; MAF = minor allele frequency; HDL = high-density lipoprotein cholesterol; TG = triglycerides; Bivariate= MultiPhen, which models genotype as a function of HDL and TG; OJIM adds an interaction term; Interaction = interaction term for the OJIM.

Table C5. HDL and triglycerides: top ten p-values for the interaction term of the ordinal joint interaction model; 2669 exonic SNPs from the NHGRI GWAS Catalog were assessed for 869 Ghanaian participants.

SNP	Chr.	Minor Allele	MAF	p-value					Gene
				Single Trait HDL	Single Trait TG	Bivariate	OJIM	Interaction	
rs7493138	14	T	0.29	0.1672	0.7369	0.3873	0.0020	0.0003	*FOXG1*
rs1512268	8	C	0.28	0.4558	0.9887	0.7306	0.0068	0.0007	*NKX3.1*
rs442177	4	T	0.47	0.1994	0.8760	0.4629	0.0065	0.0010	*AFF1**
rs774359	9	C	0.22	0.3969	0.0788	0.1605	0.0028	0.0012	*c9orf72*
rs2880058	1	A	0.24	0.4610	0.4329	0.5508	0.0103	0.0015	*NOS1AP*
rs13188386	5	A	0.37	0.8803	0.9249	0.9817	0.0191	0.0017	*GHR*
rs1812175	4	A	0.35	0.6885	0.0673	0.1144	0.0028	0.0018	*HHIP*
rs727088	18	A	0.22	0.6255	0.3390	0.5324	0.0141	0.0022	*CD226*
rs6811556	4	C	0.46	0.7231	0.5182	0.7209	0.0226	0.0028	*intergenic*
rs7689420	4	T	0.36	0.6504	0.0689	0.1115	0.0041	0.0029	*HHIP*

Genes marked with an asterisk have been previously associated with lipids in GWAS; p-values in bold were significant after Bonferroni correction; all tests were adjusted for age and sex; single trait tests were not adjusted for multiple testing (i.e. 2 tests).

Columns**:** Chr. = chromosome; MA = minor allele; MAF = minor allele frequency; HDL = high-density lipoprotein cholesterol; TG = triglycerides; Bivariate= MultiPhen, which models genotype as a function of HDL and TG; OJIM adds an interaction term; Interaction = interaction term for the OJIM.

Table C6. HDL and LDL: top ten associations for the ordinal joint interaction model; 2669 exonic SNPs from the NHGRI GWAS Catalog were assessed for 869 Ghanaian participants.

SNP	Chr.	Minor Allele	MAF	p-value					Gene
				Single Trait HDL	Single Trait LDL	Bivariate	OJIM	Interaction	
rs12740374	1	T	0.26	0.8309	**6.09E-06**	**1.20E-05**	2.77E-05	0.2898	*SORT1**
rs7528419	1	G	0.26	0.8309	**6.09E-06**	**1.20E-05**	2.77E-05	0.2898	*SORT1**
rs7499892	16	T	0.44	8.83E-05	0.1070	3.86E-05	0.0001	0.7899	*CETP**
rs629301	1	G	0.37	0.4601	0.0002	0.0003	0.0004	0.1802	*SORT1**
rs1803274	3	T	0.21	0.1531	0.0002	0.0002	0.0005	0.3780	*BCHE**
rs6445035	3	A	0.21	0.1531	0.0002	0.0002	0.0005	0.3780	*BCHE**
rs646776	1	C	0.37	0.6050	0.0002	0.0003	0.0005	0.2189	*SORT1**
rs660240	1	T	0.37	0.6050	0.0002	0.0003	0.0005	0.2189	*SORT1**
rs247616	16	T	0.26	**1.49E-05**	0.7736	0.0002	0.0007	0.8683	*CETP**
rs17197037	14	A	0.26	0.1752	0.0584	0.0671	0.0016	0.0016	*HNRNPC*
rs1405069	6	A	0.26	0.3473	0.0089	0.0286	0.0018	0.0048	*PI16*
rs9447004	6	A	0.32	0.0028	0.1872	0.0019	0.0019	0.1235	*CD109**
rs261360	20	A	0.41	0.0006	0.2208	0.0006	0.0020	0.8078	*RPS21P7*

Genes marked with an asterisk have been previously associated with lipids in GWAS; p-values in bold were significant after Bonferroni correction; all tests were adjusted for age and sex; single trait tests were not adjusted for multiple testing (i.e. 2 tests).

Columns**:** Chr. = chromosome; MA = minor allele; MAF = minor allele frequency; HDL = high-density lipoprotein cholesterol; LDL = low-density lipoprotein cholesterol; Bivariate= MultiPhen, which models genotype as a function of HDL and LDL; OJIM adds an interaction term; Interaction = interaction term for the OJIM.

Table C7. Total cholesterol and HDL: top ten associations for the univariate tests not featured in Table 5-3; 2669 exonic SNPs from the NHGRI GWAS Catalog were assessed for 869 Ghanaian participants.

				p-value					
SNP	Chr.	Minor Allele	MAF	Single Trait TC	Single Trait HDL	Bivariate	OJIM	Interaction	Gene
rs739401	11	T	0.25	0.2376	0.0005	0.0022	0.0032	0.2152	*CARS*

Note: the bivariate model had the same top 10 SNPs as the OJIM (although the order was different).

Genes marked with an asterisk have been previously associated with lipids in GWAS; p-values in bold were significant after Bonferroni correction; all tests were adjusted for age and sex; single trait tests were not adjusted for multiple testing (i.e. 2 tests).

Columns**:** Chr. = chromosome; MA = minor allele; MAF = minor allele frequency; TC = total cholesterol; HDL = high-density lipoprotein cholesterol; Bivariate= MultiPhen, which models genotype as a function of TC and HDL; OJIM adds an interaction term; Interaction = interaction term for the OJIM.

Table C8. Total cholesterol and HDL: top ten p-values for the interaction term of the ordinal joint interaction model; 2669 exonic SNPs from the NHGRI GWAS Catalog were assessed for 869 Ghanaian participants.

SNP	Chr.	Minor Allele	MAF	p-value					Gene
				Single Trait TC	Single Trait HDL	Bivariate	OJIM	Interaction	
rs1491942	12	C	0.28	0.3594	0.2937	0.2208	0.0012	0.0003	*LRRK2*
rs9263871	6	G	0.38	0.4594	0.6837	0.7523	0.0155	0.0017	*HCG27* *
rs1537415	9	G	0.27	0.3163	0.5580	0.5896	0.0133	0.0019	*GLT6D1*
rs7153703	14	G	0.38	0.8167	0.9323	0.9735	0.0250	0.0023	*FRMD6*
rs17197037	14	A	0.26	0.2058	0.1961	0.0743	0.0024	0.0025	*HNRNPC*
rs16889440	6	T	0.29	0.0505	0.4275	0.1474	0.0054	0.0029	*KIAA0319*
rs667282	15	C	0.34	0.3948	0.3970	0.3158	0.0129	0.0036	*CHRNA5*
rs1557351	18	C	0.25	0.8405	0.7114	0.8767	0.0328	0.0036	*WDR7*
rs406936	6	A	0.38	0.0553	0.0682	0.0766	0.0043	0.0046	*SKIV2L*
rs11888559	2	C	0.44	0.0531	0.9975	0.1153	0.0067	0.0049	*CYP20A1*

Genes marked with an asterisk have been previously associated with lipids in GWAS; p-values in bold were significant after Bonferroni correction; all tests were adjusted for age and sex; single trait tests were not adjusted for multiple testing (i.e. 2 tests).

Columns**:** Chr. = chromosome; MA = minor allele; MAF = minor allele frequency; TC = total cholesterol; HDL = high-density lipoprotein cholesterol; Bivariate= MultiPhen, which models genotype as a function of TC and HDL; OJIM adds an interaction term; Interaction = interaction term for the OJIM.

<u>Figure C1.</u> QQ-plot depicting robustness of ordinal and linear interaction models to outliers. Simulated data were approximately normally distributed, but with outliers, which were generated by replacing the normally distributed error term with one drawn from a *t*-distribution (df=8). The covariate (tested for interaction with the SNP) was correlated with the outcome at r=0.30. This plot is representative of results from 10 repeated runs.

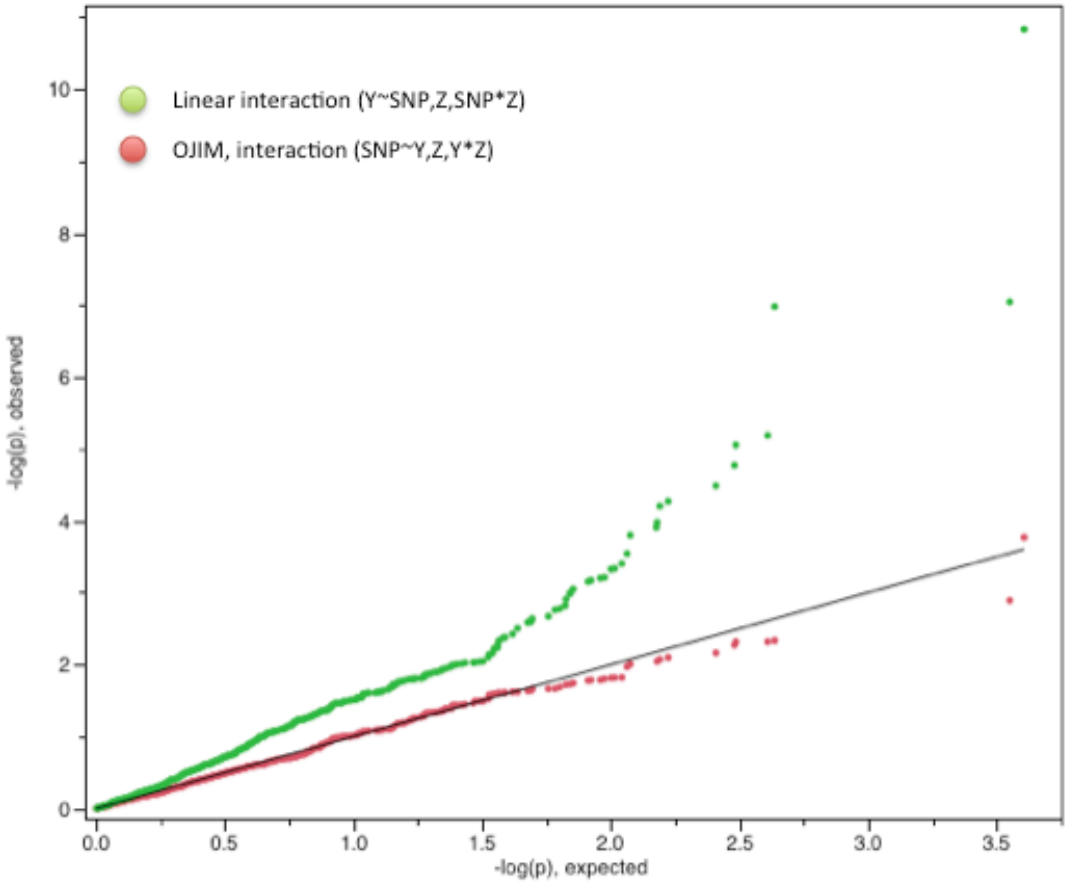
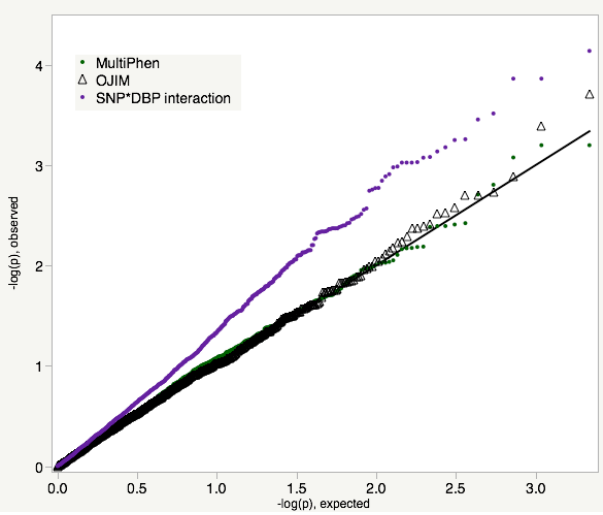


Figure C2. QQ-plots of p-values for tests assessing 2269 NHGRI SNPs

(A) QQ-plots of p-values for tests assessing 2269 NHGRI SNPs for association with systolic blood pressure (SBP) and diastolic blood pressure (DBP) in 1032 Ghanaian participants. Joint tests of SBP and DBP with MultiPhen (green) and the ordinal joint interaction model (OJIM) (black triangle); tests of the SNP-by-DBP interaction term in a linear regression model with SBP as the outcome (purple). Results for the SNP-by-SBP interaction term were less inflated for Type 1 error (not shown).



(B) Similar QQ-plots for tests of HDL and triglycerides (TG). Joint tests of HDL and TG with MultiPhen (green) and the OJIM (black triangle); tests of the SNP-by-HDL interaction term in a linear regression model with TG as the outcome (purple). Results for the SNP-by-TG interaction term were less inflated for Type 1 error (not shown).

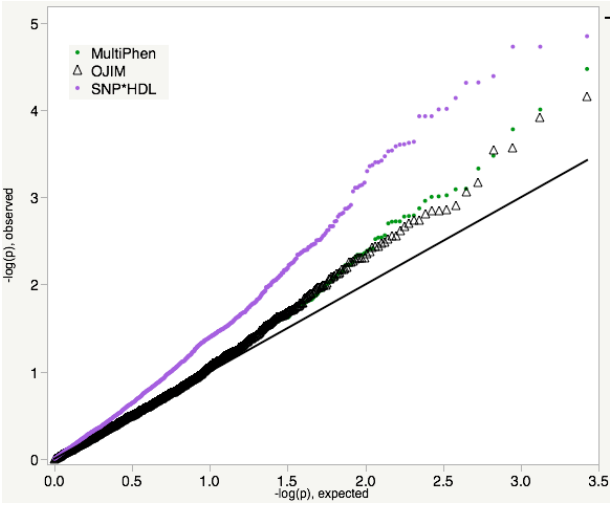
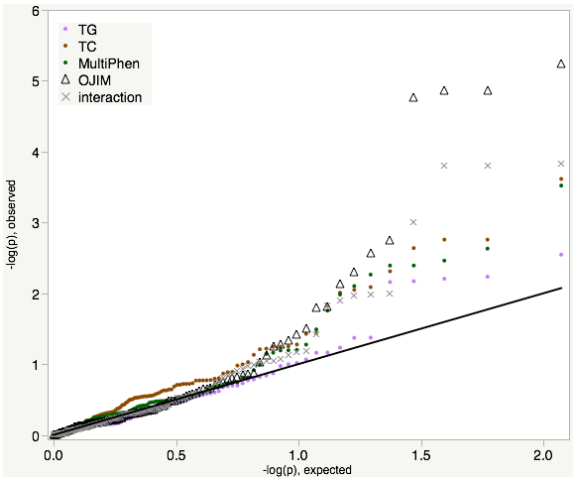
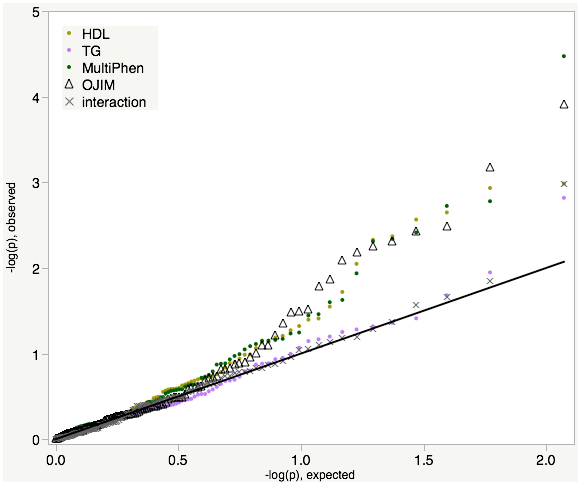


Figure C3. QQ-plots of p-values for tests assessing 116 lipid-associated SNPs for association with (A) triglycerides and total cholesterol; (B) HDL and triglycerides; and (C) total cholesterol and HDL in 1032 Ghanaian participants.

A



B



C

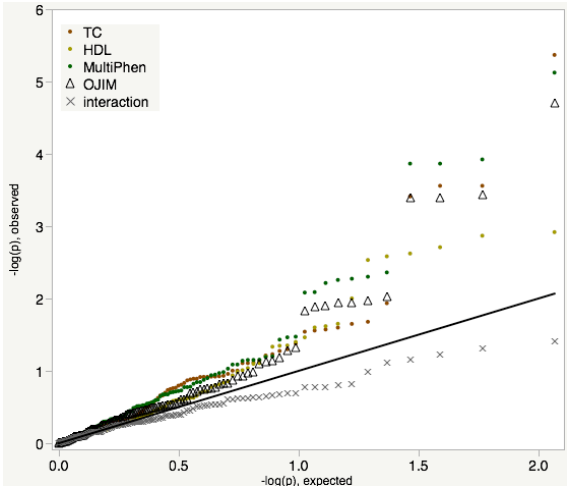


Figure C4. Six sets of ten randomly drawn phenotypes from the NHGRI GWAS Catalog. Asterisks denote phenotypes related to cardiovascular disease.

Cognitive performance	Sexual dimorphism	Hepatocellular carcinoma
Biochemical measures	Smoking behavior	HIV-1 control
Ulcerative colitis	Primary tooth development	HbA2 levels
Chronic kidney disease	Height	epilepsies
Prostate cancer	Coronary heart disease*	Lipid traits*
Erythrocytes	Hyperactive-impulsive symptoms	Kawasaki disease
Bone mineral density	HDL cholesterol*	IgG glycosylation
Hepatitis C-induced cirrhosis	Coronary heart disease*	Obesity
Folate pathway	Response to antipsychotics	Nephropathy
Bone mineral density hip	Conduct disorder interaction	Cholesterol total*
Telomere length	Schizophrenia	Cognitive performance
Dengue shock syndrome	Orofacial clefts	Information processing speed
Type 1 diabetes	ADHD	Prostate cancer
Height	Graves disease	Inattentive symptoms
Bone mineral density	Orofacial clefts	Schizophrenia
Height	Crohns disease	Radiation response
Sub-clinical atherosclerosis*	Height	Pagets disease
Type 2 diabetes	Multiple sclerosis	Obesity-related traits*
Breast cancer	Coronary heart disease*	Migraine
Fibrinogen	Personality dimensions	RR interval heartrate

REFERENCES

1. Moran, A.E. *et al.* Temporal trends in ischemic heart disease mortality in 21 world regions, 1980 to 2010: the Global Burden of Disease 2010 study. *Circulation* **129**, 1483-92 (2014).
2. Yusuf, S. *et al.* Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. *Lancet* **364**, 937-52 (2004).
3. Anand, S.S. *et al.* Risk factors for myocardial infarction in women and men: insights from the INTERHEART study. *Eur Heart J* **29**, 932-40 (2008).
4. Yusuf, S., Reddy, S., Ounpuu, S. & Anand, S. Global burden of cardiovascular diseases: Part II: variations in cardiovascular disease by specific ethnic groups and geographic regions and prevention strategies. *Circulation* **104**, 2855-64 (2001).
5. Williams, S.M. *et al.* A population-based study in Ghana to investigate inter-individual variation in plasma t-PA and PAI-1. *Ethn Dis* **17**, 492-7 (2007).
6. Yusuf, S., Reddy, S., Ounpuu, S. & Anand, S. Global burden of cardiovascular diseases: part I: general considerations, the epidemiologic transition, risk factors, and impact of urbanization. *Circulation* **104**, 2746-53 (2001).
7. Ezzati, M. *et al.* Rethinking the "diseases of affluence" paradigm: global patterns of nutritional risks in relation to economic development. *PLoS Med* **2**, e133 (2005).
8. Daar, A.S. *et al.* Grand challenges in chronic non-communicable diseases. *Nature* **450**, 494-6 (2007).
9. Manolio, T.A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747-53 (2009).
10. Lozano, R. *et al.* Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* **380**, 2095-128 (2012).
11. Moran, A.E., Roth, G.A., Narula, J. & Mensah, G.A. 1990-2010 global cardiovascular disease atlas. *Glob Heart* **9**, 3-16 (2014).
12. Wang, T.J. New cardiovascular risk factors exist, but are they clinically useful? *Eur Heart J* **29**, 441-4 (2008).
13. Wilson, P.W. *et al.* Prediction of coronary heart disease using risk factor categories. *Circulation* **97**, 1837-47 (1998).
14. Rowe, A.K., Powell, K.E. & Flanders, W.D. Why population attributable fractions can sum to more than one. *Am J Prev Med* **26**, 243-9 (2004).
15. Weissglas-Volkov, D. & Pajukanta, P. Genetic causes of high and low serum HDL-cholesterol. *J Lipid Res* **51**, 2032-57 (2010).
16. Kathiresan, S. *et al.* Polymorphisms associated with cholesterol and risk of cardiovascular events. *N Engl J Med* **358**, 1240-9 (2008).
17. Hindorff, L.A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**, 9362-7 (2009).
18. Vattikuti, S., Guo, J. & Chow, C.C. Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits. *PLoS Genet* **8**, e1002637 (2012).

19. Lloyd-Jones, D.M. *et al.* Parental cardiovascular disease as a risk factor for cardiovascular disease in middle-aged adults: a prospective study of parents and offspring. *JAMA* **291**, 2204-11 (2004).
20. Schildkraut, J.M., Myers, R.H., Cupples, L.A., Kiely, D.K. & Kannel, W.B. Coronary risk associated with age and sex of parental heart disease in the Framingham Study. *Am J Cardiol* **64**, 555-9 (1989).
21. Marenberg, M.E., Risch, N., Berkman, L.F., Floderus, B. & de Faire, U. Genetic susceptibility to death from coronary heart disease in a study of twins. *N Engl J Med* **330**, 1041-6 (1994).
22. Regitz-Zagrosek, V. & Fleck, E. Heart failure: clinical aspects from basic research. Introduction. *Eur Heart J* **15 Suppl D**, 1 (1994).
23. Hernandez, A.V., Pasupuleti, V., Deshpande, A., Bernabe-Ortiz, A. & Miranda, J.J. Effect of rural-to-urban within-country migration on cardiovascular risk factors in low- and middle-income countries: a systematic review. *Heart* **98**, 185-94 (2012).
24. Ezzati, M., Lopez, A.D., Rodgers, A., Vander Hoorn, S. & Murray, C.J. Selected major risk factors and global and regional burden of disease. *Lancet* **360**, 1347-60 (2002).
25. Kim, A.S. & Johnston, S.C. Global variation in the relative burden of stroke and ischemic heart disease. *Circulation* **124**, 314-23 (2011).
26. Kohler, H.P. & Grant, P.J. Plasminogen-activator inhibitor type 1 and coronary artery disease. *N Engl J Med* **342**, 1792-801 (2000).
27. Aaronson, P.I., Ward, J.P.T. *The Cardiovascular System at a Glance*, (Blackwell Publishing, 2007).
28. Preston, R.R., Wilson, T.E. *Lippincott's Illustrated Reviews Physiology*, (Lippincott Williams and Wilkins, 2013).
29. Kumar, V., Abbas, A.K., Aster, J.C. *Robbins and Cotran Pathologic Basis of Disease*, (2014).
30. Gebbink, M.F. Tissue-type plasminogen activator-mediated plasminogen activation and contact activation, implications in and beyond haemostasis. *J Thromb Haemost* **9 Suppl 1**, 174-81 (2011).
31. Hoylaerts, M., Rijken, D.C., Lijnen, H.R. & Collen, D. Kinetics of the activation of plasminogen by human tissue plasminogen activator. Role of fibrin. *J Biol Chem* **257**, 2912-9 (1982).
32. Thorsen, S. The mechanism of plasminogen activation and the variability of the fibrin effector during tissue-type plasminogen activator-mediated fibrinolysis. *Ann N Y Acad Sci* **667**, 52-63 (1992).
33. Brogren, H. *et al.* Platelets synthesize large amounts of active plasminogen activator inhibitor 1. *Blood* **104**, 3943-8 (2004).
34. Keijer, J. *et al.* The interaction of plasminogen activator inhibitor 1 with plasminogen activators (tissue-type and urokinase-type) and fibrin: localization of interaction sites and physiologic relevance. *Blood* **78**, 401-9 (1991).
35. Devaraj, S., Xu, D.Y. & Jialal, I. C-reactive protein increases plasminogen activator inhibitor-1 expression and activity in human aortic endothelial cells: implications for the metabolic syndrome and atherothrombosis. *Circulation* **107**, 398-404 (2003).
36. Huber, K. Plasminogen activator inhibitor type-1 (part one): basic mechanisms, regulation, and role for thromboembolic disease. *J Thromb Thrombolysis* **11**, 183-93 (2001).

37. Loscalzo, J., Schaffer, A. *Thrombosis and Hemorrhage*, (Lippincott Williams and Wilkins, 2003).
38. Chandler, W.L., Jascur, M.L. & Henderson, P.J. Measurement of different forms of tissue plasminogen activator in plasma. *Clin Chem* **46**, 38-46 (2000).
39. Brotman, M.A. *et al.* Facial emotion labeling deficits in children and adolescents at risk for bipolar disorder. *Am J Psychiatry* **165**, 385-9 (2008).
40. Schneiderman, J. *et al.* Increased type 1 plasminogen activator inhibitor gene expression in atherosclerotic human arteries. *Proc Natl Acad Sci U S A* **89**, 6998-7002 (1992).
41. Smith, A. *et al.* Which hemostatic markers add to the predictive value of conventional risk factors for coronary heart disease and ischemic stroke? The Caerphilly Study. *Circulation* **112**, 3080-7 (2005).
42. Nordt, T.K., Peter, K., Ruef, J., Kubler, W. & Bode, C. Plasminogen activator inhibitor type-1 (PAI-1) and its role in cardiovascular disease. *Thromb Haemost* **82 Suppl 1**, 14-8 (1999).
43. Meade, T.W., Ruddock, V., Stirling, Y., Chakrabarti, R. & Miller, G.J. Fibrinolytic activity, clotting factors, and long-term incidence of ischaemic heart disease in the Northwick Park Heart Study. *Lancet* **342**, 1076-9 (1993).
44. Marcucci, R. *et al.* PAI-1 and homocysteine, but not lipoprotein (a) and thrombophilic polymorphisms, are independently associated with the occurrence of major adverse cardiac events after successful coronary stenting. *Heart* **92**, 377-81 (2006).
45. Juhan-Vague, I. *et al.* Plasma plasminogen activator inhibitor-1 in angina pectoris. Influence of plasma insulin and acute-phase response. *Arteriosclerosis* **9**, 362-7 (1989).
46. Gram, J., Bladbjerg, E.M., Moller, L., Sjol, A. & Jespersen, J. Tissue-type plasminogen activator and C-reactive protein in acute coronary heart disease. A nested case-control study. *J Intern Med* **247**, 205-12 (2000).
47. Pradhan, A.D. *et al.* Tissue plasminogen activator antigen and D-dimer as markers for atherothrombotic risk among healthy postmenopausal women. *Circulation* **110**, 292-300 (2004).
48. Gorog, D.A. Prognostic value of plasma fibrinolysis activation markers in cardiovascular disease. *J Am Coll Cardiol* **55**, 2701-9 (2010).
49. Kumar, D. & Elliott, P. *Principles and practice of clinical cardiovascular genetics*, xxii, 600 p. (Oxford University Press, Oxford ; New York, 2010).
50. Teslovich, T.M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707-13 (2010).
51. O'Donnell, C.J. & Nabel, E.G. Genomics of cardiovascular disease. *N Engl J Med* **365**, 2098-109 (2011).
52. McPherson, R. Chromosome 9p21.3 locus for coronary artery disease: how little we know. *J Am Coll Cardiol* **62**, 1382-3 (2013).
53. Lowe, G. & Rumley, A. The relevance of coagulation in cardiovascular disease: what do the biomarkers tell us? *Thromb Haemost* **112**, 860-7 (2014).
54. Newton-Cheh, C. & Hirschhorn, J.N. Genetic association studies of complex traits: design and analysis issues. *Mutat Res* **573**, 54-69 (2005).
55. de Lange, M., Snieder, H., Ariens, R.A., Spector, T.D. & Grant, P.J. The genetics of haemostasis: a twin study. *Lancet* **357**, 101-5 (2001).

56. Cesari, M., Sartori, M.T., Patrassi, G.M., Vettore, S. & Rossi, G.P. Determinants of plasma levels of plasminogen activator inhibitor-1 : A study of normotensive twins. *Arterioscler Thromb Vasc Biol* **19**, 316-20 (1999).
57. Huang, J. *et al.* Genome-wide association study for circulating levels of PAI-1 provides novel insights into its regulation. *Blood* **120**, 4873-81 (2012).
58. Huang, J. *et al.* Genome-wide association study for circulating tissue plasminogen activator levels and functional follow-up implicates endothelial STXBP5 and STX2. *Arterioscler Thromb Vasc Biol* **34**, 1093-101 (2014).
59. Rimoin, D.L., Pyeritz, R.E., Korf, B.R. & Emery, A.E.H. *Emery and Rimoin's essential medical genetics*, xix, 626 pages (Academic, Amsterdam ; Boston, 2013).
60. Greenhalgh, S., Montgomery, M., Segal, S.J. & Todaro, M.P. State of world population 2007: Unleashing the potential of urban growth. *Population and Development Review* **33**, 639-640 (2007).
61. Hawkes, C. Uneven dietary development: linking the policies and processes of globalization with the nutrition transition, obesity and diet-related chronic diseases. *Globalization and Health* **2**(2006).
62. Smith, R., McCready, T. & Yusuf, S. Combination therapy to prevent cardiovascular disease: slow progress. *JAMA* **309**, 1595-6 (2013).
63. Reddy, K.S. & Yusuf, S. Emerging epidemic of cardiovascular disease in developing countries. *Circulation* **97**, 596-601 (1998).
64. Yach, D., Hawkes, C., Gould, C.L. & Hofman, K.J. The global burden of chronic diseases: overcoming impediments to prevention and control. *JAMA* **291**, 2616-22 (2004).
65. Dalal, S. *et al.* Non-communicable diseases in sub-Saharan Africa: what we know now. *Int J Epidemiol* **40**, 885-901 (2011).
66. Holmes, M.D. *et al.* Non-communicable diseases in sub-Saharan Africa: the case for cohort studies. *PLoS Med* **7**, e1000244 (2010).
67. Pobee, J.O., Larbi, E.B., Belcher, D.W., Wurapa, F.K. & Dodu, S.R. Blood pressure distribution in a rural Ghanaian population. *Trans R Soc Trop Med Hyg* **71**, 66-72 (1977).
68. Poulter, N. *et al.* Blood pressure and associated factors in a rural Kenyan community. *Hypertension* **6**, 810-3 (1984).
69. Addo, J., Smeeth, L. & Leon, D.A. Hypertension in sub-saharan Africa: a systematic review. *Hypertension* **50**, 1012-8 (2007).
70. Mbanya, J.C., Motala, A.A., Sobngwi, E., Assah, F.K. & Enoru, S.T. Diabetes in sub-Saharan Africa. *Lancet* **375**, 2254-66 (2010).
71. Kengne, A.P., Echouffo-Tcheugui, J.B., Sobngwi, E. & Mbanya, J.C. New insights on diabetes mellitus and obesity in Africa-part 1: prevalence, pathogenesis and comorbidities. *Heart* **99**, 979-83 (2013).
72. Moran, A. *et al.* The epidemiology of cardiovascular diseases in sub-Saharan Africa: the Global Burden of Diseases, Injuries and Risk Factors 2010 Study. *Prog Cardiovasc Dis* **56**, 234-9 (2013).
73. Cooney, M.T., Dudina, A., D'Agostino, R. & Graham, I.M. Cardiovascular Risk-Estimation Systems in Primary Prevention Do They Differ? Do They Make a Difference? Can We See the Future? *Circulation* **122**, 300-310 (2010).
74. D'Agostino, R.B. *et al.* General cardiovascular risk profile for use in primary care - The Framingham Heart Study. *Circulation* **117**, 743-753 (2008).

75. Berry, J.D. *et al.* Lifetime Risks of Cardiovascular Disease. *New England Journal of Medicine* **366**, 321-329 (2012).
76. Bernabe-Ortiz, A., Benziger, C.P., Gilman, R.H., Smeeth, L. & Miranda, J.J. Sex differences in risk factors for cardiovascular disease: the PERU MIGRANT study. *PLoS One* **7**, e35127 (2012).
77. Zeba, A.N., Delisle, H.F., Renier, G., Savadogo, B. & Baya, B. The double burden of malnutrition and cardiometabolic risk widens the gender and socio-economic health gap: a study among adults in Burkina Faso (West Africa). *Public Health Nutr* **15**, 2210-9 (2012).
78. BeLue, R. *et al.* An overview of cardiovascular risk factor burden in sub-Saharan African countries: a socio-cultural perspective. *Global Health* **5**, 10 (2009).
79. Lenfant, C. *et al.* Seventh report of the Joint National Committee on the Prevention, Detection, Evaluation, and Treatment of High Blood Pressure (JNC 7): resetting the hypertension sails. *Hypertension* **41**, 1178-9 (2003).
80. Mancia, G. *et al.* 2007 Guidelines for the Management of Arterial Hypertension: The Task Force for the Management of Arterial Hypertension of the European Society of Hypertension (ESH) and of the European Society of Cardiology (ESC). *J Hypertens* **25**, 1105-87 (2007).
81. Gavin, J.R. *et al.* Report of the expert committee on the diagnosis and classification of diabetes mellitus. *Diabetes Care* **22**, S5-S19 (1999).
82. Genuth, S. *et al.* Follow-up report on the diagnosis of diabetes mellitus. *Diabetes Care* **26**, 3160-3167 (2003).
83. Definition and Diagnosis of Diabetes Mellitus and Intermediate Hyperglycemia. (ed. Organization., W.H.) (2006).
84. Kavey, R.E.W. *et al.* American Heart Association Guidelines for Primary Prevention of Atherosclerotic Cardiovascular Disease beginning in childhood. *Journal of Pediatrics* **142**, 368-372 (2003).
85. Assmann, G., Schulte, H. & von Eckardstein, A. Hypertriglyceridemia and elevated lipoprotein(a) are risk factors for major coronary events in middle-aged men. *Am J Cardiol* **77**, 1179-84 (1996).
86. Cullen, P. Evidence that triglycerides are an independent coronary heart disease risk factor. *Am J Cardiol* **86**, 943-9 (2000).
87. Yamamoto-Kimura, L. *et al.* Prevalence and interrelations of cardiovascular risk factors in urban and rural Mexican adolescents. *J Adolesc Health* **38**, 591-8 (2006).
88. Physical status: the use and interpretation of anthropometry. (World Health Organisation 1995).
89. Council, T.W.A.E. Basic Education Certificate Examination. (2015).
90. Scadding, H. Junior Secondary-Schools - an Educational Initiative in Ghana. *Compare-a Journal of Comparative Education* **19**, 43-48 (1989).
91. Schünemann HJ, O.A., Vist GE, Higgins JP, Deeks JJ, Glasziou P,, Guyatt GH, o.b.o.t.C.A.a.R. & Group, M. Interpreting results and drawing conclusions.
Cochrane Handbook for Systematic Reviews of Interventions,
Version 500 [updated February 2008]. (The Cochrane Collaboration, 2008).
92. al., A.O.e. Age Standardization of Rates: A New WHO Standard (Technical Report). . in *GPE Discussion Paper Series* (World Health Organization, Geneva, 2001).

93. Mathers CD, F.D., Boerma JT The Global Burden of Disease: 2004 Update. . (World Health Organization, Geneva, Switzerland, 2008).
94. Riley, L., and Melanie Cowan. "Noncommunicable Diseases Country Profiles 2014." (World Health Organization, Geneva, 2014).
95. Commodore-Mensah, Y., Samuel, L.J., Dennison-Himmelfarb, C.R. & Agyemang, C. Hypertension and overweight/obesity in Ghanaians and Nigerians living in West Africa and industrialized countries: a systematic review. *J Hypertens* **32**, 464-72 (2014).
96. Amoah, A.G., Owusu, S.K. & Adjei, S. Diabetes in Ghana: a community based prevalence study in Greater Accra. *Diabetes Res Clin Pract* **56**, 197-205 (2002).
97. Duda, R.B. *et al.* Results of the Women's Health Study of Accra: assessment of blood pressure in urban women. *Int J Cardiol* **117**, 115-22 (2007).
98. van der Sande, M.A. *et al.* Blood pressure patterns and cardiovascular risk factors in rural and urban gambian communities. *J Hum Hypertens* **14**, 489-96 (2000).
99. Addo, J., Amoah, A.G. & Koram, K.A. The changing patterns of hypertension in Ghana: a study of four rural communities in the Ga District. *Ethn Dis* **16**, 894-9 (2006).
100. Agyemang, C., Bruijnzeels, M.A. & Owusu-Dabo, E. Factors associated with hypertension awareness, treatment, and control in Ghana, West Africa. *J Hum Hypertens* **20**, 67-71 (2006).
101. Isezuo, S.A., Sabir, A.A., Ohwovorilole, A.E. & Fasanmade, O.A. Prevalence, associated factors and relationship between prehypertension and hypertension: a study of two ethnic African populations in Northern Nigeria. *J Hum Hypertens* **25**, 224-30 (2011).
102. Oladapo, O.O. *et al.* A prevalence of cardiometabolic risk factors among a rural Yoruba south-western Nigerian population: a population-based survey. *Cardiovasc J Afr* **21**, 26-31 (2010).
103. Hendriks, M.E. *et al.* Hypertension in sub-Saharan Africa: cross-sectional surveys in four rural and urban communities. *PLoS One* **7**, e32638 (2012).
104. Agyemang, C. Rural and urban differences in blood pressure and hypertension in Ghana, West Africa. *Public Health* **120**, 525-33 (2006).
105. Bosu, W.K. Epidemic of hypertension in Ghana: a systematic review. *BMC Public Health* **10**, 418 (2010).
106. Cappuccio, F.P. *et al.* Prevalence, detection, management, and control of hypertension in Ashanti, West Africa. *Hypertension* **43**, 1017-22 (2004).
107. Vaccarino, V. *et al.* Ischaemic heart disease in women: are there sex differences in pathophysiology and risk factors? Position paper from the working group on coronary pathophysiology and microcirculation of the European Society of Cardiology. *Cardiovasc Res* **90**, 9-17 (2011).
108. Mancia, G. *et al.* 2013 ESH/ESC Guidelines for the management of arterial hypertension: the Task Force for the management of arterial hypertension of the European Society of Hypertension (ESH) and of the European Society of Cardiology (ESC). *J Hypertens* **31**, 1281-357 (2013).
109. Danaei, G. *et al.* National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2.7 million participants. *Lancet* **378**, 31-40 (2011).
110. Emerging Risk Factors, C. *et al.* Lipoprotein(a) concentration and the risk of coronary heart disease, stroke, and nonvascular mortality. *JAMA* **302**, 412-23 (2009).

111. Shaw, J.E., Sicree, R.A. & Zimmet, P.Z. Global estimates of the prevalence of diabetes for 2010 and 2030. *Diabetes Res Clin Pract* **87**, 4-14 (2010).
112. Dodu, S.R. The incidence of diabetes mellitus in Accra (Ghana); a study of 4,000 patients. *West Afr Med J* **7**, 129-34 (1958).
,113. Dodu, S.R.A.d.H., N. A Diabetes Case-Finding Survey in Ho, Ghana. *Ghana Med.J.*, 75-80 (1964).
114. Kengne, A.P., Sobngwi, E., Echouffo-Tcheugui, J.B. & Mbanya, J.C. New insights on diabetes mellitus and obesity in Africa-Part 2: prevention, screening and economic burden. *Heart* **99**, 1072-7 (2013).
115. Nicholls, S.J. *et al.* Relationship between cardiovascular risk factors and atherosclerotic disease burden measured by intravascular ultrasound. *J Am Coll Cardiol* **47**, 1967-75 (2006).
116. Nicholls, S.J. *et al.* Rate of progression of coronary atherosclerotic plaque in women. *J Am Coll Cardiol* **49**, 1546-51 (2007).
117. Schoenhard, J.A. *et al.* Male-female differences in the genetic regulation of t-PA and PAI-1 levels in a Ghanaian population. *Hum Genet* **124**, 479-88 (2008).
118. Regitz-Zagrosek, V., Brokat, S. & Tschope, C. Role of gender in heart failure with normal left ventricular ejection fraction. *Prog Cardiovasc Dis* **49**, 241-51 (2007).
119. Hill, A.G. *et al.* Health of urban Ghanaian women as identified by the Women's Health Study of Accra. *Int J Gynaecol Obstet* **99**, 150-6 (2007).
120. Renzaho, A.M. Fat, rich and beautiful: changing socio-cultural paradigms associated with obesity risk, nutritional status and refugee children from sub-Saharan Africa. *Health Place* **10**, 105-13 (2004).
121. Finucane, M.M. *et al.* National, regional, and global trends in body-mass index since 1980: systematic analysis of health examination surveys and epidemiological studies with 960 country-years and 9.1 million participants. *Lancet* **377**, 557-67 (2011).
122. Hitimana, L., Heinrigs, P., and Tremolieres, M. West African Futures. Vol. 1 1-8 (Sahel and West Africa Club Secretariat (SWAC/OECD)).
123. Bosu, W.K. An overview of the nutrition transition in West Africa: implications for non-communicable diseases. *Proc Nutr Soc*, 1-12 (2014).
124. van der Sande, M.A. *et al.* Obesity and undernutrition and cardiovascular risk factors in rural and urban Gambian communities. *Am J Public Health* **91**, 1641-4 (2001).
125. Msyamboza, K.P. *et al.* The burden of selected chronic non-communicable diseases and their risk factors in Malawi: nationwide STEPS survey. *PLoS One* **6**, e20316 (2011).
126. Farzadfar, F. *et al.* National, regional, and global trends in serum total cholesterol since 1980: systematic analysis of health examination surveys and epidemiological studies with 321 country-years and 3.0 million participants. *Lancet* **377**, 578-86 (2011).
127. Vorster, H.H. The emergence of cardiovascular disease during urbanisation of Africans. *Public Health Nutr* **5**, 239-43 (2002).
128. World urbanization prospects: the 2007 revision. (UN Department of Economic and Social Affairs, Population Division, 2008).
129. Witztum, J.L. & Steinberg, D. Role of oxidized low density lipoprotein in atherogenesis. *J Clin Invest* **88**, 1785-92 (1991).
130. Ben-Yehuda, O. & DeMaria, A.N. LDL-cholesterol targets after the ACC/AHA 2013 guidelines: evidence that lower is better? *J Am Coll Cardiol* **64**, 495-7 (2014).

131. Ridker, P.M. *et al.* Rosuvastatin to prevent vascular events in men and women with elevated C-reactive protein. *N Engl J Med* **359**, 2195-207 (2008).
132. Martin, S.S., Blumenthal, R.S. & Miller, M. LDL cholesterol: the lower the better. *Med Clin North Am* **96**, 13-26 (2012).
133. Yajnik, C.S. *et al.* Conventional and novel cardiovascular risk factors and markers of vascular damage in rural and urban Indian men. *Int J Cardiol* **165**, 255-259 (2013).
134. Glew, R.H. *et al.* Risk factors for cardiovascular disease and diet of urban and rural dwellers in northern Nigeria. *J Health Popul Nutr* **22**, 357-69 (2004).
135. Miranda, J.J., Gilman, R.H. & Smeeth, L. Differences in cardiovascular risk factors in rural, urban and rural-to-urban migrants in Peru. *Heart* **97**, 787-96 (2011).
136. Torun, B. *et al.* Rural-to-urban migration and cardiovascular disease risk factors in young Guatemalan adults. *Int J Epidemiol* **31**, 218-26 (2002).
137. Wong, N.D. Epidemiological studies of CHD and the evolution of preventive cardiology. *Nat Rev Cardiol* **11**, 276-89 (2014).
138. Morange, P.E. & Alessi, M.C. Thrombosis in central obesity and metabolic syndrome: mechanisms and epidemiology. *Thromb Haemost* **110**, 669-80 (2013).
139. Cohen, B. Urban growth in developing countries: A review of current trends and a caution regarding existing forecasts. *World Development* **32**, 23-51 (2004).
140. Grundy, S.M. *et al.* Diagnosis and management of the metabolic syndrome: an American Heart Association/National Heart, Lung, and Blood Institute Scientific Statement. *Circulation* **112**, 2735-52 (2005).
141. Alberti, K.G. *et al.* Harmonizing the metabolic syndrome: a joint interim statement of the International Diabetes Federation Task Force on Epidemiology and Prevention; National Heart, Lung, and Blood Institute; American Heart Association; World Heart Federation; International Atherosclerosis Society; and International Association for the Study of Obesity. *Circulation* **120**, 1640-5 (2009).
142. Grundy, S.M. Metabolic syndrome: a multiplex cardiovascular risk factor. *J Clin Endocrinol Metab* **92**, 399-404 (2007).
143. Eckel, R.H., Alberti, K.G., Grundy, S.M. & Zimmet, P.Z. The metabolic syndrome. *Lancet* **375**, 181-3 (2010).
144. Kahn, R. *et al.* The metabolic syndrome: time for a critical appraisal: joint statement from the American Diabetes Association and the European Association for the Study of Diabetes. *Diabetes Care* **28**, 2289-304 (2005).
145. Grundy, S.M. Does the metabolic syndrome exist? *Diabetes Care* **29**, 1689-92; discussion 1693-6 (2006).
146. Wen, C.P. *et al.* Attributable mortality burden of metabolic syndrome: comparison with its individual components. *Eur J Cardiovasc Prev Rehabil* **18**, 561-73 (2011).
147. Samaras, K. *et al.* The value of the metabolic syndrome concept in elderly adults: is it worth less than the sum of its parts? *J Am Geriatr Soc* **60**, 1734-41 (2012).
148. Godsland, I.F., Lecamwasam, K. & Johnston, D.G. A systematic evaluation of the insulin resistance syndrome as an independent risk factor for cardiovascular disease mortality and derivation of a clinical index. *Metabolism* **60**, 1442-8 (2011).
149. Kraja, A.T. *et al.* Pleiotropic genes for metabolic syndrome and inflammation. *Mol Genet Metab* **112**, 317-38 (2014).
150. Zappulla, D. Environmental stress, erythrocyte dysfunctions, inflammation, and the metabolic syndrome: adaptations to CO2 increases? *J Cardiometab Syndr* **3**, 30-4 (2008).

151. Karlsson, B., Knutsson, A. & Lindahl, B. Is there an association between shift work and having a metabolic syndrome? Results from a population based study of 27,485 people. *Occup Environ Med* **58**, 747-52 (2001).
152. Mertens, I. *et al.* Among inflammation and coagulation markers, PAI-1 is a true component of the metabolic syndrome. *Int J Obes (Lond)* **30**, 1308-14 (2006).
153. Nieuwdorp, M., Stroes, E.S., Meijers, J.C. & Buller, H. Hypercoagulability in the metabolic syndrome. *Curr Opin Pharmacol* **5**, 155-9 (2005).
154. Scheer, F.A. & Shea, S.A. Human circadian system causes a morning peak in prothrombotic plasminogen activator inhibitor-1 (PAI-1) independent of the sleep/wake cycle. *Blood* **123**, 590-3 (2014).
155. Sobel, B.E. Increased plasminogen activator inhibitor-1 and vasculopathy. A reconcilable paradox. *Circulation* **99**, 2496-8 (1999).
156. Alessi, M.C. & Juhan-Vague, I. Contribution of PAI-1 in cardiovascular pathology. *Arch Mal Coeur Vaiss* **97**, 673-8 (2004).
157. Alessi, M.C. & Juhan-Vague, I. PAI-1 and the metabolic syndrome: links, causes, and consequences. *Arterioscler Thromb Vasc Biol* **26**, 2200-7 (2006).
158. De Taeye, B., Smith, L.H. & Vaughan, D.E. Plasminogen activator inhibitor-1: a common denominator in obesity, diabetes and cardiovascular disease. *Curr Opin Pharmacol* **5**, 149-54 (2005).
159. Iwaki, T., Urano, T. & Umemura, K. PAI-1, progress in understanding the clinical problem and its aetiology. *Br J Haematol* **157**, 291-8 (2012).
160. Kjoller, L. *et al.* Plasminogen activator inhibitor-1 represses integrin- and vitronectin-mediated cell migration independently of its function as an inhibitor of plasminogen activation. *Exp Cell Res* **232**, 420-9 (1997).
161. Ma, L.J. *et al.* Prevention of obesity and insulin resistance in mice lacking plasminogen activator inhibitor 1. *Diabetes* **53**, 336-46 (2004).
162. Mertens, I. & Van Gaal, L.F. Obesity, haemostasis and the fibrinolytic system. *Obes Rev* **3**, 85-101 (2002).
163. Lopez-Alemany, R., Redondo, J.M., Nagamine, Y. & Munoz-Canoves, P. Plasminogen activator inhibitor type-1 inhibits insulin signaling by competing with alphavbeta3 integrin for vitronectin binding. *Eur J Biochem* **270**, 814-21 (2003).
164. Tamura, Y. *et al.* Plasminogen activator inhibitor-1 deficiency ameliorates insulin resistance and hyperlipidemia but not bone loss in obese female mice. *Endocrinology* **155**, 1708-17 (2014).
165. Caglayan, E., Blaschke, F., Takata, Y. & Hsueh, W.A. Metabolic syndrome-interdependence of the cardiovascular and metabolic pathways. *Curr Opin Pharmacol* **5**, 135-42 (2005).
166. Dichtl, W. *et al.* In vivo stimulation of vascular plasminogen activator inhibitor-1 production by very low-density lipoprotein involves transcription factor binding to a VLDL-responsive element. *Thromb Haemost* **84**, 706-11 (2000).
167. Brown, N.J., Agirbasli, M.A., Williams, G.H., Litchfield, W.R. & Vaughan, D.E. Effect of activation and inhibition of the renin-angiotensin system on plasma PAI-1. *Hypertension* **32**, 965-71 (1998).
168. Margaglione, M. *et al.* PAI-1 plasma levels in a general population without clinical evidence of atherosclerosis: relation to environmental and genetic determinants. *Arterioscler Thromb Vasc Biol* **18**, 562-7 (1998).

169. Asselbergs, F.W. *et al.* Gender-specific correlations of plasminogen activator inhibitor-1 and tissue plasminogen activator levels with cardiovascular disease-related traits. *J Thromb Haemost* **5**, 313-20 (2007).
170. Hawkes, C. Uneven dietary development: linking the policies and processes of globalization with the nutrition transition, obesity and diet-related chronic diseases. *Global Health* **2**, 4 (2006).
171. Grundy, S.M. *et al.* A summary of implications of recent clinical trials for the National Cholesterol Education Program Adult Treatment Panel III guidelines. *Arteriosclerosis Thrombosis and Vascular Biology* **24**, 1329-1330 (2004).
172. Hsu, C.H. *et al.* Mean arterial pressure is better at predicting future metabolic syndrome in the normotensive elderly: A prospective cohort study in Taiwan. *Preventive Medicine* **72**, 76-82 (2015).
173. Edgell, S.E. & Noon, S.M. Effect of Violation of Normality on the T-Test of the Correlation-Coefficient. *Psychological Bulletin* **95**, 576-583 (1984).
174. Kowalski, C.J. Effects of Non-Normality on Distribution of Sample Product-Moment Correlation Coefficient. *Journal of the Royal Statistical Society Series C-Applied Statistics* **21**, 1-& (1972).
175. Reinsch, C.H. Smoothing by Spline Functions. *Numerische Mathematik* **10**, 177-& (1967).
176. Popkin, B.M. & Doak, C.M. The obesity epidemic is a worldwide phenomenon. *Nutr Rev* **56**, 106-14 (1998).
177. Ruderman, N., Chisholm, D., Pi-Sunyer, X. & Schneider, S. The metabolically obese, normal-weight individual revisited. *Diabetes* **47**, 699-713 (1998).
178. Paterson, J.M. *et al.* Metabolic syndrome without obesity: Hepatic overexpression of 11 beta-hydroxysteroid dehydrogenase type 1 in transgenic mice. *Proc Natl Acad Sci U S A* **101**, 7088-7093 (2004).
179. Lear, S.A., Kohli, S., Bondy, G.P., Tchernof, A. & Sniderman, A.D. Ethnic Variation in Fat and Lean Body Mass and the Association with Insulin Resistance. *Journal of Clinical Endocrinology & Metabolism* **94**, 4696-4702 (2009).
180. Despres, J.P. *et al.* Abdominal obesity and the metabolic syndrome: Contribution to global cardiometabolic risk. *Arteriosclerosis Thrombosis and Vascular Biology* **28**, 1039-1049 (2008).
181. Reilly, M.P. & Rader, D.J. The Metabolic Syndrome: More than the Sum of Its Parts? . *Circulation* **108**, 1546-1551 (2003).
182. Lusis, A.J., Attie, A.D. & Reue, K. Metabolic syndrome: from epidemiology to systems biology. *Nat Rev Genet* **9**, 819-30 (2008).
183. Kahn, R. The metabolic syndrome (emperor) wears no clothes. *Diabetes Care* **29**, 1693-1696 (2006).
184. Sattar, N. *et al.* Can metabolic syndrome usefully predict cardiovascular disease and diabetes? Outcome data from two prospective studies. *Lancet* **371**, 1927-1935 (2008).
185. White, M.J. *et al.* Genetics of Plasminogen Activator Inhibitor-1 (PAI-1) in a Ghanaian population. *PLoS One* (2015).
186. Brage, S. *et al.* Features of the metabolic syndrome are associated with objectively measured physical activity and fitness in Danish children - the European Youth Heart Study (EYHS). *Diabetes Care* **27**, 2141-2148 (2004).

187. Franks, P.W., Ekelund, U., Brage, S., Wong, M.Y. & Wareham, N.J. Does the association of habitual physical activity with the metabolic syndrome differ by level of cardiorespiratory fitness? *Diabetes Care* **27**, 1187-1193 (2004).
188. Forouhi, N.G., Luan, J., Cooper, A., Boucher, B.J. & Wareham, N.J. Baseline serum 25-hydroxy vitamin D is predictive of future glycemic status and insulin resistance - The Medical Research Council Ely Prospective Study 1990-2000. *Diabetes* **57**, 2619-2625 (2008).
189. Beekman, M. *et al.* Heritabilities of apolipoprotein and lipid levels in three countries. *Twin Res* **5**, 87-97 (2002).
190. Souren, N.Y. *et al.* Anthropometry, carbohydrate and lipid metabolism in the East Flanders Prospective Twin Survey: heritabilities. *Diabetologia* **50**, 2107-16 (2007).
191. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**, 565-9 (2010).
192. Eichler, E.E. *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* **11**, 446-50 (2010).
193. Mackay, T.F. *et al.* The Drosophila melanogaster Genetic Reference Panel. *Nature* **482**, 173-8 (2012).
194. Huang, W. *et al.* Epistasis dominates the genetic architecture of Drosophila quantitative traits. *Proc Natl Acad Sci U S A* **109**, 15553-9 (2012).
195. Taylor, M.B. & Ehrenreich, I.M. Higher-order genetic interactions and their contribution to complex traits. *Trends Genet* **31**, 34-40 (2015).
196. Ehrenreich, I.M. *et al.* Dissection of genetically complex traits with extremely large pools of yeast segregants. *Nature* **464**, 1039-42 (2010).
197. Bloom, J.S., Ehrenreich, I.M., Loo, W.T., Lite, T.L. & Kruglyak, L. Finding the sources of missing heritability in a yeast cross. *Nature* **494**, 234-7 (2013).
198. Deutschbauer, A.M. & Davis, R.W. Quantitative trait loci mapped to single-nucleotide resolution in yeast. *Nat Genet* **37**, 1333-40 (2005).
199. Gerke, J., Lorenz, K., Ramnarine, S. & Cohen, B. Gene-environment interactions at nucleotide resolution. *PLoS Genet* **6**, e1001144 (2010).
200. Weiss, L.A., Pan, L., Abney, M. & Ober, C. The sex-specific genetic architecture of quantitative traits in humans. *Nat Genet* **38**, 218-22 (2006).
201. Ober, C., Loisel, D.A. & Gilad, Y. Sex-specific genetic architecture of human disease. *Nat Rev Genet* **9**, 911-22 (2008).
202. Gilks, W.P., Abbott, J.K. & Morrow, E.H. Sex differences in disease genetics: evidence, evolution, and detection. *Trends Genet* **30**, 453-63 (2014).
203. Aschard, H., Vilhjalmsson, B.J., Joshi, A.D., Price, A.L. & Kraft, P. Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. *Am J Hum Genet* **96**, 329-39 (2015).
204. Manning, A.K. *et al.* A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat Genet* **44**, 659-69 (2012).
205. Briollais, L. & Durrieu, G. Application of quantile regression to recent genetic and -omic studies. *Hum Genet* **133**, 951-66 (2014).
206. Williams, P.T. Quantile-specific penetrance of genes affecting lipoproteins, adiposity and height. *PLoS One* **7**, e28764 (2012).

207. Mitchell, J.A., Hakonarson, H., Rebbeck, T.R. & Grant, S.F. Obesity-susceptibility loci and the tails of the pediatric BMI distribution. *Obesity (Silver Spring)* **21**, 1256-60 (2013).
208. Cordell, H.J. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet* **10**, 392-404 (2009).
209. Hill, W.G., Goddard, M.E. & Visscher, P.M. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet* **4**, e1000008 (2008).
210. Khoury, M.J. & Wacholder, S. Invited commentary: from genome-wide association studies to gene-environment-wide interaction studies--challenges and opportunities. *Am J Epidemiol* **169**, 227-30; discussion 234-5 (2009).
211. Murcray, C.E., Lewinger, J.P. & Gauderman, W.J. Gene-environment interaction in genome-wide association studies. *Am J Epidemiol* **169**, 219-26 (2009).
212. Andreasen, C.H. *et al.* Low physical activity accentuates the effect of the FTO rs9939609 polymorphism on body fat accumulation. *Diabetes* **57**, 95-101 (2008).
213. Frayling, T.M. *et al.* A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316**, 889-94 (2007).
214. Scuteri, A. *et al.* Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLoS Genet* **3**, e115 (2007).
215. Wei, W.H., Hemani, G. & Haley, C.S. Detecting epistasis in human complex traits. *Nat Rev Genet* **15**, 722-33 (2014).
216. Aliev, F., Latendresse, S.J., Bacanu, S.A., Neale, M.C. & Dick, D.M. Testing for Measured Gene-Environment Interaction: Problems with the use of Cross-Product Terms and a Regression Model Reparameterization Solution. *Behavior Genetics* **44**, 165-181 (2014).
217. Kodaman, N. *et al.* Human and Helicobacter pylori coevolution shapes the risk of gastric disease. *Proc Natl Acad Sci U S A* **111**, 1455-60 (2014).
218. Solovieff, N., Cotsapas, C., Lee, P.H., Purcell, S.M. & Smoller, J.W. Pleiotropy in complex traits: challenges and strategies. *Nat Rev Genet* **14**, 483-95 (2013).
219. Sivakumaran, S. *et al.* Abundant pleiotropy in human complex diseases and traits. *Am J Hum Genet* **89**, 607-18 (2011).
220. Cotsapas, C. *et al.* Pervasive sharing of genetic effects in autoimmune disease. *PLoS Genet* **7**, e1002254 (2011).
221. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. Genome-wide complex trait analysis (GCTA): methods, data analyses, and interpretations. *Methods Mol Biol* **1019**, 215-36 (2013).
222. Visscher, P.M. *et al.* Statistical power to detect genetic (co)variance of complex traits using SNP data in unrelated samples. *PLoS Genet* **10**, e1004269 (2014).
223. Ferreira, M.A. & Purcell, S.M. A multivariate test of association. *Bioinformatics* **25**, 132-3 (2009).
224. Zhou, X. & Stephens, M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat Methods* **11**, 407-9 (2014).
225. Aschard, H. *et al.* Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. *Am J Hum Genet* **94**, 662-76 (2014).
226. O'Reilly, P.F. *et al.* MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. *PLoS One* **7**, e34861 (2012).

227. Galesloot, T.E., van Steen, K., Kiemeney, L.A., Janss, L.L. & Vermeulen, S.H. A comparison of multivariate genome-wide association methods. *PLoS One* **9**, e95923 (2014).
228. Tang, C.S. & Ferreira, M.A. A gene-based test of association using canonical correlation analysis. *Bioinformatics* **28**, 845-50 (2012).
229. Allison, D.B. *et al.* Multiple phenotype modeling in gene-mapping studies of quantitative traits: power advantages. *Am J Hum Genet* **63**, 1190-201 (1998).
230. Global Lipids Genetics, C. *et al.* Discovery and refinement of loci associated with lipid levels. *Nat Genet* **45**, 1274-83 (2013).
231. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).
232. Musunuru, K. *et al.* From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466**, 714-U2 (2010).
233. Zeggini, E. *et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet* **40**, 638-45 (2008).
234. Rodriguez, S. *et al.* Haplotypic analyses of the IGF2-INS-TH gene cluster in relation to cardiovascular risk traits. *Hum Mol Genet* **13**, 715-25 (2004).
235. Yang, Q., Kathiresan, S., Lin, J.P., Tofler, G.H. & O'Donnell, C.J. Genome-wide association and linkage analyses of hemostatic factors and hematological phenotypes in the Framingham Heart Study. *Bmc Medical Genetics* **8 Suppl 1**, S12 (2007).
236. Melzer, D. *et al.* A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genet* **4**, e1000072 (2008).
237. Yu, B. *et al.* Association of genome-wide variation with highly sensitive cardiac troponin-T levels in European Americans and Blacks: a meta-analysis from atherosclerosis risk in communities and cardiovascular health studies. *Circ Cardiovasc Genet* **6**, 82-8 (2013).
238. Marroni, F. *et al.* A genome-wide association scan of RR and QT interval duration in 3 European genetically isolated populations: the EUROSPAN project. *Circ Cardiovasc Genet* **2**, 322-8 (2009).
239. Egecioglu, E. *et al.* Growth hormone receptor deficiency in mice results in reduced systolic blood pressure and plasma renin, increased aortic eNOS expression, and altered cardiovascular structure and function. *American Journal of Physiology-Endocrinology and Metabolism* **292**, E1418-E1425 (2007).
240. Schoeps, A. *et al.* Identification of new genetic susceptibility loci for breast cancer through consideration of gene-environment interactions. *Genet Epidemiol* **38**, 84-93 (2014).
241. Figueiredo, J.C. *et al.* Genome-wide diet-gene interaction analyses for risk of colorectal cancer. *PLoS Genet* **10**, e1004228 (2014).
242. Almli, L.M. *et al.* Correcting systematic inflation in genetic association tests that consider interaction effects: application to a genome-wide association study of posttraumatic stress disorder. *JAMA Psychiatry* **71**, 1392-9 (2014).
243. Voorman, A., Lumley, T., McKnight, B. & Rice, K. Behavior of QQ-plots and genomic control in studies of gene-environment interaction. *PLoS One* **6**, e19416 (2011).
244. Ingelsson, E. *et al.* Clinical utility of different lipid measures for prediction of coronary heart disease in men and women. *JAMA* **298**, 776-85 (2007).

245. Lawson, H.A. *et al.* Genetic Effects at Pleiotropic Loci Are Context-Dependent with Consequences for the Maintenance of Genetic Variation in Populations. *PLoS Genet* **7**(2011).
246. Kathiresan, S. *et al.* Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans (vol 40, pg 189, 2008). *Nat Genet* **40**, 1384-1384 (2008).
247. Sandhu, M.S. *et al.* LDL-cholesterol concentrations: a genome-wide association study. *Lancet* **371**, 483-491 (2008).
248. Wallace, C. *et al.* Genome-wide association study identifies genes for biomarkers of cardiovascular disease: Serum urate and dyslipiclemia. *Am J Hum Genet* **82**, 139-149 (2008).
249. Kathiresan, S. *et al.* Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat Genet* **41**, 56-65 (2009).
250. Nakayama, K., Bayasgalan, T. & Yamanaka, K. Large scale replication analysis of loci associated with lipid concentrations in a Japanese population (vol 46, pg 370, 2009). *Journal of Medical Genetics* **46**, 861-861 (2009).
251. Muendlein, A. *et al.* Significant impact of chromosomal locus 1p13.3 on serum LDL cholesterol and on angiographically characterized coronary atherosclerosis. *Atherosclerosis* **206**, 494-499 (2009).
252. Angelakopoulou, A. *et al.* Comparative analysis of genome-wide association studies signals for lipids, diabetes, and coronary heart disease: Cardiovascular Biomarker Genetics Collaboration. *Eur Heart J* **33**, 393-407 (2012).
253. Adeyemo, A. *et al.* Transferability and Fine Mapping of genome-wide associated loci for lipids in African Americans. *Bmc Medical Genetics* **13**(2012).
254. Samani, N.J. *et al.* Genomewide association analysis of coronary artery disease. *New England Journal of Medicine* **357**, 443-453 (2007).
255. Arvind, P., Nair, J., Jambunathan, S., Kakkar, V.V. & Shanker, J. CELSR2-PSRC1-SORT1 gene expression and association with coronary artery disease and plasma lipid levels in an Asian Indian cohort. *Journal of Cardiology* **64**, 339-346 (2014).
256. Lee, J.Y. *et al.* A genome-wide association study of a coronary artery disease risk variant. *Journal of Human Genetics* **58**, 120-126 (2013).
257. Willer, C.J. *et al.* Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet* **40**, 161-169 (2008).
258. Kathiresan, S. *et al.* Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat Genet* **41**, 334-341 (2009).
259. Wang, A.Z., Li, L., Zhang, B., Shen, G.Q. & Wang, Q.K. Association of SNP rs17465637 on Chromosome 1q41 and rs599839 on 1p13.3 with Myocardial Infarction in an American Caucasian Population. *Annals of Human Genetics* **75**, 475-482 (2011).
260. O'Donnell, C.J. *et al.* Genome-Wide Association Study for Coronary Artery Calcification With Follow-Up in Myocardial Infarction. *Circulation* **124**, 2855-U255 (2011).
261. Bauer, R.C., Stylianou, I.M. & Rader, D.J. Functional validation of new pathways in lipoprotein metabolism identified by human genetics. *Curr Opin Lipidol* **22**, 123-128 (2011).
262. Gustafsen, C. *et al.* The Hypercholesterolemia-Risk Gene SORT1 Facilitates PCSK9 Secretion. *Cell Metabolism* **19**, 310-318 (2014).

263. Strong, A. & Rader, D.J. Sortilin as a Regulator of Lipoprotein Metabolism. *Current Atherosclerosis Reports* **14**, 211-218 (2012).
264. Kjolby, M. *et al.* Sort1, Encoded by the Cardiovascular Risk Locus 1p13.3, Is a Regulator of Hepatic Lipoprotein Export. *Cell Metabolism* **12**, 213-223 (2010).
265. Linsel-Nitschke, P. *et al.* Genetic variation at chromosome 1p13.3 affects sortilin mRNA expression, cellular LDL-uptake and serum LDL levels which translates to the risk of coronary artery disease. *Atherosclerosis* **208**, 183-189 (2010).
266. Strong, A. *et al.* Hepatic sortilin regulates both apolipoprotein B secretion and LDL catabolism. *J Clin Invest* **122**, 2807-16 (2012).
267. Westerterp, M. & Tall, A.R. SORTILIN: many headed hydra. *Circ Res* **116**, 764-6 (2015).
268. Lusis, A.J. & Pajukanta, P. A treasure trove for lipoprotein biology. *Nat Genet* **40**, 129-130 (2008).
269. Oldoni, F., Sinke, R.J. & Kuivenhoven, J.A. Mendelian disorders of high-density lipoprotein metabolism. *Circ Res* **114**, 124-42 (2014).
270. Kraja, A.T. *et al.* A bivariate genome-wide approach to metabolic syndrome: STAMPEED consortium. *Diabetes* **60**, 1329-39 (2011).
271. Murakami, T. *et al.* Triglycerides are major determinants of cholesterol esterification/transfer and HDL remodeling in human plasma. *Arterioscler Thromb Vasc Biol* **15**, 1819-28 (1995).
272. Greene, D.J., Skeggs, J.W. & Morton, R.E. Elevated triglyceride content diminishes the capacity of high density lipoprotein to deliver cholesteryl esters via the scavenger receptor class B type I (SR-BI). *J Biol Chem* **276**, 4804-11 (2001).
273. Waterworth, D.M. *et al.* Genetic variants influencing circulating lipid levels and risk of coronary artery disease. *Arterioscler Thromb Vasc Biol* **30**, 2264-76 (2010).
274. Kuivenhoven, J.A. & Groen, A.K. Beyond the genetics of HDL: why is HDL cholesterol inversely related to cardiovascular disease? *Handb Exp Pharmacol* **224**, 285-300 (2015).
275. Klimentidis, Y.C., Chougule, A., Arora, A., Frazier-Wood, A.C. & Hsu, C.H. Triglyceride-Increasing Alleles Associated with Protection against Type-2 Diabetes. *PLoS Genet* **11**, e1005204 (2015).
276. Pendergrass, S.A. *et al.* The use of phenome-wide association studies (PheWAS) for exploration of novel genotype-phenotype relationships and pleiotropy discovery. *Genet Epidemiol* **35**, 410-22 (2011).
277. Global status report on non-communicable diseases. (ed. Alawan, A.) (World Health Organization, 2011).
278. Kawasaki, T., Dewerchin, M., Lijnen, H.R., Vermylen, J. & Hoylaerts, M.F. Vascular release of plasminogen activator inhibitor-1 impairs fibrinolysis during acute arterial thrombosis in mice. *Blood* **96**, 153-60 (2000).
279. Benyamin, B. *et al.* Variants in TF and HFE explain approximately 40% of genetic variation in serum-transferrin levels. *Am J Hum Genet* **84**, 60-5 (2009).
280. Asselbergs, F.W. *et al.* Genetic architecture of tissue-type plasminogen activator and plasminogen activator inhibitor-1. *Semin Thromb Hemost* **34**, 562-8 (2008).
281. Tsantes, A.E. *et al.* Association between the plasminogen activator inhibitor-1 4G/5G polymorphism and venous thrombosis. A meta-analysis. *Thromb Haemost* **97**, 907-13 (2007).

282. Maes, H.H., Neale, M.C. & Eaves, L.J. Genetic and environmental factors in relative body weight and human adiposity. *Behavior Genetics* **27**, 325-51 (1997).
283. Schousboe, K. *et al.* Twin study of genetic and environmental influences on adult body size, shape, and composition. *Int J Obes Relat Metab Disord* **28**, 39-48 (2004).
284. Hasselbalch, A.L. Genetics of dietary habits and obesity - a twin study. *Dan Med Bull* **57**, B4182 (2010).
285. van den Berg, L. *et al.* Heritability of dietary food intake patterns. *Acta Diabetol* **50**, 721-6 (2013).
286. Sakkinen, P.A. *et al.* Analytical and biologic variability in measures of hemostasis, fibrinolysis, and inflammation: assessment and implications for epidemiology. *Am J Epidemiol* **149**, 261-7 (1999).
287. McClellan, J. & King, M.C. Genetic heterogeneity in human disease. *Cell* **141**, 210-7 (2010).
288. Nilsson, L. *et al.* VLDL activation of plasminogen activator inhibitor-1 (PAI-1) expression: involvement of the VLDL receptor. *J Lipid Res* **40**, 913-9 (1999).
289. Reilly, S.L., Ferrell, R.E. & Sing, C.F. The gender-specific apolipoprotein E genotype influence on the distribution of plasma lipids and apolipoproteins in the population of Rochester, MN. III. Correlations and covariances. *Am J Hum Genet* **55**, 1001-18 (1994).
290. Maxwell, T.J. *et al.* APOE modulates the correlation between triglycerides, cholesterol, and CHD through pleiotropy, and gene-by-gene interactions. *Genetics* **195**, 1397-405 (2013).
291. Yang, J. *et al.* FTO genotype is associated with phenotypic variability of body mass index. *Nature* **490**, 267-72 (2012).
292. Shen, X., Pettersson, M., Ronnegard, L. & Carlborg, O. Inheritance beyond plain heritability: variance-controlling genes in Arabidopsis thaliana. *PLoS Genet* **8**, e1002839 (2012).
293. Reilly, S.L., Ferrell, R.E., Kottke, B.A. & Sing, C.F. The gender-specific apolipoprotein E genotype influence on the distribution of plasma lipids and apolipoproteins in the population of Rochester, Minnesota. II. Regression relationships with concomitants. *Am J Hum Genet* **51**, 1311-24 (1992).
294. Curia, M.C. *et al.* Increased variance in germline allele-specific expression of APC associates with colorectal cancer. *Gastroenterology* **142**, 71-77 e1 (2012).
295. Gibson, G. & Wagner, G. Canalization in evolutionary genetics: a stabilizing theory? *Bioessays* **22**, 372-80 (2000).
296. Gibson, G. Decanalization and the origin of complex disease. *Nat Rev Genet* **10**, 134-40 (2009).
297. Hughes, T.R. *et al.* Functional discovery via a compendium of expression profiles. *Cell* **102**, 109-26 (2000).
298. Moore, J.H. & Williams, S.M. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *Bioessays* **27**, 637-46 (2005).
299. Jimenez-Gomez, J.M., Corwin, J.A., Joseph, B., Maloof, J.N. & Kliebenstein, D.J. Genomic analysis of QTLs and genes altering natural variation in stochastic noise. *PLoS Genet* **7**, e1002295 (2011).

300. Xu, Z. & Taylor, J.A. SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. *Nucleic Acids Res* **37**, W600-5 (2009).
301. Fieller, E.C., Hartley, H.O. & Pearson, E.S. Tests for Rank Correlation Coefficients .1. *Biometrika* **44**, 470-481 (1957).
302. Sobota, R.S. *et al.* Addressing population-specific multiple testing burdens in genetic association studies. *Annals of Human Genetics* **79**, 136-47 (2015).
303. Gur-Wahnon, D. *et al.* The plasminogen activator system: involvement in central nervous system inflammation and a potential site for therapeutic intervention. *J Neuroinflammation* **10**, 124 (2013).
304. Adeyemo, A. *et al.* A genome-wide association study of hypertension and blood pressure in African Americans. *PLoS Genet* **5**, e1000564 (2009).
305. Srikumar, N. *et al.* PAI-1 in human hypertension: relation to hypertensive groups. *Am J Hypertens* **15**, 683-90 (2002).
306. Cuomo, O. *et al.* A critical role for the potassium-dependent sodium-calcium exchanger NCKX2 in protection against focal ischemic brain damage. *J Neurosci* **28**, 2053-63 (2008).
307. Pinsky, D.J. *et al.* Coordinated induction of plasminogen activator inhibitor-1 (PAI-1) and inhibition of plasminogen activator gene expression by hypoxia promotes pulmonary vascular fibrin deposition. *J Clin Invest* **102**, 919-28 (1998).
308. Woloszynek, J.C., Coleman, T., Semenkovich, C.F. & Sands, M.S. Lysosomal dysfunction results in altered energy balance. *J Biol Chem* **282**, 35765-71 (2007).
309. Di Malta, C., Fryer, J.D., Settembre, C. & Ballabio, A. Astrocyte dysfunction triggers neurodegeneration in a lysosomal storage disorder. *Proc Natl Acad Sci U S A* **109**, E2334-42 (2012).
310. Golovanova, N.K., Gracheva, E.V., Il'inskaya, O.P., Tararak, E.M. & Prokazova, N.V. Sialidase activity in normal and atherosclerotic human aortic intima. *Biochemistry (Mosc)* **67**, 1230-4 (2002).
311. Venerando, B. *et al.* Acidic and neutral sialidase in the erythrocytes of patients with type 2 diabetes: an answer to comments by Richard et al. *Blood* **101**, 2071 (2003).
312. Natori, Y., Ohkura, N., Nasui, M., Atsumi, G. & Kihara-Negishi, F. Acidic sialidase activity is aberrant in obese and diabetic mice. *Biol Pharm Bull* **36**, 1027-31 (2013).
313. Serrano, R. *et al.* Tissue-specific PAI-1 gene expression and glycosylation pattern in insulin-resistant old rats. *Am J Physiol Regul Integr Comp Physiol* **297**, R1563-9 (2009).
314. Yagami, A. *et al.* IL-33 mediates inflammatory responses in human lung tissue cells. *J Immunol* **185**, 5743-50 (2010).
315. Goljan, E.F. *Rapid Review Pathology*, (Mosby Elsevier, Philadelphia PA, 2011).
316. Xiao, W., Hsu, Y.P., Ishizaka, A., Kirikae, T. & Moss, R.B. Sputum cathelicidin, urokinase plasminogen activation system components, and cytokines discriminate cystic fibrosis, COPD, and asthma inflammation. *Chest* **128**, 2316-26 (2005).
317. Kim, Y.H. *et al.* Artery and vein size is balanced by Notch and ephrin B2/EphB4 during angiogenesis. *Development* **135**, 3755-3764 (2008).
318. Stefansson, S., McMahon, G.A., Petitclerc, E. & Lawrence, D.A. Plasminogen activator inhibitor-1 in tumor growth, angiogenesis and vascular remodeling. *Curr Pharm Des* **9**, 1545-64 (2003).

319. Kameyama, H. *et al.* The mRNA expressions and immunohistochemistry of factors involved in angiogenesis and lymphangiogenesis in the early stage of rat skin incision wounds. *Leg Med (Tokyo)* **17**, 255-60 (2015).
320. Quezada-Calvillo, R. *et al.* Luminal starch substrate "Brake" on maltase-glucoamylase activity is located within the glucoamylase subunit. *Journal of Nutrition* **138**, 685-692 (2008).