HAL
open science

Genetic determinants of cardiovascular disease : heritability and genetic risk score

Elias Levy Itshak Salfati

To cite this version:

Elias Levy Itshak Salfati. Genetic determinants of cardiovascular disease : heritability and genetic risk score. Cardiology and cardiovascular system. Université René Descartes - Paris V, 2014. English. NNT : 2014PA05S014 . tel-01127451

HAL Id: tel-01127451

https://theses.hal.science/tel-01127451v1

Submitted on 7 Mar 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THESE de DOCTORAT DE L'UNIVERSITE PARIS DESCARTES

Spécialité
Pathologies Cardio-vasculaire

Ecole doctorale 273
Biologie et biotechnologie (B2T)

Présentée par

Elias Levy Itshak Salfati

Pour obtenir le grade de
Docteur de L'UNIVERSITE PARIS DESCARTES

Genetic Determinants of Cardiovascular Disease:
Heritability and Genetic Risk Score

Devant le jury composé de:

Dr. Bernard Lévy	Directeur de Thèse
Dr. François Cambien	Rapporteur
Pr. Jean-Yves Boëlle	Rapporteur
Dr. Emmanuel Messas	Examinateur
Pr. Joël Ménard	Examinateur
Pr. Alain Tedgui	Examinateur

ABSTRACT

Complex diseases such as cardiovascular disease (CVD) are influenced by both genetic and environmental factors. Estimation of an individual's cardiovascular risk usually involves measurement of risk factors correlated with risk of CVD (e.g. age, sex, smoking, blood pressure, and total cholesterol). Lately, several biomarkers have been evaluated for their ability to improve prediction of cardiovascular disease beyond traditional risk factors. The interest in novel loci is propelled notably by emerging discoveries from the advent of genome-wide association studies (GWAS) of genetic variants associated with risk for common diseases. GWAS has greatly enhanced our knowledge of the genetic architecture of cardiovascular disease, yielding over 50 variants confirmed to be associated with CVD to date, as well as over 200 associated with traditional cardiovascular risk factors (e.g. lipids, blood pressure, body mass index, and type 2 diabetes mellitus). This recent and continuing success in discovering increasing numbers of robustly associated genetic markers has led to reassessment of whether genetic data can provide clinically useful information by refining risk prediction and moderating disease risk through a more efficient application of prevention strategies. In this thesis, we first address novel approach to survey the genetic architecture of hypertension (i.e. major risk factor for premature CVD), then construct risk prediction models for coronary artery disease (CAD; i.e. most common type of CVD) and finally establish a common genetic basis of the strongest predictor of clinical complications of CAD, subclinical atherosclerosis, to add incremental prognostic value above traditional risk scores across a range of ages.

We show that, for first visit measurements, the heritability is ~25%/~45% and ~30%/~37% for systolic (SBP) and diastolic blood pressure (DBP) in European (N=8,901) and African (N=2,860) ancestry individuals from the Atherosclerosis Risk in Communities (ARIC) cohort, respectively, in accord with prior studies. Then we present a means to combine a polygenic risk score - genetic effects among an ensemble of markers - with an independent assessment of clinical risk using a log-link function. We apply the method to the prediction of coronary heart disease (CHD) in the ARIC cohort. The addition of a genetic risk score (GRS) to a clinical risk score (CRS) improves both discrimination and calibration for CHD in ARIC and subsequently reveal how this genetic information influences risk assessment and thus potentially clinical management. Finally, Among 1561 cases and 5068 controls, from several clinical and genetic datasets available through the NCBI's database of Genotypes and Phenotypes (dbGAP), we found a one SD increase in the genetic risk score of 49 CAD SNPs was associated with a 28% increased risk of having advanced subclinical coronary atherosclerosis ($p = 1.43 \times 10\text{-}16$). This increase in risk was significant in every 15-year age stratum ($.01 > p > 9.4 \times 10\text{-}7$) and was remarkably similar across all age strata (p test of heterogeneity = 0.98). We obtained near identical results and levels of significance when we restricted the genetic risk score to 32 SNPs not associated with traditional risk factors.

Accordingly, common variation largely recapitulates the known heritability of blood pressure traits. The vast majority of this heritability varies by chromosome, depending on its length, and is largely concentrated in intronic and intergenic regions of the genome but widely distributed across the common allele frequency spectrum. Respectively, our proposed method to combine genetic information at established susceptibility loci with a non-genetic risk prediction tool facilitates the standardized incorporation of a GRS in risk assessment. Lastly, multi-locus GRS derived from the high-risk alleles of SNPs associated with clinical complications of CAD is strongly associated with the presence of advanced subclinical atherosclerosis. This susceptibility to plaque formation is life long, remarkably homogenous, and not driven by exposure to traditional risk factors.

Table of Contents

CHAPTER 1: INTRODUCTION **3**

1.1 Blood pressure, major cardiovascular risk factor 3

1.2 Hypertension, major Cardiovascular Disease risk factor 4

1.3 The Genetics of Hypertension 4

1.4 Genome-Wide Association Studies (GWAS) of Hypertension 5

1.5 The Heritability (h^2) of Hypertension 6

1.6 The hunt for missing Heritability 7

1.7 Development of Atherosclerosis 8

1.8 The most common type of Cardiovascular Disease 9

1.9 The strongest predictor of Coronary Artery Disease: CAC 9

1.10 Genome-Wide Association Studies of CAD 10

1.11 Clinical and Genetic Risk Assessment of Cardiovascular Disease 12

CHAPTER 2: THE HERITABILITY OF BLOOD PRESSURE **14**

Preface to the Manuscript **14**

2.1 Abstract: **17**

2.3 Results: **20**

2.4 Discussion: **25**

2.5 Materials and Methods: **28**

Legends to Figures: **32**

Legends to Tables: **38**

Legend to Supplementary Table: **45**

CHAPTER 3: CLINICAL AND GENETIC RISK ASSESSMENT OF CARDIOVASCULAR DISEASE **52**

Preface to the Manuscript **52**

3.1 Abstract **55**

3.2 Introduction **56**

3.3 Methods **57**

3.3.1 SNP Selection & Weighting 57

3.3.2 Prospective Cohort for testing Genetic Risk Scores 58

3.3.3 Clinical Risk Score Assessment 59
3.3.4 Imputation of ARIC raw genotype data to 1000 genomes 59
3.3.5 GRS Construction 60
3.3.6 Combining Clinical and Genetic Risk 60
3.3.7 Evaluation of performance of risk scores 62
3.3.8 Risk Reports 63
3.4 Results 64
3.4.1 ARIC cohort exclusions 64
3.4.2 Risk Scores 64
3.4.3 Performance of risk scores and sensitivity analyses 64
3.4.4 Risk Reports 65
3.5 Discussion 66
Figure Legends 70
CHAPTER 4: GENETIC RISK ASSESSMENT OF CORONARY PLAQUE BURDEN 77
Preface to the Manuscript 77
4.1 Abstract 80
4.2 Introduction 81
4.3 Material and Methods 82
4.3.1 Study Population 82
4.3.2 Case definition 82
4.3.3 SNP Selection and Imputation 83
4.3.4 Construction of the GRS 83
4.3.5 Statistical Analysis 84
4.4 Results 85
4.5 Discussion 86
Legends to Figures: 91
Legends to Tables: 95
Legend to Supplementary Figures: 98
CHAPTER 5: DISCUSSION 101
References 106

CHAPTER 1: INTRODUCTION

1.1 Blood pressure, major cardiovascular risk factor

"The measurement of [blood pressure] is likely the clinical procedure of greatest importance that is performed in the sloppiest manner." [1]

Blood pressure is the pressure within the major arterial system of the body measured in millimeters of mercury (mm Hg) with a sphygmomanometer and usually expressed as the systolic - contraction of the ventricles, heart beats - over the diastolic pressure (filling of the heart with blood between ventricles contractions, heart rests between beats). In other words, systolic pressure is the maximum blood pressure and diastolic pressure is the minimum pressure recorded just prior to the next contraction.

Optimal blood pressure is less than 120 mm Hg systolic and 80 mm Hg and ideally perfused adequately vital organs[2] without causing damages. Organs inadequately perfused experience ischemic damage and are unable to perform efficiently. For instance, poor renal perfusion may trigger renal failure with thorough metabolic outcomes. In contrast, elevated blood pressures may cause end organ damage with harmful consequences such as heart attack, stroke, kidney failure and dissecting aneurysm. Blood pressure is considered high if it is 140 mm Hg systolic and/or 90 mm Hg diastolic or higher[2].

Low blood pressure is manifested by fatigue or shortness of breath on effort. Since perfusion pressure is inadequate, the increased oxygen and nutrient demand by exercising muscles cannot be satisfied and can cause symptoms of dizziness and fainting. The most frequent causes of low blood pressure are due to dehydration or reduced blood volume.

1.2 Hypertension, major Cardiovascular Disease risk factor

Untreated high blood pressure (BP) is an important major cardiovascular risk factor for ischemic heart disease, cardiac and renal failure and accounts for a large fraction of morbidity and mortality worldwide[3]. In North America, nearly one third of the adults over 20 years of age suffer from Hypertension (HTN) [3].

Essentially, blood pressure is a quantitative trait controlled by cardiac output[4], peripheral vascular resistance and blood volume. This trait varies between subjects by a variety of environmental and physiologic factors such as age, BMI (body mass index), smoking and physical activity; yet adjustment for these covariates only explain a small proportion of blood pressure variability. Application of genetics and genomics offer subsequently a major opportunity to elucidate the remaining fraction of blood pressure variation.

1.3 The Genetics of Hypertension

The nature of the inherited basis of hypertension has been questioned in the 1950s, since the classic work of Lord Robert Platt and Sir George Pickering[5]. Platt claimed that hypertension was influenced by a single Mendelian genetic variant responsible of a dichotomous frequency distribution curve of blood pressure levels. Pickering, while recognizing the importance of inherited factors in blood pressure regulation, countered that multifactorial genetic inheritance drove a continuous, unimodal blood pressure distribution with the hypertensive and non-hypertensive segments defined by an arbitrary line. Platt's hypothesis is supported by rare variants with large effects that cause monogenic hypertension syndromes, whereas Pickering's theory is substantiated by variants (polygenic) with small effect sizes that collectively contribute to essential hypertension.

Over the years, researchers have gained insight into the genetic architecture of hypertension as a result of technologic advances that permit the genotyping of million single nucleotide polymorphisms (SNPs) on a single microarray[6]. These genomic tools enable the investigation of a considerable proportion of the common human genetic

variation throughout the genome. According to numerous studies, it appears that common variants act on common disease at many loci[7-10], explaining little individually but explaining a much larger share of the trait or disease collectively. Previous investigations of complex genetic disease by candidate gene studies or linkage analysis were not designed toward identification of variants with these features[11]. The genome-wide association study (GWAS) offers the first opportunity to test such hypotheses[12, 13].

1.4 Genome-Wide Association Studies (GWAS) of Hypertension

Association testing of every single SNP against blood pressure traits opens the way for an unbiased interrogation of genetic causes of these traits. The basic methodology used to test for association between blood pressure and hundreds of thousands of SNPs disseminated throughout the human genome is rudimentary: an association statistic is calculated to assess the relationship between each SNP and the phenotype of interest, generally by linear regression for continuous phenotypes or by logistic regression for dichotomous phenotypes. In other words, it compares the frequency of SNPs in cases and controls; high frequency of the SNP in cases suggests close proximity to a genetic risk variant. The number of tests performed is equivalent to the number of SNPs; although the number of independent tests is lower since many SNPs are correlated. Given the large numbers of tests carried-out, highly significant results can be obtained by chance. Subsequently this burden of multiple testing required stringent thresholds[14, 15], current practice seems to prefer a threshold of $P = 5 \times 10^{-8}$ (based on an adjusted p-value of 0.05 for one-million tests).

In order to infer genotypes at un-typed SNPs, genotypes used for GWAS are imputed to the 1000 Genomes Project datasets[15, 16], bringing the number of variants up to ~30 million. Imputation backbone of the 1000 Genomes project, therefore increase statistical power and extend the allele frequency spectrum analyzed.

The first GWAS to identify HTN variants was carried out by the Wellcome Trust Case Control Consortium (WTCCC) in 2007 using 2,000 cases and 3,000 shared controls[17]. From this large-scale analysis no variants reached genome-wide significance,

calling out the need for larger sample sizes. Since then, several consortia and individual studies have identified over 49 variants associated with SBP, DBP, or HTN[18-28]. These discovery efforts were carried out using samples of European descent, such as the CHARGE (Cohorts for Heart and Aging Research in Genomic Epidemiology) consortium, the Global BP Gen (Global BP Genetics) consortium, and the ICBP (International Consortium for BP).

To date, several loci have been identified for SBP, DBP, and HTN as a result of the contribution of the ICBP effort and many other studies. The ICBP-GWAS experiment of 2011 included 69 395 individuals and further replication genotyping in up to 133 661 subjects[18]. This study described 29 SNPs with genome-wide significance and replicated 13 loci identified in previous effort. As of now, 49 causal variants are associated to blood pressure traits and interestingly all of the alleles increasing SBP/DBP also increased the risk of HTN. Of the 49 variants significantly associated with SBP, DBP and HTN, a minority is near a gene related to BP. The remaining variants are localized in genomic regions that were previously unsuspected for their link with BP. The effect sizes for each individual genetic variant are small, typically 1 mmHg for SBP and 0.5 mmHg for DBP. Even collectively, the variance tagged by the 49 associated SNPs explains only 1-2% of the expected narrow-sense heritability (h^2) for SBP and DBP[18], defined as the ratio of the additive genetic values (combined effect of all loci) to the total phenotypic variance.

1.5 The Heritability (h^2) of Hypertension

Heritability is often estimated to summarize the proportion of phenotypic variation due to variation in genetic factors[29, 30]. Given its definition the heritability always lies between 0 and 1. Current estimate of the heritability of BP is approximately 25 times larger than the variation accounted by GWAS SNPs[18]. This mismatch between the high heritability estimates from quantitative genetic analyses and the small fraction of variance explained through GWAS findings has been designated as the 'missing heritability'[31, 32]. This discrepancy is likely due to rare genetic variants ("common disease–rare variant" hypothesis[33, 34]) or because genotyped SNPs are in incomplete linkage disequilibrium (LD) with the causal variants[35]. Since non-additive genetic effects

(gene-by-gene or gene-by-environment interactions) do not contribute to the narrow-sense heritability. Part of these "missing" genetic factors contribute to the estimated genetic effect, but are not detected in GWAS analyses, which capture only additive effects of common SNPs with minor allele frequencies (MAF) of 5%. Lack of complete LD might, for instance, occur if causal variants have lower MAF than genotyped SNPs. A second suggested explanation is that multiple additional common genetic variants contributing to the genetic effect observed in quantitative genetic studies have such small effect sizes (fail to pass stringent significance thresholds) that they remain undetected in large data sets used in contemporary GWAS analyses.

1.6 The hunt for missing Heritability

GWASs have successfully identified thousands of SNPs significantly associated with complex traits and diseases in human by examining each SNP individually for significance; yet these variants typically account for only a small fraction of the genetic variation[36]. Alternatively, the narrow sense heritability explained by the common SNPs may be estimated by adapting a linear mixed-effects model for all SNPs simultaneously[37]. The effects of the SNPs are treated statistically as random, and the variance explained by all the SNPs together is estimated. This approach subsequently decomposes the phenotypic variance into genetic and residual variance components. Usually, the estimate of heritability is applied to related individuals where the genetic relationships are assessed by using family pedigree or SNP data. Recent study pointed out that heritability could be estimated using genetic relationships obtained from autosomal SNPs for unrelated individuals since degree of genetic resemblance for common SNPs at the whole-genome level is normally distributed among unrelated individuals[35]. The main difference between the latter and the former estimates is due to the difference in LD between the common SNP and the rest of the genome, with the assumption that closely related individuals would be in greater LD than unrelated subjects. Thus, heritability estimated with the genetic relationships of unrelated individuals is attributed to the common variants while that estimated with genetic relationships of related individuals is attributed to the entire genome. Furthermore, estimate of the genetic variance using common markers in unrelated individuals is directly comparable to results from GWAS,

since both are based on the same experimental design. While the method does not identify single variants, it quantifies the overall contribution from the additive effects of SNPs in currently available DNA array.

This alternative approach designed to overcome the weaknesses of GWAS can provide an unbiased estimate of the variance explained by all SNPs, given that only a small fraction of the BP heritability is currently explained by genome-wide significant SNPs[18].

Untreated elevated blood pressure is not only a well-established cardiovascular risk factor but predisposes to and accelerates atherosclerosis[38-40].

1.7 Development of Atherosclerosis

Atherosclerosis is a degenerative inflammatory disorder characterized by the progressive deposition of lipids and fibrous matrix in the arterial wall, which accounts for substantial cardiovascular morbidity and mortality[3, 41, 42]. The first stages of atherosclerosis are characterized by the loss of the normal barrier function of the endothelium, lipoprotein abnormalities that promote the accumulation of lipoproteins, particularly low-density lipoproteins (LDL), in the intimal region[43]. In response to the lipoprotein accumulation, dysfunctional endothelial cells express a number of adhesion molecules and selectins that promote the binding of circulating monocytes to vascular endothelial cells[42, 43]. Monocytes are then exposed to chemokine that induces the transmigration of bound monocytes into the sub-endothelial space. In the inflamed intima, a cytokine or growth factor give rise to the differentiation of monocytes into macrophages. This step is critical for the development of atherosclerosis[44]. Macrophages are subsequently exposed to modified LDL and express scavenger receptors that bind and promote the ingestion of oxidized LDL. As the macrophage progressively accumulates cholesterol, the macrophage takes on the appearance of a lipid-laden foam cell. In concert, T-lymphocytes infiltrate the developing lesion site from both the intimal and adventitial aspects of the vessel wall, where it secretes inflammatory cytokines and growth factors[45]. This provides a signal for smooth muscle cells to alter their

cytoskeleton, migrate from the media into the intimal space, where they proliferate and secrete extracellular matrix components that form a fibrous cap over the developing lesion[42, 43, 46]. The ongoing inflammatory response in the vascular wall continues to provide signals for further LDL uptake and leucocyte infiltration, creating conditions for further growth of the atherosclerotic lesion[47]. Over time, the atherosclerotic lesion continues to expand at its base via the same mechanisms that led to formation of the initial fatty streak.

1.8 The most common type of Cardiovascular Disease

The stability of the advanced atherosclerotic lesion or plaque (lipid core bounded by a fibrous cap) depends on its cellular and extracellular contents. Plaques with small lipid cores, thick fibrous caps, few inflammatory cells and a preponderance of smooth muscle cells are typically stable; conversely, those with large lipid cores, thin fibrous caps, numerous macrophages and relatively few smooth muscle cells are most likely to rupture (vulnerable plaque). Atherosclerotic lesion rupture usually are caused by resident activated macrophages, T cells, at sites of plaque rupture that secrete several types of molecules such as cytokines and vasoactive molecules that can destabilize lesions, which inhibit the formation of stable fibrous caps, attack collagen in the cap, and initiate thrombus (blood clot) formation[47, 48]. Thereafter, the disrupted plaque serves as a scaffold to allow platelet aggregation and coagulation. The thrombus size depends on the extent of plaque rupture as well as activity of the endogenous fibrinolytic pathway. When sufficiently large, the thrombus can either partially or completely occlude the coronary vessel lumen and precipitate an acute coronary event (e.g. unstable angina, myocardial infarction (MI), and sudden death). Coronary artery disease (CAD) is almost always due to atheromatous narrowing and subsequent occlusion of the vessel[49].

1.9 The strongest predictor of Coronary Artery Disease: CAC

The prevalence and extent of atherosclerosis development increase with age[50]. Age is used as a surrogate for coronary plaque burden, but plaque burden is the true risk factor for coronary heart disease related morbidity and mortality[51]. Because plaque burden can vary among individuals at any given age, accurate measurement of subclinical

atherosclerosis may provide a better method for predicting risk for acute cardiovascular events.

A variety of invasive and non-invasive techniques are available to measure atherosclerosis and subclinical atherosclerosis. These techniques can ascertain parameters such as luminal diameter or stenosis, vessel wall thickness, plaque volume, and the specific distribution and localization of atherosclerotic disease. Accordingly, Computed Tomography (CT) Scan is the only noninvasive test to evaluate the lumen and wall of the coronary artery with high sensitivity and specificity for calcium detection, and capable of quantifying coronary artery calcification (CAC) [52]. This measures is translated into an Agatston score, calculated by multiplying the lesion area by a density factor derived from the maximal Hounsfield units in this area. Coronary calcium reflects plaque burden, because calcium deposits are related to the lipid and apoptotic remnants of the plaque. This calcification of the atherosclerotic plaque (deposition of calcium phosphate in the vessel wall) is limited to the sub-intimal space and can appear as early as the second decade of life, soon after the formation of fatty streaks[53]. First seen in the lipid core of the atheroma it occurs via an active process that resembles bone formation and is controlled by complex enzymatic and cellular pathways, including osteoblast-like cells, cytokines, transcription factors and bone morphogenetic proteins, which are typically involved in bone calcification, are also involved in the process of vascular calcification. Subsequently, the presence of calcium in coronary arteries is pathognomonic of atherosclerosis[54]. CAC is an independent cardiovascular risk factor that adds prognostic information when considered in conjunction with other risk factors.

However, although CT scan can localize coronary plaques within the coronary tree and provide a quantitative measure of relative disease severity, it can be used to ascertain highly vulnerable patient rather than the susceptibility of individual plaques to rupture.

1.10 Genome-Wide Association Studies of CAD

In late 2007, three independent GWAS for CAD identified a significant association signal on chromosome 9p21[55-57]. No prior genetic studies had implicated this

region of the genome. Moreover, the SNPs in the locus that were associated with coronary disease were not associated with any traditional cardiovascular risk factors. Thus, it appears that the genetic mechanism underlying the association signal is operating through a novel pathway. Subsequent studies established an association between the 9p21 locus with MI and other vascular phenotypes such as abdominal aortic aneurysm, intracranial aneurysm, and peripheral arterial disease, suggesting that the sequence variations may interfere with vascular tissue development[58-60]. The 9p21 locus illustrates the difficulty of linking some of the genetic associations identified by GWAS with pathological mechanism. No annotated genes are present in the minimal region of association as defined by linkage disequilibrium, the closest genes to the locus, CDKN2A, CDKN2B, and ARF, are more than 100 kb away from the index SNPs (SNPs with the highest level of association), making it unclear how the causal DNA variant(s) might influence coronary disease.

One possibility is that the loci harbor non-gene transcripts that regulate other genes or lies in a regulatory element (e.g., a transcriptional enhancer) that affects the transcription of a gene or genes that are ultimately responsible for the phenotype[61].

Lately, a large consortium of investigators focused on coronary disease (the Coronary Artery DIsease Genome-Wide Replication and Meta-analysis plus The Coronary Artery Disease (C4D) Genetics, or CARDIoGRAMplusC4D Consortium) has assembled 63,746 cases of coronary disease and 130,681 control samples and discovered 15 additional associated loci[62], reaching genome-wide significance, taking the number of susceptibility loci for CAD to 46[55, 63-70], and a further 104 independent variants ($r2 < 0.2$) strongly associated with CAD at a 5% false discovery rate (FDR).

GWAS results represent a rich source of information for treatment research that forms bridges between genome science and clinical and public health practice[71, 72]. Given the large number of genome-wide studies, sufficient data exist to support such translational research for a number of common chronic health conditions, including CAD[73, 74]. Infrastructure is now in place at the start of the translational pipeline, with

GWAS data extracted and curated in continuously updated catalog[75, 76]. Likewise, at the other end of the pipeline, evidence from translational research is estimated to establish the clinical benefit of genomic information and issue guidelines for clinical practice[77]. However, significant gaps remain in the middle of the translational pipeline, and approaches are needed to support research at this intersection, so that population-based samples with rich environmental and phenotypic measurements can be used to follow up disease markers identified in GWAS. In accordance, systematic approaches are needed to scrutinize the results of various association studies and distill the most promising set of markers for further investigation.

1.11 Clinical and Genetic Risk Assessment of Cardiovascular Disease

CAD is a public health problem, which is highly prevalent; it is a significant source of morbidity and mortality under strong genetic influence and lately GWAS started to elucidate its molecular genetic roots. Consequently, risk assessment plays a critical clinical role in prevention strategies and in therapy for CAD at the individual level, and is key to future efforts in personalized medicine in this area. Current risk prediction models are established on traditional risk factors (TRFs) such as age, sex, smoking, lipid levels, and blood pressure, and although extensively validated, these models have limitations[78]. Lately, multiple studies have evaluated the ability of "emerging risk factors," including biomarkers, and genetic variants, to improve CHD risk assessment beyond the use of TRFs[79]. Of the biomarkers that could be objectively and systematically measured, genetic variants have some unique features in that they do not change over time. Subsequently, multi-locus profiles of genetic risk, so-called "genetic risk scores" (GRS) can be used to translate discoveries from GWAS into tools for population health research[80, 81]. GRS summarize risk-associated variation across the genome by aggregating information from multiple-risk SNPs (summing up the number of disease-associated alleles). Since GRS pool information from numerous SNPs, each individual SNP is less important to the summary measurement, and thereby counteract the lack of linkage for any one SNP. For the same reason, GRS is less sensitive to minor allele frequencies for individual SNPs. As the number of SNPs included in a GRS increases, the distribution of values approaches normality, even when individual risk

alleles are relatively uncommon[82]. Therefore, the GRS can be an efficient and effective means of constructing genome-wide risk measurements from GWAS findings.

To address these challenges, data from the population-based Atherosclerosis Risk in the Communities (ARIC) Study were used to estimate the genomic contributions to blood pressure heritability. Then, we proposed a method to facilitate the standardized incorporation of a GRS in risk assessment for complex traits with CAD as example. Finally, I examined the association between a GRS of high-risk alleles associated with clinically significant complications of CAD and the presence of subclinical atherosclerosis.

CHAPTER 2: THE HERITABILITY OF BLOOD PRESSURE

Preface to the Manuscript

This manuscript presents the first part of a study investigating the genetic basis of cardiovascular disease by examining the heritability of the leading cardiovascular risk factor.

This study focuses on determining whether common variant capture a large proportion of blood pressure traits variability. We felt that it was important to answer this question first because such estimate could establish the total contribution of genotyped markers on current SNP arrays for blood pressure.

We utilized data on genotyped common SNPs and imputed SNPs respectively, in European ancestry and African ancestry from ARIC population, separately, and a mixed linear model for analyses.

Subjects under anti-hypertensive treatments were adjusted for potential medication effects by adding 10 and 5 mm Hg to observed systolic (SBP) and diastolic (DBP) blood pressure measurements, respectively.

The variance explained by common variant in this study was a function of chromosome size, minor allele frequency (MAF), functional annotation (coding, intronic and intergenic; cardiovascular/renal or other genes), and markers enriched for functional candidates (GWAS blood pressure loci; Cardio-Metabochip SNPs) respectively. These genomic partitions reflect actual genetic architecture of blood pressure variance from several angles.

In this context, the general objective of this first study is to estimate the proportion of variance tagged by common SNPs for blood pressure traits.

The specific objective of this first study is:

1. To identify the extent to which common variants can explain the amounts and distribution of SBP and DBP variation within the genome and with respect to allele frequency, coding versus non-coding DNA and sites of gene expression.

2. To compare the variance explained by directly genotyped SNPs to the variance explained by genotyped and imputed SNPs.

3. To compare the genetic variance captured by common SNPs for European ancestry to the genetic variance attributed by common SNPs for African ancestry.

4. To compare whether more stringent definition of unrelated individuals affect the estimate of blood pressure heritability.

Direct estimates of the genomic contributions to blood pressure heritability within a population cohort (ARIC)

Elias Salfati[1,2], Alanna Morrison[3], Eric Boerwinkle[3] and Aravinda Chakravarti[1, 4]

[1]Center for Complex Disease Genomics, McKusick - Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205

[2]Ecole Doctorale B2T, IUH; Université Paris 7, 75010 Paris , France

[3]Human Genetics Center, University of Texas Health Science Center, Houston, TX 77030.

2.1 Abstract:

Blood pressure (BP) is a heritable trait with multiple environmental and genetic contributions with current heritability estimates from twin and family studies being ~ 40%. Here, we use genome-wide polymorphism data from the Atherosclerosis Risk in Communities (ARIC) study to estimate BP heritability from genomic relatedness among cohort members. We utilized data on 656,362 and 772,638 genotyped common single nucleotide polymorphisms (SNPs), and up to 7,558,733 and 9,578,528 imputed SNPs, in 8,901 European ancestry (EA) and 2,860 African Ancestry (AA) ARIC participants, respectively, and a mixed linear model for analyses. We show that, for first visit measurements, the heritability is ~25%/~45% and ~30%/~37% for systolic (SBP) and diastolic blood pressure (DBP) in European and African ancestry individuals, respectively, in accord with prior studies. A new finding is that common variation largely recapitulates the known heritability of BP traits. The vast majority of this heritability varies by chromosome, depending on its length, and is largely concentrated in intronic and intergenic regions of the genome but widely distributed across the common allele frequency spectrum. Interestingly, the majority of this heritability arises from loci harboring currently known cardiovascular and renal genes. Recent meta-analyses of large-scale genome-wide association studies (GWASs) and admixture mapping have identified ~50 loci associated with BP and hypertension (HTN) and yet they account for only a small fraction (~2%) of the heritability. Consequently, elucidation of BP genes will require focused analysis of cis-regulatory elements controlling cardiovascular and renal gene expression.

2.2 Introduction:

Blood pressure (BP) is an established risk factor for multiple cardiovascular diseases (CVD) and, worldwide, about one-tenth of adult global death is attributable to high blood pressure or essential hypertension[83]. It's a truism that BP, studied through systolic (SBP) or diastolic (DBP) measures or clinically defined hypertension (HTN), is a complex, polygenic trait that is influenced by both genetic and environmental factors[84]. BP heritability is moderate, and ~40% across studies (18), and has classically been estimated from twin and family studies. Molecular genetic analyses of BP genetics have been challenging with the exception of Mendelian hypo- and hypertension syndromes that show large BP variation in individuals harboring loss- and gain-of-function mutations in numerous renal genes[85]. These latter studies convincingly demonstrate that renal salt-water homeostasis is key to maintaining blood pressure control and is rate limiting. Nevertheless, it is unknown whether loss of renal salt-water homeostasis is primary or secondary to elevated BP arising from other mechanisms. Several environmental factors that influence BP levels, such as alcohol consumption, dietary salt-intake, physical activity and stress, are also known but the biochemical paths of their action remain incompletely described. Identification of the genes that influence inter-individual variation in BP thus remains a key and important challenge since this can lead to discovery of new etiological pathways.

In recent years, genetic advancements have made it feasible to conduct a comprehensive search for genes underlying a trait. To date, large-scale genome-wide association studies (GWAS), and other genome-wide analyses, have identified ~50 single nucleotide polymorphisms (SNPs) associated ($P < 5 \times 10^{-8}$) with genetic risk factors contributing to inter-individual variation in BP[18-20, 22, 23, 25-28, 86]. By design, the vast majority of these genetic variants is common (>10%) in the general population and have small ($<0.05\sigma$, where σ^2 is the residual phenotype variance) allelic effects, and collectively these loci explain only a small (<5%) fraction of the phenotypic variance (i.e. heritability) [87]. This substantial gap between the overall and identified heritability has led to a great deal of speculation as to the causes for this "missing" heritability, including our failure to assess effects at rare variants and copy number polymorphisms[88, 89].

Nevertheless, before we entertain new genetic hypotheses for complex trait architecture it is first necessary to answer what is the total contribution of all *common genetic variation* to BP heritability? The typical approach to providing this answer has been through summing the contributions of individual SNPs showing genome-wide significant associations: this approach leads to a severe underestimate since GWASs suffer from a high false negative rate in its attempt to control the false positive rate. This false negative rate arises from the majority of genetic effects being too small to reach statistical significance and incomplete linkage disequilibrium between genotyped markers and causal variants.

Newer statistical methods allow a robust answer to this question by *estimating* the trait residual variance explained by all common SNPs taken together and by considering them as random effects in a mixed linear model[35, 37]. Indeed, these analyses can be conducted on all genomic polymorphisms or those restricted to specific subgroups, such as individual chromosomes, allele frequency class or functional annotation, to assess relative contributions from these subgroups. Visscher and colleagues have demonstrated that some complex traits arise largely from allelic effects of common variants[35, 90-93]. We use their approach to ask: is inter-individual BP variation mostly due to polymorphic additive genetic factors? We also investigate the proportion of inter-individual BP variation captured by common SNPs as a function of chromosome size, minor allele frequency (MAF) of genotyped variants, by functional annotation (coding, intronic and intergenic; cardiovascular, renal or other genes), and by markers enriched for functional candidates (GWAS BP Loci; Cardio-Metabochip SNPs). Finally, we also used longitudinal phenotype data, and assessing the effect of long-term average (LTA) BP, to detect additional genetic variance through reducing measurement error[94]. These analyses demonstrate that the majority (>50%) of both SBP and DBP heritability is from common (MAF>10%) genetic variation almost exclusively in non-coding (intronic and intergenic) DNA and at loci enriched for cardiovascular and renal genes. Consequently, genetic etiologies of BP and HTN are addressable through independent identification of cardiovascular and renal genes and focused identification of the underlying variants and

genes. We propose specific approaches for identifying the causal factors for BP, approaches that do not depend merely on larger GWAS studies but require specific understanding of the cis-regulatory architecture of the human genome.

2.3 Results:

The majority of our analyses are on the full set of 8,901 EA and 2,860 AA unrelated individuals within ARIC (Table 1; Table 2). The pairwise genomic relationship matrix (GRM) was estimated for these individuals using the high quality (call rate > 95%; MAF ≥ 1%; HWE $P > 10^{-6}$) autosomal genotypes at 656,362 and 772,638 directly genotyped SNPs and also including all 7,558,733 and 9,578,528 imputed markers with MAF ≥ 1% (imputation $R^2 \geq 0.3$) in EA and AA participants, respectively. To avoid phenotypic resemblance due to non-additive genetic effects and common environmental influences, we excluded one of each pair of individuals with an estimated genetic relationship > 0.025 (equivalent to 2nd cousins). Consequently, we created a second dataset that included only 6,914 and 1,763 genetically "unrelated" EA and AA participants, correspondingly, to assess whether they impacted our conclusions. The BP distributions of the 8,901 EA and 2,860 AA individuals are not statistically different from the 6,914 EA and 1,763 AA, respectively, suggesting that the use of either set would be representative of the population's BP features (Table 1, Table 2). Relatedness between participants using genotyped and genotyped and imputed SNPs followed normal distributions with mean -0.00015 (s.d.= 0.0044) and -0.00014 (s.d.=0.0043), respectively, and showed trivial differences. Consequently, for the remaining analyses we used the set of 8,901 EA and 2,860 AA individuals to maximize the available sample sizes.

The GRMs were fitted to a mixed linear model (MLM) to SBP and DBP and restricted maximum likelihood (REML) methods were used to estimate the proportion of variance explained by genetic markers. Two types of analyses were performed that estimated the proportion of variance explained by the sum of that on individual chromosomes and by the whole-genome: for the first analysis, we fit 22 pairwise relationship matrices simultaneously (joint analysis) while in the second we merged these relationship matrices into one GRM (combined analysis). Estimates from both analyses

were very similar. The phenotypic variance explained by only genotyped SNPs (~600k SNPs in EA, ~700k SNPs in AA) was 0.25 in EA (SE = 0.05, $P = 2x10^{-8}$) and 0.45 in AA (SE = 0.12, $P = 1.1x10^{-5}$) for systolic and was 0.31 in EA (SE = 0.05, $P = 2\ x10^{-15}$) and 0.29 in AA for diastolic (SE = 0.11, $P = 7\ x10^{-5}$) blood pressure, and were highly significant (Table 3; Table 4). These estimates were nearly identical to the variance explained by considering all imputed and genotyped SNPs (~7m SNPs in EA, ~9m in AA), and were 0.23 in EA (SE = 0.05, $P = 7\ x10^{-7}$) and 0.40 in AA (SE = 0.23, $P = 4x10^{-3}$) for SBP and 0.32 in EA (SE = 0.05, $P = 3x10^{-14}$) and 0.37 in AA (SE = 1.2, $P = 1x10^{-3}$) for DBP and, once again, were highly significant (Supplementary Table 1; Supplementary Table 2). Therefore, the estimated variances are stable, largely from the effects of polymorphisms (MAF ≥ 1%) and do not appear to be dependent on the number of SNPs used. This result is not surprising since imputation increased the numbers of markers but included those that were highly correlated to the primary genotyped set of common alleles.

The apportioning of variance explained by individual chromosomes clearly demonstrated that although there is a general yet significant correlation between chromosome length and variance explained (SBP r_{cor} = 0.26 (EA) / 0.56 (AA); DBP r_{cor} = 0.31 (EA) / 0.42 (AA)), individual chromosomes differed considerably in their contributions to BP variation (Figure 1, Figure 2). Moreover, there is a high but not absolute concordance between the variances explained for both SBP and DBP by each chromosome. In EA the highest proportion of genetic variance captured by chromosome for SBP is from three chromosomes: chromosomes 2 (h^2~3%; SE ~0.011), 4 (h^2~2.5%; SE ~0.011) and 12 (h^2~2.2%; SE ~0.012). Likewise, in AA the three chromosomes that account for the largest fraction of genetic variance for SBP were chromosomes 2 (h^2~11%; SE ~0.04), 5 (h^2~5.9%; SE ~0.04) and 11 (h^2~5.1%; SE ~0.03), capturing nearly half of the genetic variance. With respect to DBP, the most prominent contributions of genetic variation were from four chromosomes in EA: chromosomes 2 (h^2~3.1%; SE ~0.015), 4 (h^2~2.5%; SE ~0.014), 11 (h^2~3%; SE ~0.013) and 16 (h^2~2.3%; SE ~0.013); whereas for AA, five chromosomes accounted for the highest variances, namely, chromosomes 2 (h^2~4%; SE ~0.04), 3 (h^2~3.6%; SE ~0.032), 5

(h^2~6.2%; SE ~0.038), 11 (h^2~3.5%; SE ~0.03) and 13 (h^2~5.4%; SE ~0.033), tagging over 65% of the genetic variance.

Blood pressure is a naturally varying phenotype. Thus, we assessed whether using the BP measurements from multiple (2-4) visits across time, as a Long Term Average (LTA), would lead to different conclusions by reducing measurement error[95, 96]. We used the same set of directly genotyped SNPs and LTA for SBP and DBP for similar analyses in EA and AA subjects (Supplementary Table 3). The proportion of genetic variance captured by chromosome for SBP showed a high correlation with first visit measurements BP (EA: r_{cor} ~ 0.73, P = 8.84 x10^{-5}; AA: r_{cor} ~ 0.78, P = 1.28 x10^{-5}) with greater variation explained by chromosomes 4 (~40% increase in both EA and AA), 10 (~40% increase in EA), 16 (~100% increase in AA) and 17 (~60% increase in EA). For DBP, the LTA measurements are smaller than those from first visit values in EA (r_{cor} ~ 0.54, P = 9.4 x10^{-3}) with the majority of the variation explained by chromosomes 2, 6, 10 and 11 (~ 50% to 75% decrease). In contrast, in AA, the variance explained by DBP-LTA is 50% more than first visit (r_{cor} ~ 0.65, P = 9.4 x10^{-4}), with the majority of the variation explained by chromosomes 1, 3 and 13 (~ 30% to 75% increase). Consequently, the pattern of the variance explained by each chromosome through the whole-genome differs significantly between BP-LTA measurements versus first visit measurements for DBP more than SBP.

A second feature of the chromosome-specific and whole-genome estimates of the BP heritability is that the latter is expected to be the sum of the chromosome-specific estimates unless there are very strong interaction effects. In these data, the chromosome sum and whole genome estimates are both 25% (EA)/45% (AA) for SBP and 31%(EA)/29%(AA) for DBP for directly genotyped SNPs (Table 3; Table 4); for the genotyped and imputed SNPs these comparisons are somewhat more discrepant at 23%(EA)/49%(AA) and 32%(EA)/37%(AA) for SBP and DBP, respectively. These observations suggest that BP variation is essentially additive in nature and largely arise from the contribution of polymorphisms (MAF ≥ 1%).

If genetic effects are approximately equal for all contributory alleles then the variation in contribution by these loci is highly dependent on their frequency (proportional to heterozygosity). Therefore, we analyzed the variance contributions by minor allele frequency (MAF) by binning each allele into five equal 10% frequency classes between 0 and 50%. These analyses (Table 5, Table 6) once again demonstrate that the heritability estimates do not significantly differ irrespective of whether we consider genotyped or genotyped and imputed SNPs or whether chromosome-sum or whole-genome estimates are considered. Moreover, the values are nearly identical to those latter obtained (Table 3,4; Figure 1, 2). Given the standard errors of the estimates, the general conclusions is that the variance explained, for both SBP and DBP, is roughly equivalent for all MAF classes with minor differences. For SBP, the estimated genetic variance for the five MAF categories ranged from 0.0 to 0.08 (SE 0.02–0.03) the highest proportion being from SNPs with MAFs 0.1-0.2; for DBP, the estimated genetic variance for the five MAF categories ranged from 0.03 to 0.073 (SE 0.02–0.03) in EA and from 0.0 to 0.28 (SE 0.14–0.13), the highest proportion being from SNPs with MAFs 0.1-0.2 in EA and with MAFs 0.2-0.3 in AA; for DBP, the estimated genetic variance for the five MAF categories ranged from 0.03 to 0.073 (SE 0.02–0.03) in EA and from 0.0 to 0.11 (SE 0.1–0.13) in AA, the highest proportion being from SNPs with MAFs 0.2-0.3 in EA and AA. The only noticeable feature is the low heritability for SBP for uncommon alleles (MAF <10%) but the substantial heritability for DBP for this same class. The more remarkable feature is the rough equivalency of heritability by MAF class despite there being a greater number of polymorphisms as MAF decreases and the consequent expectation that there are larger numbers of causal alleles at lower MAFs, alleles that are also considered to be of larger effect[35]. These results suggest that either an equivalent number of causal alleles exist irrespective of allele frequency or that causal alleles of higher frequencies explain more of the phenotypic variation.

We also investigated the likely locations of these common alleles modulating BP variation: were they preferentially located within genes in exons, introns and UTRs or were largely in the non-coding intergenic regions (Figure 3, Figure 4). Overall, irrespective of whether only genotyped SNPs or both genotyped and imputed SNPs were

analyzed, or SBP or DBP were considered, the contribution of SNPs within exons and UTRs were small, and less than 10% of the total, while intronic and intergenic regions contributed equally. Intriguingly, addition of the imputed variants increased the variation explained by exons and UTRs probably reflecting the increase in lower frequency alleles associated with functional genic regions[97] Accordingly, the fraction of alleles under 5% and 1% MAF in the genotyped and imputed data were 48% and 30%, respectively, in contrast to values of 9% and 0.2% for the genotyped-only SNPs. A second functional annotation that can be used to identify specific genomic locations associated with BP variation is the expression site of the gene (Figure 5). Do known cardiovascular and renal genes explain a significant fraction of BP heritability? Our analyses show that ~5,000 annotated cardiovascular and renal genes harboring 237,173 genotyped SNPs explained ~7-10% of SBP/DBP variance in EA. Thus, these candidate genes explain a third of the SBP/DBP variance and is greater than that expected either from their proportional gene number (25%) or the number of SNPs within these genes (~35%).

We also estimated the genetic variance captured by known genome-wide significant BP Loci (identified in EA), including markers in strong LD, and discovered that GWAS loci account for a small proportion of BP variance in EA (~1% for SBP/DBP), as expected. Finally, we used the Cardio-Metabochip SNPs, or genetic variants selected based on GWAS meta-analyses of 23 cardiovascular and metabolic traits[98]. This analysis shows that over 50% of the BP variance in EA was explained by these set of SNPs despite them accounting for only 5% of the genotyped SNPs; in AA, these markers explained 20% and 45% of the variance for SBP and DBP, respectively (Supplementary Table 4). Similarly there is a high concordance between the variance explained by each chromosome for all genotyped SNPs and those genotyped markers from the Cardio-Metabochip array (EA: r_{cor}~0.65, p= 6.91 x 10^{-4}; AA: r_{cor}~0.58, p= 1.41 x 10^{-3}). The highest proportion of genetic variance captured by chromosome for SBP in the Cardio-Metabochip is from four chromosomes in EA: chromosomes 1 (h^2~1.6%; SE ~0.006), 2 (h^2~1.6%; SE ~0.006), 3 (h^2~1.6%; SE ~0.005) and 4 (h^2~1.9%; SE ~0.006). In comparison, for AA, four different chromosomes accounted for the majority of the variance: chromosomes 2(h^2~3.7%; SE ~0.02), 12 (h^2~2.3%; SE ~0.014),

13 (h^2~1.7%; SE ~0.013) and 18 (h^2~1.7%; SE ~0.01) With respect to DBP, the most prominent contributions of genetic variation were from four chromosomes in EA: chromosomes 1 (h^2~2.2%; SE ~0.006), 2 (h^2~1.5%; SE ~0.006), 3 (h^2~2.4%; SE ~0.006) and 4 (h^2~1.7%; SE ~0.006). Where in AA, the most prominent proportion of genetic variation was from four other chromosomes: chromosomes 2 (h^2~2.5%; SE ~0.017), 5 (h^2~3%; SE ~0.017), 12 (h^2~1.4%; SE ~0.012) and 18 (h^2~2.6%; SE ~0.012)

2.4 Discussion:

There is still considerable debate as to the relative importance of common versus rare variants in the inter-individual variation of complex traits[88]. Contemporary statistical methods now allow a direct estimate of the heritability from genome-wide marker data from unrelated phenotyped individuals[35, 37]. These estimates can be compared to the classical estimates obtained from family and twin data, and these methods also allow heritability estimates from any subset of the genome to test genetic and etiological hypotheses. In this study, we explored this question for systolic (SBP) and diastolic blood pressure (DBP) by identifying the extent to which common variants can explain the amounts and distribution of SBP and DBP variation within the genome and with respect to allele frequency, coding versus non-coding DNA and sites of gene expression. We used single nucleotide polymorphism (SNP) data from the population cohort ARIC (Atherosclerosis Risk in Communities) to demonstrate that the heritability for SBP and DBP is 25% (EA)/45% (AA) and 30% (EA and AA), respectively (Table 3, Table 4). These estimates were robust and did not depend on whether we used directly genotyped SNPs or included a larger number of imputed SNPs (Table 3, Table 4) or whether we used a more stringent definition of unrelated individuals in the estimation (Supplementary Tables 1 and 2). These estimates compare favorably to estimates obtained from family[99] and twin studies[100] of adults, that vary between 42% and 39-40% for SBP and DBP, respectively. One aspect to consider is that these genome-wide heritability estimates are quite accurate and have coefficients of variation of <15%. Clearly, these estimates can be made more accurate with increasing numbers of samples. However, increasing numbers of SNPs, beyond the basic set of 600k polymorphic genotyped markers, apparently do not matter much since imputation did not affect the

estimates greatly (Table 3, Table 4). Thus, the heritability estimates are not greatly affected by numbers of SNPs.

The BP heritability estimates provided here strongly assert that the majority of inter-individual variation in BP can be attributed to polymorphisms since the directly genotyped SNPs used had 1% or greater allele frequency. Genome-wide association studies, to date, have identified ~50 BP loci with a combined effect of ~2% of the phenotypic variance[2-14]. It is well known that current GWAS are underpowered and that many BP loci remain undetected after stringent control for statistical significance. Our results suggest that the vast majority of these causal factors are indeed common (polymorphic) and remain undetected: there may indeed be up to 1000 or more BP loci leading to inter-individual phenotypic variation. The typical SBP/DBP allelic effect is ~0.05σ (where σ^2 is the phenotypic variance) so that the variance explained per SNP is $2pq(0.05\sigma)^2$ or 0.0008 or 0.08% of the variance for detected alleles (average allele frequency ~20%). If this value were typical, then 300 such loci would explain 25% of the phenotypic variance; one would infer a larger number of genes since most loci would explain a smaller variance and there is likely a statistical distribution of allelic effects.

Our analyses also shed light on some of the properties of these putative causal BP alleles. First, the additive chromosomal-level and the joint genome-level analyses provide near identical estimates suggesting that BP alleles are additive in genetic action (Figure 1,2), which is not surprising given their small effects since interaction effects will be notoriously difficult to identify at such effect sizes. The numbers of such factors are generally proportional to chromosome size although some individual chromosomes do harbor a surfeit of BP loci. The reasons for such spacial clustering are unknown and will be important to unravel. Second, the vast majority of these causal alleles reside in non-coding DNA, within introns and inter-genic DNA (Figure 3,4). These common BP alleles are under-represented within coding regions and there is a tendency for rarer alleles to be within genes. Third, causal alleles are widely distributed throughout the allele frequency spectrum with a tendency of a deficiency of rarer allele contributors to SBP and a greater presence for DBP. Fourth, mapping of susceptibility loci, associated with one or more

metabolic and cardiovascular traits, explained over half of the proportion of BP variance tagged by common SNPs. Fifth, LTA measurement, expected to describe a more accurate estimate of an individual's long-term BP value, captured additional variance as compared to single visit BP. These conclusions are quantitatively an underestimate of the additive genomic influence because it is limited to SNPs with a minor allele frequency beyond 1% and other SNPs reliably ($R^2 > 0.3$) imputed from them; rarer variants and not well-imputed SNPs are, therefore, excluded. Also, causal SNPs that were not highly correlated with the SNPs on the genotyping array or after imputation were also missed.

The above conclusions would need to be replicated in independent sets of individuals. Nevertheless, the main challenge in complex trait genetics remains the specific identification of the causal non-coding alleles and the genes they affect underlying SBP and DBP variability. Although GWAS are increasing in sample size and identifying a greater number of loci it is unlikely that they can achieve saturation identification. Our analyses suggest that the vast majority of these alleles are common, distributed across the genome, non-coding and not associated with known cardiovascular or renal genes. We propose that alternate methods be considered. One such method might be the partitioning of the genome and its variation into three segments: coding, regulatory and unknown functions. Searches for genetic variation, either by the analyses methods here or by direct GWAS, affecting phenotypes within these segments may lead to the identification of a larger number of causal factors.

In conclusion, we estimated additive genetic variation that is captured by genotyped and imputed SNPs for BP, and partitioned this variation according to chromosome, MAF, gene annotation. We provide compelling evidence that a substantial proportion of variance for blood pressure trait is explained by common SNPs, and thereby, give insights into the genetic architecture of BP trait. However, it is likely that variants other than the ones considered here and those with small effect need to be considered in addition to common SNPs.

2.5 Materials and Methods:

In the present study, phenotype data were available for 15,792 participants from the ARIC (Atherosclerosis Risk in Communities) study. ARIC is a population-based, prospective epidemiologic study of cardiovascular disease in European ancestry (EA) and African ancestry (AA) volunteers aged 45-64 years at baseline (1987-89), conducted in four US communities[101]. This analysis is focused on both European ancestry (EA) and African ancestry (AA) study participants. Cohort members completed up to four clinic examinations between 1987 and 1998, that were conducted approximately three years apart. Clinical examinations for ARIC participants assessed cardiovascular risk factors and diet, undertook various clinical and laboratory measurements, and measured numerous social variables (education, income, etc.). Genome-wide SNP genotypes in 9,747 self-identified EA and 3,207 self-identified AA subjects were obtained using the Affymetrix Gene Chip Human Mapping Array Set 6.0. The genotype data were used to exclude some samples from analyses for the following reasons: 1) discordance with previous genotype data (n=171 in EA; n=11 in AA), 2) mismatch between genotype- and phenotype-based gender (n=12 in EA), 3) previously unrecognized but suspected first or second degree relative of another participant (n=355 in EA), 4) genetic outlier as assessed by average Identity by State (IBS) statistics (>8 standard deviations along any of first 10 principal components in EIGENSTRAT with 5 iterations using PLINK) (n=308 in EA; n=336 in AA). This led to an exclusion of 846 EA and 347 AA participants, resulting in a retained dataset of 8,901 and 2,860 unrelated EA and AA subjects, respectively. In parallel, to check if shared environmental effects and/or causal variants not captured could further bias our variance estimates, we also tested a more stringent cut-off, after estimation of the pairwise genetic relationship using all autosomal markers, by excluding one of each pair of individuals with an estimated genetic relatedness of >0.025 (kinship less than 2nd cousins). This led to an exclusion of 2,833 unrelated EA and 1,123 unrelated AA subjects, respectively, resulting in a retained dataset of correspondingly 6,914 and 1,763 genetically "unrelated" EA and AA participants [90].

Blood pressures were measured by a random zero mercury sphygmomanometer following a standard protocol described elsewhere[102]. The phenotypes used for this

analysis were SBP, DBP and HTN at the first examination (visit). Subjects under antihypertensive treatments were adjusted for potential medication effects by adding 10 and 5 mm Hg to observed SBP and DBP measurements, respectively. Hypertensive participants were defined as either having SBP ≥ 140 mm Hg or DBP ≥ 90 mmHg or using an antihypertensive drug at the time of examination. We fit regression models for SBP, DBP and HTN, separately, after adjusting for the following covariates: sex, age, age squared and body mass index (BMI). The blood pressure traits analyzed here were the residuals from this regression.

To obtain the long-term averaged (LTA) BP traits, we averaged repeated BP measurements for study participants; individuals with four repeated BP measures at least 1 year apart were included in our analyses. At each study visit, we fit a linear regression model by using covariate adjustment in a manner identical to what has been done in first visit to generate visit-specific BP residuals. These residual values were subsequently averaged over all available visits, and the final averaged residual was the LTA trait analyzed (formulated LTA SBP, LTA DBP).

Quality control on genotypes has been described elsewhere[101]. Nevertheless, after pruning 308 EA and 334 AA individual participants from the raw data as genetic outliers, we performed imputation in EA and AA subjects to the 1000 Genomes reference panel. Imputation in EA participants was performed using a hidden Markov model as implemented in the software packages MaCH (v1.0.16) and Minimac (v4.6) [103]. Each chromosome was first phased to estimate haplotypes using MaCH and then the phased haplotypes were used, along with the 1000G Interim Phase I Haplotype reference panel, resulting in >37 million SNPs for imputation using Minimac. Imputation in AA participants was completed with IMPUTE version 2 (v2.1.2) and involved a two-step procedure for each chromosome: phasing to generate haplotype followed by imputation using similar reference panels. Measured SNPs used for imputation were restricted to those with the following features: MAF >0.5%, missing data per SNP < 5%, and Hardy-Weinberg equilibrium (HWE) $P > 10^{-5}$. Of the 839,048 genotyped markers, 656,362 and 772,638 genotyped autosomal SNPs in EA and AA, respectively, passed the initial

quality filters and were used for imputation. In this study, sex chromosomes were excluded from the analysis of blood pressure.

Descriptive statistics and regression of the phenotype on age, age-squared, sex, BMI and PCs, were carried out using R version 2.6.0 (The R Foundation for Statistical Computing). The statistical method utilized here is detailed in Yang *et al*[37]. Briefly, a genetic relationship matrix (GRM) for each pair of individuals was calculated as the sum of the products of SNP coefficients between two individuals scaled by the SNP heterozygosity for all genotyped and imputed SNPs across the genome. Subsequently, the GRM was used in a linear mixed model to estimate the variance captured by all the autosomal SNPs via restricted maximum likelihood analysis. This was expressed as a linear function of the total amount of the additive effects due to SNPs associated with causal markers and residual effects: $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \sum_{i=1}^{p} \boldsymbol{\mathcal{V}}_{A_i} + \boldsymbol{\varepsilon}$ where $\boldsymbol{y}$ is an $\mathbf{N} \times \mathbf{1}$ vector of systolic or diastolic blood pressure measurements with $\mathbf{N}$ being the sample size, $\boldsymbol{\beta}$ a vector for fixed effects such as sex, age, age squared and BMI, $\boldsymbol{X}$ the genotype incidence matrix relating to individuals, $\boldsymbol{\mathcal{V}}_{A}$ a vector of random additive genetic effects partitioned on aggregate of all autosomal SNPs estimated from whole-genome markers ($\boldsymbol{p}$ = 1; $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \sum_{i=1}^{1} \boldsymbol{\mathcal{V}}_{A_i} + \boldsymbol{\varepsilon}$). The proportion of variance explained by whole-genome markers is the narrow-sense heritability, i.e., $h^2 = \sigma_A^2 / \sigma_P^2$ where σ_P^2 is the total phenotypic variance.

The variance estimate from the entire genome can also be partitioned into non-overlapping subsets of SNPs defined by any specific criteria: if $\boldsymbol{p}$ such classes are considered then $\boldsymbol{var}(\boldsymbol{\mathcal{V}}_{A}) = \sum_{i=1}^{p} \boldsymbol{M}_{A_i} + \boldsymbol{\sigma}_A^2$ where $\mathbf{M}_A$ is the GRM estimated from the whole-genome genotyped or imputed markers, $\boldsymbol{\sigma}_A^2$ is the variance explained by all SNPs, and $\boldsymbol{\varepsilon}$ is a vector of random error effects. The specific partitions we considered were chromosomal number ($\boldsymbol{p}$ = 22; autosomes), minor allele frequency (MAF) class ($\boldsymbol{p}$ = 5; 0-50% in 10% intervals), functional annotation based on location ($\boldsymbol{p}$ = 4; UTRs, coding, intronic, intergenic), and functional annotation based on gene expression related to known cardiovascular or renal associated-genes[104]and genes annotated by traits and Cardiovascular and Renal Genes related from EMBL-EBI[105, 106]. Finally, the variance

estimate from the entire genome was partitioned by blood pressure loci (using NHGRI GWAS catalog[107] - http://www.genome.gov/gwastudies - and literature SNPs (3,5,6,7,8,10,11,12,13,14)) including markers with strong LD (LD ≥ 0.8) using SNAP[108] and by Cardio-Metabochip SNPs[98]respectively.

Acknowledgements:

We thank the participants and staff of the ARIC study for their important contributions to this study. This work was supported by grant HL086694 from the National Heart, Lung, and Blood Institute (NHLBI) to A.C.. The Atherosclerosis Risk in Communities Study was carried out as a collaborative study supported by NHLBI contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C), R01HL087641, R01HL59367 and R01HL086694; National Human Genome Research Institute contract U01HG004402; and National Institutes of Health contract HHSN268200625226C. Infrastructure was partly supported by Grant Number UL1RR025005, a component of the National Institutes of Health and NIH Roadmap for Medical Research.

Conflict of Interest Disclosure: None

Legends to Figures:

Figure 1: Estimates of the variance explained by SNPS by chromosome (h^2_c) for SBP (red) and DBP (blue) by joint analysis of 8,901 EA individuals. The trait analyzed is first-visit SBP and DBP with analyses for genotyped SNPs only.

Figure 2: Estimates of the variance explained by SNPS by chromosome (h^2_c) for SBP (red) and DBP (blue) by joint analysis of 2,860 AA individuals. The trait analyzed is first-visit SBP and DBP with analyses for genotyped SNPs only.

Figure 3: Estimates of the variance explained of SBP (red) and DBP (blue) by functional annotation class (UTR, exon, intron, intergenic) by joint analysis using genotyped or all (genotyped and imputed) SNPs in 8,901 EA individuals.

Figure 4: Estimates of the variance explained of SBP (red) and DBP (blue) by functional annotation class (UTR, exon, intron, intergenic) by joint analysis using genotyped or all (genotyped and imputed) SNPs in 2,860 AA individuals.

Figure 5: Proportion of the genetic variance explained by SNPs for SBP and DBP for genotyped SNPs within known gene annotations (Cardio-Metabochip=Markers associated with metabolic traits; Cardio-Metabochip.CV=Markers associated with cardiovascular traits; Cardio/Renal Genes= Cardiovascular and Renal SNPs; Non-Annotated=Marker with no known association with cardiovascular or renal tissue).

Figure 1.

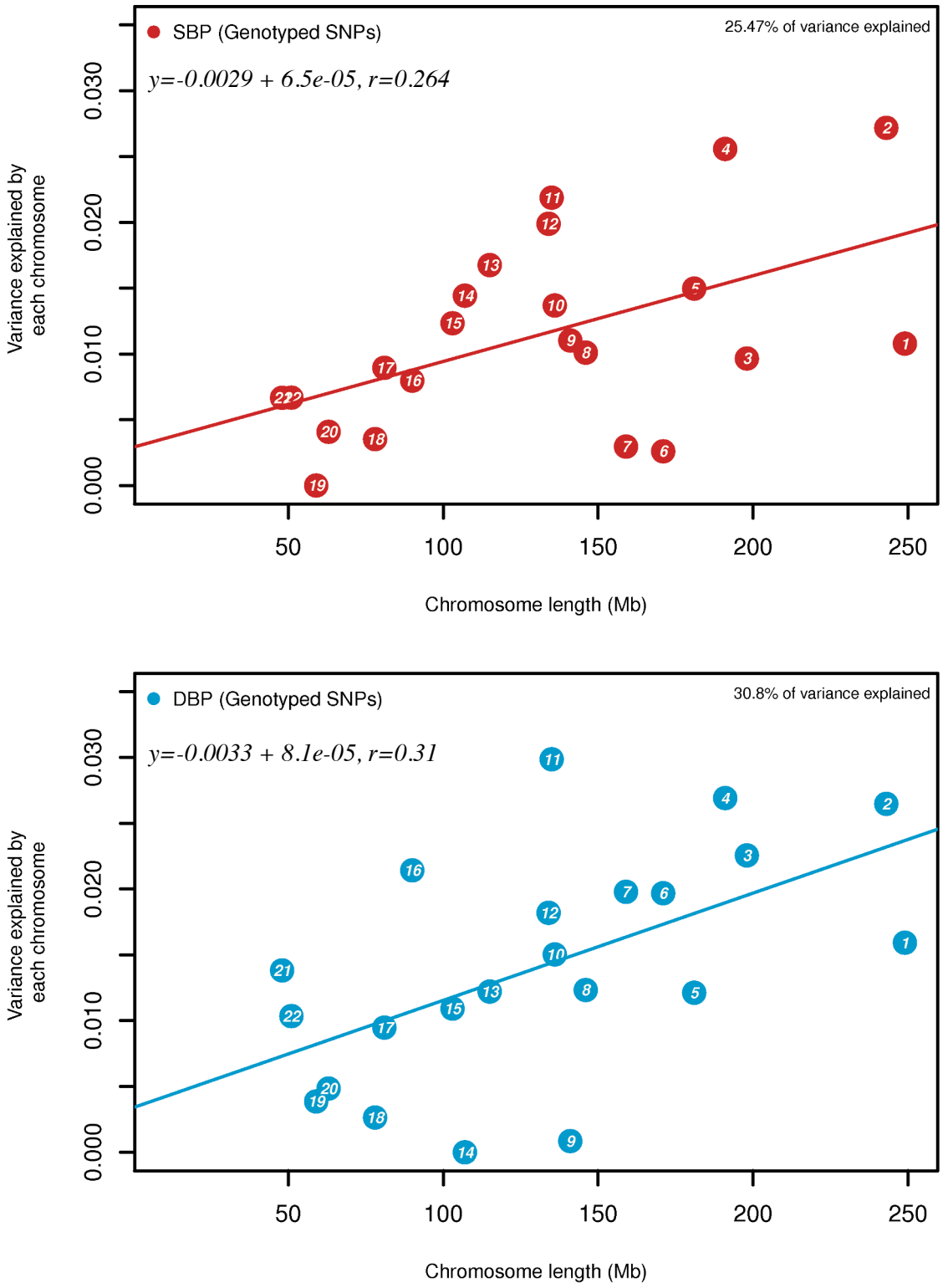
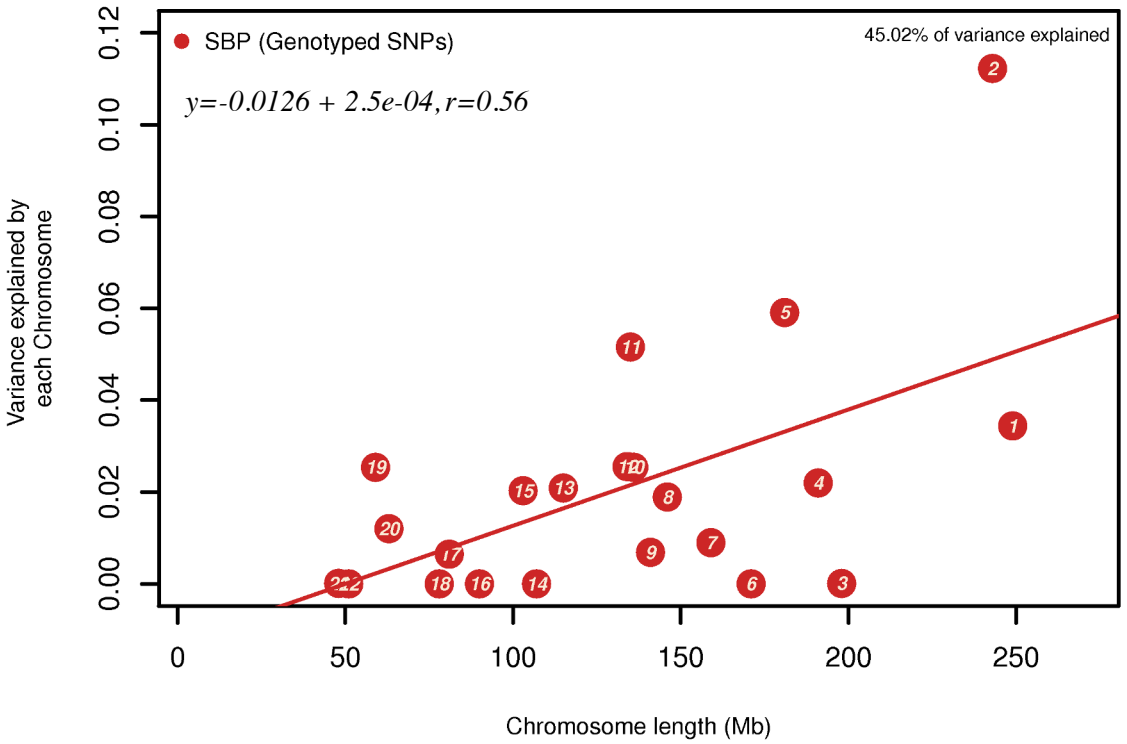

SBP (Genotyped SNPs)
25.47% of variance explained
y=-0.0029 + 6.5e-05, r=0.264
Variance explained by each chromosome
0.000
0.010
0.020
0.030
50
100
150
200
250
Chromosome length (Mb)
DBP (Genotyped SNPs)
30.8% of variance explained
y=-0.0033 + 8.1e-05, r=0.31
Variance explained by each chromosome
0.000
0.010
0.020
0.030
50
100
150
200
250
Chromosome length (Mb)

Figure 2.


SBP (Genotyped SNPs)
45.02% of variance explained
y=-0.0126 + 2.5e-04, r=0.56
Variance explained by each Chromosome
Chromosome length (Mb)

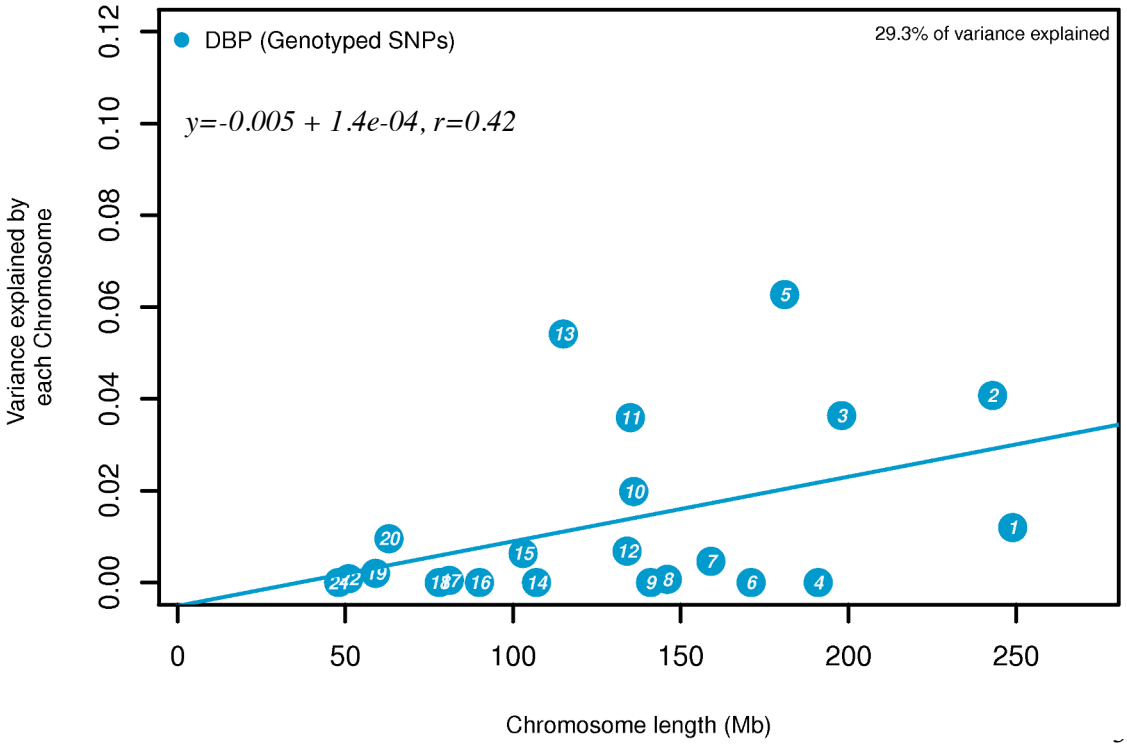

DBP (Genotyped SNPs)
29.3% of variance explained
y=-0.005 + 1.4e-04, r=0.42
Variance explained by each Chromosome
Chromosome length (Mb)

Figure 3.

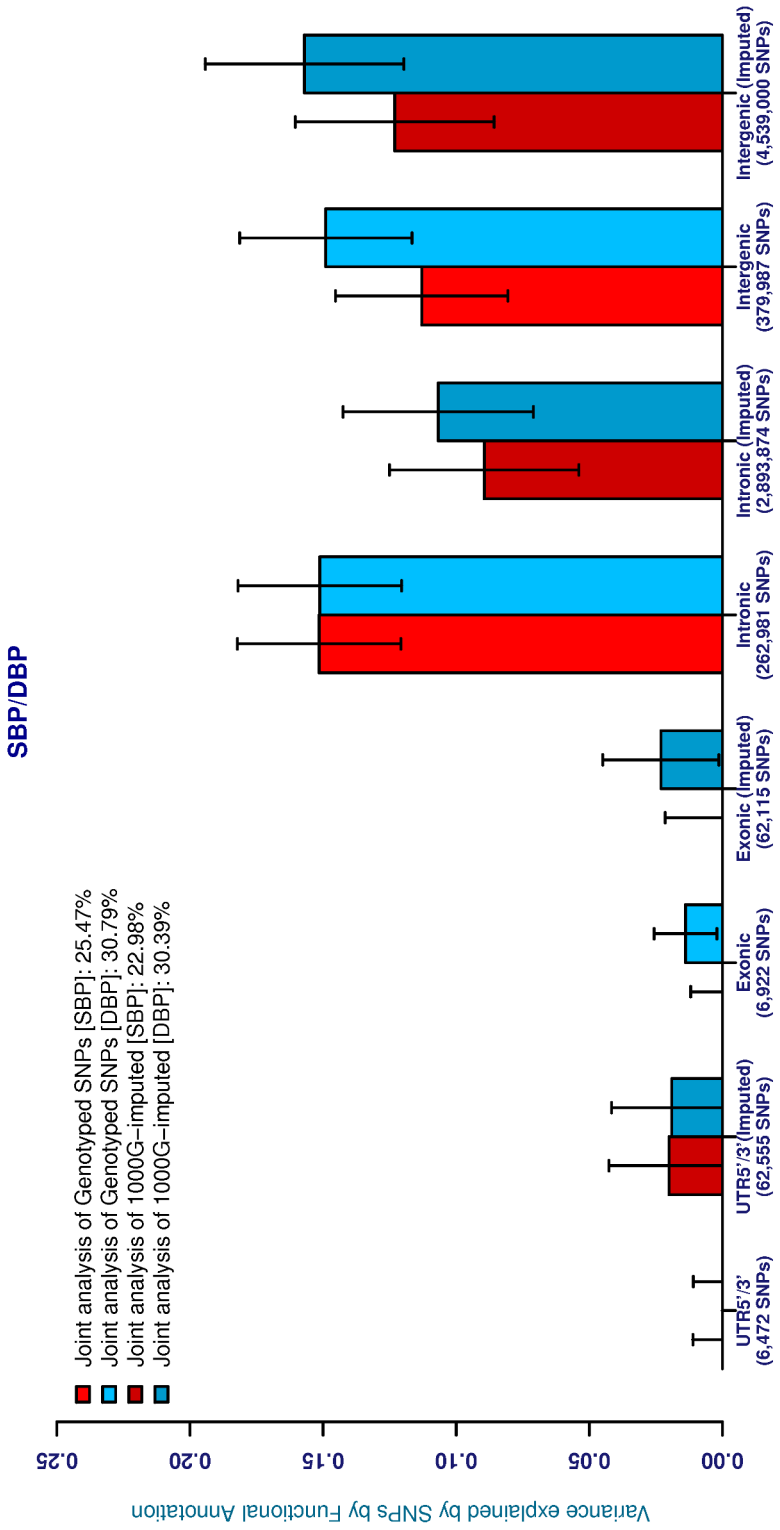

SBP/DBP
Joint analysis of Genotyped SNPs [SBP]: 25.47%
Joint analysis of Genotyped SNPs [DBP]: 30.79%
Joint analysis of 1000G–imputed [SBP]: 22.98%
Joint analysis of 1000G–imputed [DBP]: 30.39%
Variance explained by SNPs by Functional Annotation
0.00
0.05
0.10
0.15
0.20
0.25
UTR5'/3' (6,472 SNPs)
UTR5'/3' (Imputed) (62,555 SNPs)
Exonic (6,922 SNPs)
Exonic (Imputed) (62,115 SNPs)
Intronic (262,981 SNPs)
Intronic (Imputed) (2,893,874 SNPs)
Intergenic (379,987 SNPs)
Intergenic (Imputed) (4,539,000 SNPs)

Figure 4.

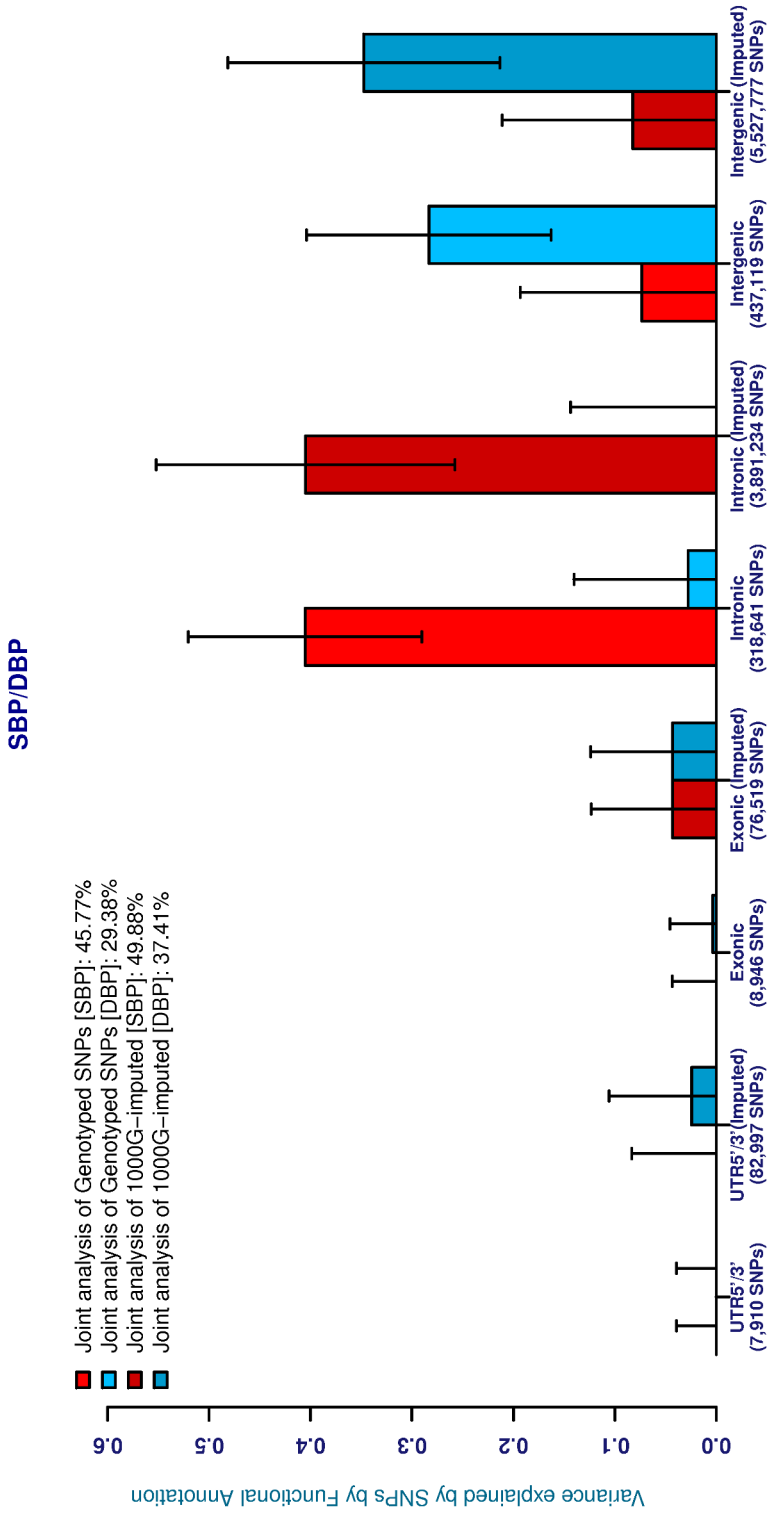
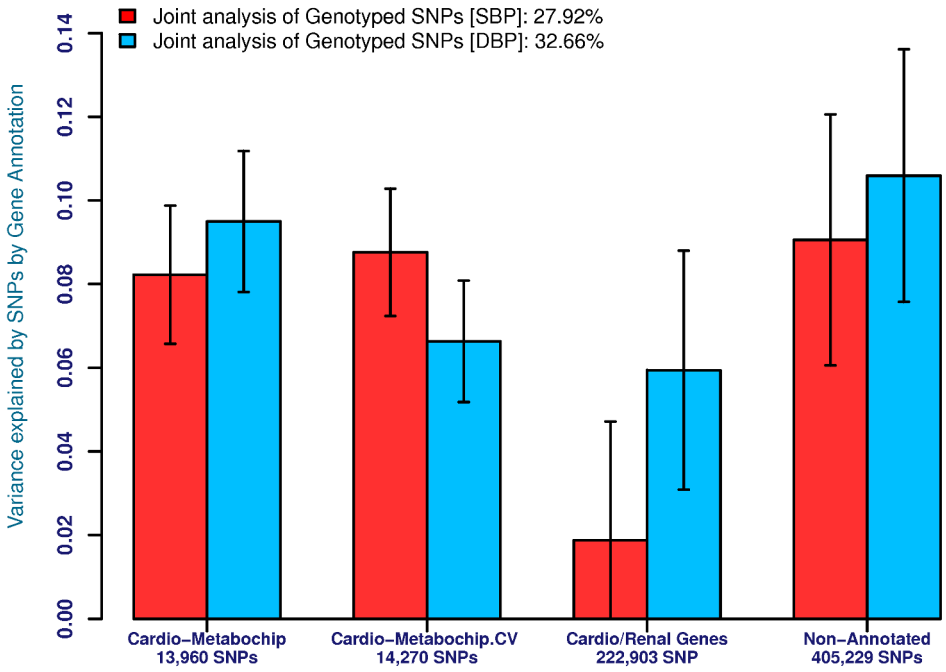

SBP/DBP
Joint analysis of Genotyped SNPs [SBP]: 45.77%
Joint analysis of Genotyped SNPs [DBP]: 29.38%
Joint analysis of 1000G–imputed [SBP]: 49.88%
Joint analysis of 1000G–imputed [DBP]: 37.41%
Variance explained by SNPs by Functional Annotation
0.0
0.1
0.2
0.3
0.4
0.5
0.6
UTR5'/3' (7,910 SNPs)
UTR5'/3' (Imputed) (82,997 SNPs)
Exonic (8,946 SNPs)
Exonic (Imputed) (76,519 SNPs)
Intronic (318,641 SNPs)
Intronic (Imputed) (3,891,234 SNPs)
Intergenic (437,119 SNPs)
Intergenic (Imputed) (5,527,777 SNPs)

Figure 5.


Joint analysis of Genotyped SNPs [SBP]: 27.92%
Joint analysis of Genotyped SNPs [DBP]: 32.66%
Variance explained by SNPs by Gene Annotation
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Cardio−Metabochip
13,960 SNPs
Cardio−Metabochip.CV
14,270 SNPs
Cardio/Renal Genes
222,903 SNP
Non−Annotated
405,229 SNPs

Legends to Tables:

Table 1: Summary statistics of ARIC European ancestry (EA) subjects used.

Table 2: Summary statistics of ARIC African ancestry (AA) subjects used.

Table 3: Proportion of the genetic variance explained by each chromosome and the whole genome using 8,901 EA individuals. The trait analyzed is first-visit SBP and DBP with separate analyses for genotyped SNPs only and all genotyped and imputed SNPs. The heritability (h^2), its standard error (SE) and significance value (P) are shown.

Table 4: Proportion of the genetic variance explained by each chromosome and the whole genome using 2,860 AA individuals. The trait analyzed is first-visit SBP and DBP with separate analyses for genotyped SNPs only and all genotyped and imputed SNPs. The heritability (h^2), its standard error (SE) and significance value (P) are shown.

Table 5: Proportion of the genetic variance explained as a function of minor allele frequency (MAF) class of 8,901 EA individuals. The trait analyzed is first-visit SBP and DBP with separate analyses for genotyped SNPs only and all genotyped and imputed SNPs. The heritability (h^2), its standard error (SE) and significance value (P) are shown.

Table 6: Proportion of the genetic variance explained as a function of minor allele frequency (MAF) class of 2,860 AA individuals. The trait analyzed is first-visit SBP and DBP with separate analyses for genotyped SNPs only and all genotyped and imputed SNPs. The heritability (h^2), its standard error (SE) and significance value (P) are shown.

ARIC European ancestry (EA) subjects					
Trait	**N**	**Mean**	**SE**	**Min**	**Max**
SBP	8901	121.29	19.44	61	221
DBP	8901	73.54	11.52	12	139
AGE	8901	54.27	5.70	44	66
BMI	8901	26.97	4.83	14.38	56.26
Trait	**Males**	**Mean**	**SE**	**Min**	**Max**
SBP	4197	122.99	18.28	61	203
DBP	4197	75.59	11.37	12	130
AGE	4197	54.67	5.70	44	66
BMI	4197	27.46	3.96	17.21	56.26
Trait	**Females**	**Mean**	**SE**	**Min**	**Max**
SBP	4704	119.76	20.30	72	221
DBP	4704	71.71	11.34	27	139
AGE	4704	53.90	5.67	44	66
BMI	4704	26.54	5.46	14.38	55.20

ARIC European ancestry (EA) unrelated subjects					
Trait	**N**	**Mean**	**SE**	**Min**	**Max**
SBP	6914	120.81	19.12	61	207
DBP	6914	73.50	11.40	12	128
AGE	6914	54.20	5.71	44	66
BMI	6914	26.84	4.70	14.91	53.87
Trait	**Males**	**Mean**	**SE**	**Min**	**Max**
SBP	3275	122.89	18.14	61	201
DBP	3275	75.64	11.26	12	123
AGE	3275	54.64	5.69	44	66
BMI	3275	27.45	3.95	17.21	53.87
Trait	**Females**	**Mean**	**SE**	**Min**	**Max**
SBP	3639	118.93	19.77	72	207
DBP	3639	71.57	11.18	27	128
AGE	3639	53.81	5.70	44	66
BMI	3639	26.29	5.23	14.91	51.36

Table 1.

ARIC African ancestry (AA) subjects					
Trait	**N**	**Mean**	**SE**	**Min**	**Max**
SBP	2871	111.93	23.49	73	257
DBP	2871	84.01	13.70	34	152
AGE	2871	24.32	5.76	44	66
BMI	2871	25.42	6.06	14.20	59.33
Trait	**Males**	**Mean**	**SE**	**Min**	**Max**
SBP	1068	135.52	23.14	88	241
DBP	1068	86.29	14.19	51	149
AGE	1068	53.71	5.94	44	66
BMI	1068	28.01	4.84	15.46	54.40
Trait	**Females**	**Mean**	**SE**	**Min**	**Max**
SBP	1803	134.24	23.69	73	257
DBP	1803	82.64	13.35	34	152
AGE	1803	53.19	5.64	44	65
BMI	1803	30.69	6.47	14.20	59.33

ARIC African ancestry (AA) unrelated subjects					
Trait	**N**	**Mean**	**SE**	**Min**	**Max**
SBP	1737	135.02	23.57	84	257
DBP	1737	84.16	13.87	34	152
AGE	1737	53.35	5.76	44	66
BMI	1737	29.66	6.01	15.46	59.33
Trait	**Males**	**Mean**	**SE**	**Min**	**Max**
SBP	676	135.28	22.87	88	235
DBP	676	86.11	14.23	51	138
AGE	676	53.68	5.97	44	66
BMI	676	28.21	4.85	15.46	51.16
Trait	**Females**	**Mean**	**SE**	**Min**	**Max**
SBP	1061	134.86	24.01	84	257
DBP	1061	82.92	13.51	34	152
AGE	1061	53.15	5.61	44	65
BMI	1061	30.59	6.48	16.24	59.33

Table 2.

EA cases (N=8,901)		SBP						DBP					
SNPs		Genotyped			Genotyped + imputed			Genotyped			Genotyped + imputed		
Chr	L (Mb)	h^2	SE	*P*	h^2	SE	*P*	h^2	SE	*P*	h^2	SE	*P*
1	249.25	0.0108	0.0110		0.0111	0.0117		0.0159	0.0110		0.0080	0.0116	
2	243.20	0.0272	0.0113		0.0289	0.0123		0.0265	0.0113		0.0290	0.0124	
3	198.02	0.0097	0.0100		0.0087	0.0102		0.0226	0.0103		0.0200	0.0104	
4	191.15	0.0256	0.0105		0.0254	0.0113		0.0269	0.0105		0.0240	0.0112	
5	180.92	0.0150	0.0098		0.0128	0.0101		0.0121	0.0096		0.0121	0.0100	
6	171.12	0.0026	0.0092		0.0000	0.0090		0.0197	0.0099		0.0184	0.0095	
7	159.14	0.0030	0.0088		0.0000	0.0095		0.0198	0.0098		0.0238	0.0108	
8	146.36	0.0101	0.0084		0.0111	0.0089		0.0123	0.0088		0.0088	0.0089	
9	141.21	0.0110	0.0086		0.0094	0.0092		0.0008	0.0081		0.0031	0.0090	
10	135.53	0.0137	0.0094		0.0116	0.0099		0.0150	0.0093		0.0136	0.0098	
11	135.01	0.0219	0.0088		0.0201	0.0091		0.0299	0.0094		0.0305	0.0101	
12	133.85	0.0199	0.0092		0.0231	0.0101		0.0182	0.0090		0.0226	0.0099	
13	115.17	0.0167	0.0082		0.0135	0.0083		0.0122	0.0080		0.0094	0.0080	
14	107.35	0.0144	0.0078		0.0173	0.0083		0.0000	0.0069		0.0000	0.0073	
15	102.53	0.0123	0.0078		0.0093	0.0075		0.0109	0.0075		0.0105	0.0074	
16	90.35	0.0080	0.0076		0.0052	0.0083		0.0214	0.0083		0.0268	0.0094	
17	81.20	0.0089	0.0069		0.0091	0.0073		0.0095	0.0071		0.0082	0.0072	
18	78.08	0.0035	0.0071		0.0035	0.0076		0.0026	0.0069		0.0025	0.0075	
19	59.13	0.0000	0.0053		0.0000	0.0057		0.0039	0.0055		0.0059	0.0060	
20	63.03	0.0041	0.0065		0.0045	0.0072		0.0048	0.0068		0.0124	0.0078	
21	48.13	0.0067	0.0054		0.0074	0.0059		0.0138	0.0060		0.0139	0.0063	
22	51.30	0.0067	0.0054		0.0021	0.0055		0.0103	0.0056		0.0124	0.0063	
Total	2,881.03	***0.2516***			***0.2340***			***0.3093***			***0.3157***		
Combined		***0.2547***	***0.0377***	***$1.0x10^{-13}$***	***0.2299***	***0.0395***	***$9.0x10^{-11}$***	***0.3080***	***0.036***	***$2.0x10^{-16}$***	***0.304***	***0.038***	***$2.0x10^{-16}$***

Table 3.

AA cases (N=2,860)		SBP						DBP					
SNPs		Genotyped			Genotyped + imputed			Genotyped			Genotyped + imputed		
Chr	L (Mb)	h^2	SE	*P*	h^2	SE	*P*	h^2	SE	*P*	h^2	SE	*P*
1	249.25	0.0344	0.0338		0.0261	0.0381		0.0120	0.0281		0.0041	0.0358	
2	243.20	0.1122	0.0455		0.1305	0.0514		0.0407	0.0420		0.0420	0.0488	
3	198.02	0.0001	0.0300		0.0000	0.0391		0.0364	0.0323		0.0390	0.0380	
4	191.15	0.0220	0.0381		0.0166	0.0419		0.0000	0.0380		0.0000	0.0448	
5	180.92	0.0591	0.0401		0.0717	0.0453		0.0627	0.0389		0.0874	0.0450	
6	171.12	0.0000	0.0301		0.0000	0.0370		0.0000	0.0278		0.0000	0.0365	
7	159.14	0.0090	0.0309		0.0306	0.0377		0.0046	0.0313		0.0000	0.0359	
8	146.36	0.0189	0.0270		0.0539	0.0384		0.0007	0.0253		0.0361	0.0376	
9	141.21	0.0069	0.0248		0.0186	0.0343		0.0000	0.0213		0.0102	0.0330	
10	135.53	0.0254	0.0351		0.0195	0.0375		0.0198	0.0326		0.0364	0.0379	
11	135.01	0.0516	0.0339		0.0645	0.0342		0.0359	0.0302		0.0517	0.0334	
12	133.85	0.0255	0.0317		0.0214	0.0314		0.0068	0.0242		0.0048	0.0266	
13	115.17	0.0209	0.0314		0.0000	0.0317		0.0541	0.0338		0.0359	0.0344	
14	107.35	0.0000	0.0212		0.0000	0.0282		0.0000	0.0231		0.0000	0.0291	
15	102.53	0.0203	0.0250		0.0048	0.0238		0.0063	0.0196		0.0059	0.0263	
16	90.35	0.0000	0.0290		0.0042	0.0316		0.0000	0.0277		0.0000	0.0303	
17	81.20	0.0065	0.0219		0.0034	0.0256		0.0005	0.0206		0.0066	0.0242	
18	78.08	0.0000	0.0275		0.0000	0.0300		0.0000	0.0250		0.0046	0.0303	
19	59.13	0.0254	0.0219		0.0270	0.0210		0.0020	0.0137		0.0076	0.0178	
20	63.03	0.0120	0.0230		0.0101	0.0257		0.0095	0.0222		0.0120	0.0257	
21	48.13	0.0001	0.0121		0.0000	0.0159		0.0000	0.0146		0.0000	0.0195	
22	51.30	0.0000	0.0212		0.0016	0.0211		0.0008	0.0163		0.0061	0.0202	
Total	2,881.03	***0.4501***			***0.5045***			***0.293***			***0.391***		
Combined		***0.4577***	***0.1168***	***$1.1x10^{-5}$***	***0.4987***	***0.1259***	***$7.9x10^{-6}$***	***0.2939***	***0.1071***	***$7.0x10^{-5}$***	***0.3741***	***0.121***	***$3.9x10^{-5}$***

Table 4.

EA cases (N=8,901)	SBP						DBP					
SNPs	**Genotyped**			**Genotyped + imputed**			**Genotyped**			**Genotyped + imputed**		
MAF	***h^2***	***(s.e)***	***P***	***h^2***	***(s.e)***	***P***	***h^2***	***(s.e)***	***P***	***h^2***	***(s.e)***	***P***
0 - 0.1	0.000	0.028		0.022	0.035		0.062	0.029		0.094	0.038	
0.1 - 0.2	0.078	0.029		0.065	0.027		0.067	0.029		0.069	0.028	
0.2 - 0.3	0.032	0.028		0.055	0.026		0.073	0.029		0.071	0.027	
0.3 - 0.4	0.067	0.028		0.032	0.025		0.071	0.028		0.051	0.025	
0.4 - 0.5	0.058	0.024		0.033	0.022		0.033	0.024		0.011	0.022	
Total	***0.2357***			***0.2077***			***0.3062***			***0.2955***		
Combined	***0.2547***	***0.0377***	***$1.0x10^{-13}$***	***0.2295***	***0.0395***	***$1.0x10^{-10}$***	***0.3079***	***0.0367***	***$2.0x10^{-15}$***	***0.3039***	***0.0387***	***$2.0x10^{-16}$***

Table 5.

AA cases (N=2,860)	SBP						DBP					
SNPs	**Genotyped**			**Genotyped + imputed**			**Genotyped**			**Genotyped + imputed**		
MAF	***h^2***	***(s.e)***	***P***	***h^2***	***(s.e)***	***P***	***h^2***	***(s.e)***	***P***	***h^2***	***(s.e)***	***P***
0 - 0.1	0.0048	0.1045		0.2422	0.1562		0.0000	0.1061		0.1040	0.1525	
0.1 - 0.2	0.0000	0.1418		0.0230	0.1112		0.0971	0.1393		0.0732	0.1102	
0.2 - 0.3	0.2899	0.1315		0.0742	0.0948		0.1077	0.1310		0.1010	0.0970	
0.3 - 0.4	0.1142	0.1224		0.0229	0.0842		0.0200	0.1183		0.0000	0.0835	
0.4 - 0.5	0.0430	0.1075		0.1159	0.0745		0.0960	0.1099		0.0631	0.0735	
Total	***0.4519***			***0.4781***			***0.3208***			***0.3414***		
Combined	***04577***	***0.1168***	***$1.1x10^{-5}$***	***0.4987***	***0.1259***	***$7.9x10^{-6}$***	***0.2939***	***0.1071***	***$7.0x10^{-5}$***	***0.3741***	***0.121***	***$3.9x10^{-5}$***

Table 6.

Legend to Supplementary Table:

Table S1: Proportion of the genetic variance explained by each chromosome and the whole genome using 6,914 EA unrelated individuals. The trait analyzed is first-visit SBP and DBP with separate analyses for genotyped SNPs only and all genotyped and imputed SNPs. The heritability (h^2), its standard error (SE) and significance value (P) are shown.

Table S2: Proportion of the genetic variance explained by each chromosome and the whole genome using 1,763 AA unrelated individuals. The trait analyzed is first-visit SBP and DBP with separate analyses for genotyped SNPs only and all genotyped and imputed SNPs. The heritability (h^2), its standard error (SE) and significance value (P) are shown.

Table S3: Proportion of the genetic variance explained by each chromosome using 8,874 EA and 2,749 AA unrelated individuals respectively. The trait analyzed is long-term average SBP and DBP for genotyped SNPs only. The heritability (h^2), its standard error (SE) and significance value (P) are shown.

Table S4: Proportion of the genetic variance explained by each chromosome using 8,901 EA and 2,860 AA unrelated individuals respectively. The trait analyzed is first-visit SBP and DBP for Cardio-Metabochip genotyped SNPs only. The heritability (h^2), its standard error (SE) and significance value (P) are shown

Table S5: Estimates of the SBP and DBP heritability from classical family and twin studies; the trait measured, relatives utilized, sample size, heritability estimate and citation to the study are provided.

TRAIT (N=6,914)		SBP						DBP					
SNPs		Genotyped			Genotyped + imputed			Genotyped			Genotyped + imputed		
Chr	L (Mb)	h^2	SE	*P*	h^2	SE	*P*	h^2	SE	*P*	h^2	SE	*P*
1	249.25	0.0091	0.0139		0.0113	0.0148		0.0127	0.0137		0.0061	0.0144	
2	243.20	0.0387	0.0148		0.0397	0.0162		0.0302	0.0146		0.0331	0.0160	
3	198.02	0.0104	0.0128		0.0043	0.0127		0.0266	0.0133		0.0259	0.0137	
4	191.15	0.0348	0.0136		0.0317	0.0146		0.0477	0.0141		0.0419	0.0150	
5	180.92	0.0049	0.0119		0.0071	0.0122		0.0038	0.0118		0.0054	0.0125	
6	171.12	0.0000	0.0117		0.0000	0.0113		0.0147	0.0122		0.0204	0.0121	
7	159.14	0.0054	0.0116		0.0098	0.0125		0.0135	0.0122		0.0169	0.0133	
8	146.36	0.0080	0.0103		0.0169	0.0118		0.0083	0.0105		0.0105	0.0114	
9	141.21	0.0189	0.0113		0.0131	0.0119		0.0012	0.0103		0.0000	0.0113	
10	135.53	0.0108	0.0118		0.0085	0.0123		0.0316	0.0126		0.0247	0.0130	
11	135.01	0.0196	0.0111		0.0145	0.0107		0.0186	0.0115		0.0169	0.0119	
12	133.85	0.0256	0.0120		0.0261	0.0130		0.0145	0.0112		0.0222	0.0124	
13	115.17	0.0219	0.0105		0.0153	0.0105		0.0238	0.0106		0.0207	0.0108	
14	107.35	0.0156	0.0098		0.0163	0.0102		0.0020	0.0090		0.0000	0.0094	
15	102.53	0.0039	0.0093		0.0005	0.0084		0.0140	0.0095		0.0132	0.0094	
16	90.35	0.0048	0.0094		0.0000	0.0105		0.0218	0.0103		0.0261	0.0116	
17	81.20	0.0079	0.0087		0.0080	0.0090		0.0062	0.0086		0.0042	0.0084	
18	78.08	0.0000	0.0091		0.0000	0.0097		0.0000	0.0090		0.0000	0.0097	
19	59.13	0.0003	0.0066		0.0000	0.0073		0.0040	0.0070		0.0025	0.0075	
20	63.03	0.0012	0.0081		0.0016	0.0087		0.0022	0.0081		0.0059	0.0089	
21	48.13	0.0052	0.0068		0.0085	0.0075		0.0108	0.0072		0.0136	0.0078	
22	51.30	0.0074	0.0070		0.0053	0.0074		0.0143	0.0074		0.0209	0.0085	
Total	2,881.03	***0.2542***			***0.2386***			***0.3226***			***0.3310***		
Combined		***0.2528***	***0.0487***	***$2.0x10^{-8}$***	***0.2319***	***0.0511***	***$7.0x10^{-7}$***	***0.3158***	***0.0470***	***$2.0x10^{-15}$***	***0.3198***	***0.0498***	***$3.0x10^{-14}$***

Table S1.

TRAIT (N=1,737)		SBP						DBP					
SNPs		Genotyped			Genotyped + imputed			Genotyped			Genotyped + imputed		
Chr	L (Mb)	h^2	SE	*P*	h^2	SE	*P*	h^2	SE	*P*	h^2	SE	*P*
1	249.25	0.0000	0.0557		0.0000	0.0731		0.0000	0.0513		0.0029	0.0648	
2	243.20	0.0282	0.0698		0.1053	0.0870		0.0000	0.0692		0.0275	0.0843	
3	198.02	0.0000	0.0570		0.0000	0.0700		0.0535	0.0543		0.0681	0.0640	
4	191.15	0.0000	0.0638		0.0000	0.0734		0.0000	0.0586		0.0000	0.0683	
5	180.92	0.0220	0.0570		0.0000	0.0658		0.0213	0.0480		0.0175	0.0623	
6	171.12	0.0000	0.0587		0.0000	0.0673		0.0039	0.0429		0.0000	0.0552	
7	159.14	0.0151	0.0567		0.0000	0.0633		0.0239	0.0563		0.0000	0.0631	
8	146.36	0.0519	0.0527		0.0801	0.0637		0.0123	0.0512		0.0498	0.0609	
9	141.21	0.0015	0.0353		0.0143	0.0533		0.0025	0.0333		0.0188	0.0519	
10	135.53	0.0000	0.0578		0.0000	0.0625		0.0000	0.0382		0.0031	0.0505	
11	135.01	0.0248	0.0530		0.0603	0.0516		0.0322	0.0476		0.0371	0.0496	
12	133.85	0.0635	0.0510		0.0277	0.0474		0.0197	0.0417		0.0153	0.0467	
13	115.17	0.0342	0.0485		0.0139	0.0522		0.0638	0.0534		0.0137	0.0535	
14	107.35	0.0661	0.0491		0.0229	0.0444		0.0141	0.0342		0.0036	0.0419	
15	102.53	0.0314	0.0396		0.0188	0.0412		0.0186	0.0332		0.0197	0.0419	
16	90.35	0.0000	0.0305		0.0000	0.0400		0.0000	0.0279		0.0000	0.0414	
17	81.20	0.0082	0.0377		0.0000	0.0396		0.0248	0.0409		0.0000	0.0412	
18	78.08	0.0000	0.0407		0.0000	0.0466		0.0000	0.0359		0.0157	0.0464	
19	59.13	0.0162	0.0332		0.0029	0.0282		0.0292	0.0339		0.0535	0.0351	
20	63.03	0.0120	0.0396		0.0027	0.0458		0.0008	0.0390		0.0082	0.0475	
21	48.13	0.0064	0.0245		0.0069	0.0321		0.0194	0.0291		0.0026	0.0260	
22	51.30	0.0570	0.0406		0.0124	0.0360		0.0396	0.0376		0.0127	0.0338	
Total	2,881.03	***0.4386***			***0.3679***			***0.3796***			***0.3698***		
Combined		***0.4246***	***0.2051***	***$2.0x10^{-3}$***	***0.4095***	***0.2361***	***$4.0x10^{-3}$***	***0.3861***	***0.1885***	***$3.0x10^{-3}$***	***0.3785***	***0.2152***	***$1.0x10^{-3}$***

Table S2.

TRAIT		SBP - LTA						DBP - LTA					
SNPs		EA cases (N=8,874)			AA cases (N=2,749)			EA cases (N=8,874)			AA cases (N=2,749)		
Chr	L (Mb)	h^2	SE	*P*	h^2	SE	*P*	h^2	SE	*P*	h^2	SE	*P*
1	249.25	0.0013	0.0111		0.0459	0.0402		0.0102	0.0114		0.0469	0.0421	
2	243.20	0.0242	0.0118		0.0698	0.0448		0.0107	0.0115		0.0188	0.0384	
3	198.02	0.0157	0.0103		0.0000	0.0302		0.0268	0.0108		0.0814	0.0388	
4	191.15	0.0390	0.0115		0.0471	0.0417		0.0230	0.0108		0.0011	0.0352	
5	180.92	0.0126	0.0102		0.0690	0.0426		0.0238	0.0105		0.0457	0.0369	
6	171.12	0.0000	0.0094		0.0000	0.0313		0.0024	0.0094		0.0019	0.0263	
7	159.14	0.0077	0.0097		0.0202	0.0354		0.0160	0.0101		0.0195	0.0345	
8	146.36	0.0168	0.0092		0.0267	0.0294		0.0118	0.0089		0.0151	0.0311	
9	141.21	0.0197	0.0094		0.0286	0.0302		0.0057	0.0087		0.0224	0.0267	
10	135.53	0.0247	0.0103		0.0152	0.0332		0.0003	0.0088		0.0226	0.0328	
11	135.01	0.0177	0.0089		0.0642	0.0366		0.0176	0.0092		0.0033	0.0243	
12	133.85	0.0215	0.0098		0.0129	0.0287		0.0250	0.0099		0.0320	0.0322	
13	115.17	0.0060	0.0079		0.0343	0.0339		0.0032	0.0078		0.0792	0.0357	
14	107.35	0.0200	0.0084		0.0000	0.0186		0.0070	0.0079		0.0000	0.0216	
15	102.53	0.0021	0.0075		0.0162	0.0270		0.0120	0.0079		0.0000	0.0258	
16	90.35	0.0081	0.0078		0.0460	0.0331		0.0176	0.0085		0.0000	0.0268	
17	81.20	0.0193	0.0080		0.0094	0.0220		0.0111	0.0077		0.0008	0.0183	
18	78.08	0.0044	0.0075		0.0035	0.0287		0.0008	0.0070		0.0110	0.0283	
19	59.13	0.0000	0.0053		0.0296	0.0225		0.0111	0.0065		0.0219	0.0204	
20	63.03	0.0042	0.0068		0.0104	0.0232		0.0056	0.0072		0.0000	0.0208	
21	48.13	0.0057	0.0057		0.0051	0.0146		0.0105	0.0060		0.0045	0.0162	
22	51.30	0.0054	0.0055		0.0011	0.0190		0.0112	0.0060		0.0189	0.0232	
Total	2,881.03	***0.2761***			***0.5524***			***0.2631***			***0.4469***		
Combined		***0.2760***	***0.04***	***$2.0x10^{-13}$***	***0.509***	***0.12***	***$1.58x10^{-6}$***	***0.2578***	***0.04***	***$1.00x10^{-11}$***	***0.4568***	***0.11***	***$3.53x10^{-7}$***

Table S3.

TRAIT (N=1,737)		SBP - Metabochip						DBP - Metabochip					
SNPs		EA cases (N=8,901)			AA cases (N=2,860)			EA cases (N=8,901)			AA cases (N=2,860)		
Chr	L (Mb)	h^2	SE	*P*	h^2	SE	*P*	h^2	SE	*P*	h^2	SE	*P*
1	249.25	0.0139	0.0063		0.0000	0.0158		0.0173	0.0064		0.0000	0.0152	
2	243.20	0.0139	0.0060		0.0370	0.0195		0.0125	0.0058		0.0260	0.0176	
3	198.02	0.0149	0.0058		0.0070	0.0156		0.0215	0.0064		0.0054	0.0136	
4	191.15	0.0182	0.0060		0.0000	0.0157		0.0145	0.0058		0.0026	0.0156	
5	180.92	0.0052	0.0044		0.0149	0.0162		0.0051	0.0046		0.0307	0.0171	
6	171.12	0.0032	0.0039		0.0000	0.0161		0.0023	0.0038		0.0000	0.0144	
7	159.14	0.0078	0.0044		0.0000	0.0143		0.0072	0.0046		0.0000	0.0139	
8	146.36	0.0027	0.0038		0.0000	0.0113		0.0027	0.0039		0.0000	0.0104	
9	141.21	0.0039	0.0045		0.0000	0.0129		0.0028	0.0044		0.0000	0.0126	
10	135.53	0.0097	0.0047		0.0034	0.0112		0.0097	0.0049		0.0000	0.0146	
11	135.01	0.0103	0.0045		0.0000	0.0133		0.0080	0.0046		0.0006	0.0110	
12	133.85	0.0129	0.0052		0.0234	0.0142		0.0113	0.0049		0.0149	0.0128	
13	115.17	0.0095	0.0043		0.0171	0.0133		0.0021	0.0037		0.0082	0.0120	
14	107.35	0.0093	0.0042		0.0054	0.0124		0.0029	0.0037		0.0000	0.0120	
15	102.53	0.0148	0.0050		0.0064	0.0103		0.0102	0.0045		0.0000	0.0094	
16	90.35	0.0026	0.0038		0.0046	0.0123		0.0110	0.0045		0.0059	0.0115	
17	81.20	0.0032	0.0033		0.0023	0.0103		0.0018	0.0029		0.0003	0.0088	
18	78.08	0.0039	0.0037		0.0169	0.0118		0.0016	0.0033		0.0261	0.0123	
19	59.13	0.0004	0.0025		0.0012	0.0084		0.0021	0.0026		0.0000	0.0085	
20	63.03	0.0089	0.0039		0.0005	0.0100		0.0055	0.0035		0.0016	0.0095	
21	48.13	0.0043	0.0029		0.0000	0.0089		0.0074	0.0033		0.0011	0.0086	
22	51.30	0.0035	0.0028		0.0000	0.0078		0.0035	0.0028		0.0000	0.0078	
Total	2,881.03	***0.1772***			***0.1401***			***0.1630***			***0.1201***		
Combined		***0.1772***	***0.04***	***$2.0x10^{-12}$***	***0.1401***	***0.058***	***$1.68x10^{-6}$***	***0.1630***	***0.04***	***$1.40x10^{-9}$***	***0.1201***	***0.059***	***$9.58x10^{-7}$***

Table S4.

TRAIT	SAMPLE SIZE	HERITABILITY	RELATIVES	STUDY	REF
SBP	1,585	0.42	FAMILY	FHS	*31*
SBP	1,617	0.42	TWINS	Chinese Twins	*32*
DBP	1,294	0.39	FAMILY	FHS	*31*
DBP	1,617	0.4	TWINS	Chinese Twins	*32*

Table S5.

Abbreviations:

BP	Blood Pressure
ARIC	Atherosclerosis Risk in Communities
SNP	Single Nucleotide Polymorphisms
EA	European Ancestry
AA	African Ancestry
SBP	Systolic Blood Pressure
DBP	Diastolic Blood Pressure
GWAS	Genome-Wide Association Studies
HTN	Hypertension
CVD	Cardiovascular Disease
MAF	Minor Allele Frequency
BMI	Body-Mass Index
HWE	Hardy-Weinberg Equilibrium
GRM	Genetic Relationship Matrix
CK	Known Cardiovascular/Renal Genes
MLM	Mixed Linear Model
REML	Restricted Maximum Likelihood
LTA	Long-Term Average

CHAPTER 3: CLINICAL AND GENETIC RISK ASSESSMENT OF CARDIOVASCULAR DISEASE

Preface to the Manuscript

This manuscript presents the second part of a study investigating the genetic basis of cardiovascular disease by examining a clinical and genetic risk assessment of the most common cardiovascular disease.

This second study is a methodological paper that explores whether genetic risk score with an independent clinical risk score improves both discrimination and calibration for coronary heart disease. The first manuscript showed that a strong genetic basis is captured by common variants to explain blood pressure variability. Given these findings and the fact that a vast majority of enriched markers tagged most of the blood pressure genetic variance, we felt that it was important to answer this question next because such combination could and illustrate a means to present genetic risk information to subjects and/or their health care provider.

We illustrate this approach in the context of coronary heart disease using a genetic risk score constructed from the most promising association signals reported to date for this disease. We demonstrate how one might interpret a genetic risk score and easily incorporate it into a clinical risk assessment.

We selected genotyped and imputed markers from the most recent and largest multi–stage meta-analysis of GWAS for coronary artery disease conducted by the CARDIoGRAMplusC4D consortium to construct the weighted genetic risk score in White/European subjects from ARIC population. Then we calculated two clinical risk scores to assess clinical risk at 10 years. The first was the well-known "external"

Framingham Risk Score for 10-year risk of coronary heart disease. The second score was developed "internally" within the ARIC and tested and incorporated the same FRS risk factor variables using cross-validation. Subjects with one or more missing FRS risk factors were excluded from the analysis.

In this context, the general objective of this second study provides a means to communicate the effect on risk of genetic data when combined with clinical data.

The specific objective of this second study is:

1. To aggregate a collection of genetic alleles associated with coronary artery disease into a single number, which can then be used for genetic risk assessment.

2. To present a simple method to combine a clinical and genetic assessment.

3. To evaluate the performance of clinical and genetic risk scores under different constructions based on metrics of effect change, discrimination, and calibration.

4. To illustrate one means to provide a risk report about an individual's clinical and genetic risk of disease.

Simple, standardized incorporation of genetic risk into non-genetic risk prediction tools for complex traits: coronary heart disease as an example

Benjamin A. Goldstein[1], Joshua W. Knowles[1], Elias Salfati[1], John P. A. Ioannidis[1-3], and Themistocles L. Assimes[1*]

[1]Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA

[2]Department of Health Research and Policy, Stanford University School of Medicine, Stanford, CA, USA

[3]Department of Statistics, Stanford University School of Humanities and Sciences, Stanford, CA, USA

3.1 Abstract

Purpose: Genetic risk assessment is becoming an important component of clinical decision-making. Genetic Risk Scores (GRSs) allow the composite assessment of genetic risk in complex traits. A technically and clinically pertinent question is how to most easily and effectively combine a GRS with an assessment of clinical risk derived from established non-genetic risk factors as well as to clearly present this information to patient and health care providers.

Materials &Methods: We illustrate a means to combine a GRS with an independent assessment of clinical risk using a log-link function. We apply the method to the prediction of coronary heart disease (CHD) in the Atherosclerosis Risk in Communities (ARIC) cohort. We evaluate different constructions based on metrics of effect change, discrimination, and calibration.

Results: The addition of a GRS to a clinical risk score (CRS) improves both discrimination and calibration for CHD in ARIC. Results are similar regardless of whether external vs. internal coefficients are used for the CRS, risk factor single nucleotide polymorphisms (SNPs) are included in the GRS, or subjects with diabetes at baseline are excluded. We outline how to report the construction and the performance of a GRS using our method and illustrate a means to present genetic risk information to subjects and/or their health care provider.

Conclusion: The proposed method facilitates the standardized incorporation of a GRS in risk assessment.

Keywords: genetic risk scores; personalized medicine; coronary heart disease; electronic health records

3.2 Introduction

As genotyping technologies become more common, the interpretation of genetic risk is becoming a bigger component of clinical decision-making. A particular challenge is the interpretation of such genetic information in the context of other clinical health information. Recently, the electronic MEdical Records and GEnomics (eMERGE) network outlined challenges and opportunities for integrating genetic data into an electronic health records[109] system. One issue identified was the automated interpretation of genetic data[110-113]. The sheer size of genomic data provides many interpretative challenges, particularly in the age of whole genome sequencing with billions of variant base pairs, many of which are *de novo*.

Genetic Risk Scores (GRSs) are one tool for automating the rendition of one's genetic risk. They provide a means to aggregate the health related risk of a collection of genetic alleles into a single number, which can then be used for risk assessment. Using results from genome-wide association studies, one typically combines the observed (or meta-analyzed) log odds-ratio of the risk associated single nucleotide polymorphisms (SNPs). Such scores have been formulated for a variety of complex traits including coronary heart disease (CHD), diabetes, multiple sclerosis and schizophrenia[109, 114, 115]. Overall, GRSs have been shown to modestly improve risk assessment using both traditional and more recently developed model performance metrics[116, 117].

We anticipate individuals will increasingly approach their physicians with questions regarding their genetic risk of common diseases as high density genetic profiling becomes progressively more routinely available. In this paper, we consider the emerging scenario where a hospital system decides to incorporate genetic data into their EHR for the purposes of clinical risk assessment. One obstacle hampering the effective incorporation of GRSs into clinical practice is the lack of clarity in how to most readily combine a GRS with a clinical risk assessment. Here, we describe a relatively straightforward method to combine genetic information at established susceptibility loci with a non-genetic risk prediction tool. We illustrate this approach in the context of CHD using a GRS constructed from the most promising association signals reported to date for

this disease. We emphasize that the goal of this study is neither to validate the utility of a GRS in risk prediction nor to assess the best way to construct a GRS but rather to demonstrate how one might interpret a GRS and easily incorporate it into a clinical risk assessment. A GRS can be constructed in a variety of ways[118]. One may select SNPs and define their respective high-risk allele either through the investigation of SNP effects within the cohort itself or within external studies that are typically much larger but not necessarily prospective in nature. One may also weigh the high-risk allele by its effect size observed internally or externally. In this study, we used the weighted approach deriving both the SNPs and weights from external sources. Lastly, we illustrate one way to present risk prediction analyses incorporating GRSs to patients and health care providers.

3.3 Methods

3.3.1 SNP Selection & Weighting

We selected SNPs from the most recent and largest multi–stage meta-analysis of GWAS for coronary artery disease conducted by the CARDIoGRAMplusC4D consortium to construct the GRS[119]. The study included 63,746 cases and 130,681 controls. The vast majority of the subjects included in this meta-analysis reported white/European ancestry. The meta-analysis added 15 new CHD susceptibility loci and confirmed nearly all loci that had previously reached genome-wide significance. The investigators also identified secondary signals at four established loci. Supplementary table 9 of the CARDIoGRAMplusC4D manuscript lists all uncorrelated SNPs (r2<0.2) with an estimated FDR <5%[119]. From this list, we selected the 50 SNPs identified by the consortium as validated SNPs because they had reached a genome-wide level of statistical significance in either the CARDIOGRAMplusC4D meta-analysis or in any previous GWAS.

We expect a subset of SNPs to be influencing the risk of CHD through traditional risk factors as the CARDIOGRAMplusC4D meta-analysis adjusted only for age and sex. Indeed, the CARDIoGRAMplusC4D investigators determined that 12 and 5 of these 50 SNPs likely influence CHD risk through effects on lipids and blood pressure based on

their strong association with these traits in the Global Lipids Genetics Consortium and the International Consortium of Blood Pressure meta-analyses of GWAS, respectively[119]. For the purposes of this study, we classified these 17 SNPs as "risk factor SNPs". The remaining 33 SNPs were classified as "non-risk factor SNPs".

3.3.2 Prospective Cohort for testing Genetic Risk Scores

We selected the AtherosclerosisRisk in Communities Study (ARIC) study to develop and test a GRS constructed with the 50 SNPs of interest. The ARIC Study is an ongoing prospective investigation of atherosclerosis and its clinical sequelae involving 15,792 white and black persons aged 45–64 years at recruitment (1987–1989[120]). Detailed descriptions of the study designs, IRB consent process, sampling procedures, methods, definitions of cardiovascular outcomes, and approach to statistical analyses is published elsewhere[121, 122].

We selected ARIC for several reasons including the availability of individual level genome-wide data for all participants through the National Institutes of Health (National Human Genome Research Institute) controlled access database of Genotypes and Phenotypes (dbGaP), a prolonged follow up with > 1000 incident cases, and no overlap of incident cases with prevalent cases that were included in the CARDIoGRAMplusC4D consortium study[119]. The Affymetrix 6.0 array was used to genotype all participants of the ARIC study.

All white/Europeans without a history of CHD, myocardial infarction, or heart failure at baseline among the ARIC cohort subjects in dbGAP were eligible for study inclusion. Incident CHD was defined by the recording for the first time of either non-fatal or fatal myocardial infarction (“mi04”, “fatchd04”), CHD related revascularization procedure (“in_by04p”), or silent MI detected by ECG (“in_04s”).

The outcome of interest was incident CHD within 10 years. Those without a positive event who died or were lost to follow up prior to their 10th year anniversary of follow up were removed from analysis. All others were deemed event free at 10-years

regardless of whether they developed incident CHD sometime after their 10 year anniversary of follow up.

3.3.3 Clinical Risk Score Assessment

We calculated two clinical risk scores (CRSs) to assess clinical risk at 10 years. The first was the well-known "external" Framingham Risk Score (FRS) for 10-year risk of CHD. The score is based on one's gender, age, total cholesterol, HDL cholesterol, blood pressure, and diabetes and smoking status. Ten-year risk of CHD was calculated using the published regression coefficients[123]. The second score was developed "internally" within the ARIC and tested and incorporated the same FRS risk factor variables using cross-validation (see below). Subjects with one or more missing FRS risk factors were excluded from the analysis.

3.3.4 Imputation of ARIC raw genotype data to 1000 genomes

We imputed individual level genotype data from ARIC to the latest build of the 1000 genomes project (1kGP) used a hidden Markov model to minimize the need to use proxy SNPs in the construction of the GRS[124, 125]. We first phased each chromosome using MaCH (v1.0.16) by running 20 rounds of the Markov sampler and considering 200 haplotypes (states) when updating each individual. We then used phased haplotypes in each chromosome and the latest release of the 1kGPcosmopolitan panel (version 3 March 2012 release, 246 AFR + 181 AMR + 286 ASN + 379 EUR) to impute all SNPs in the cosmopolitan panel using the OpenMP protocol based multi-threaded version of Minimac (v4.6) with 20 rounds and 300 states for each chromosome. Genotyped SNPs used for imputation were restricted to those with the following features: MAF >0.1%, missing data per SNP < 2%, and Hardy-Weinberg equilibrium (HWE) $p > 10^{-6}$. Of the 841,820 autosomal genotyped markers, 543,653 passed the initial quality filters and were used for the imputation of over 37 million SNPs in ARIC. We used GTOOL (Genetics Software Suite, (c) 2007, The University of Oxford) to convert Minimac dosage files to best guess genotype calls.

3.3.5 GRS Construction

We calculated the GRS for an individual in the typical approach as a weighted sum of the number of high-risk alleles [1]:

$$GRS = \sum_{i \in GRS}^{50} w_i \sum_{j=1}^{2} RA_{ij} \qquad (1)$$

where the inside summation, RA_{ij}, is the count of high risk alleles and the weight, w_i, is the meta-analyzed log odds-ratio for SNP i. We used the corresponding "combined beta" (i.e. the beta across the stage 1 and 2 CARDIOGRAMplusC4D meta-analysis) to weigh the SNP when constructing the GRS. We carefully identified the high-risk allele for each SNP. We used the GTOOL genotype calls to count high-risk alleles for all SNPs in each individual after first dropping SNPs with a low imputation quality ($r^2 < 0.3$).

There are two primary assumptions in such a construction. Since this summation is over marginal effects, each effect is assumed to be independent. The second is that the effects are linearly additive, i.e. there are no interactions. For the first assumption, care was taken to select SNPs that are not in linkage disequilibrium (i.e. correlated) with one another in white/European descent participants ($r^2 < 0.2$). While the second assumption is likely violated, it is also reasonable to assume that marginal effects capture a majority of genetic risk for CHD[126, 127]. When using the GRS we standardize it to have a mean of 0 and standard deviation of 1.

3.3.6 Combining Clinical and Genetic Risk

We present a simple and easy way to combine one's CRS and GRS by using the following model [2]:

$$\log(P(CHD|\text{Clinical \& Genetic Factors})) = \alpha + \beta_1 CRS + \beta_2 GRS \qquad (2)$$

This is a standard generalized linear model, where the outcome is a binary (0 – 1) indicator for incident CHD within 10 years and the predictor variables are the CRS and GRS, respectively. The CRS represents either a calculated risk due to non-genetic clinical factors (as in FRS) or a summation over multiple clinical risk factors (when using internal coefficients). We emphasize the use of a log link function instead of the more frequently used logistic link function (as in logistic regression). This allows the two coefficients of interest (β_1 and β_2) to represent log relative risks (RR), making the following transformation more straightforward. However, we note that using the logistic link one could perform a similar transformation. After exponentiating equation [2] we obtain:

$$\begin{aligned} P(CHD|\text{Clinical \& Genetic}) &= e^{\alpha+\beta_1 CRS} \times e^{\beta_2 GRS} \\ &= P(CHD|\text{Clinical}) \times RR^{GRS}_{(GRS)} \end{aligned} \quad (3)$$

In the second line, we have combined the intercept (α) with the effect due to clinical factors. This is generally well captured by a CRS (like FRS) that incorporates the prevalence of disease in the general population. Since we are multiplying the estimated effects for the GRS and CRS, the primary assumption is that the GRS is linearly independent of the CRS. This assumption would potentially be violated if the GRS consisted of SNPs that were thought to act entirely or largely through effects on non-genetic clinical risk factors measured at baseline. However the impact is mitigated by controlling for the CRS while estimating the RR for the GRS in equation [2]. Therefore, to calculate a probability of CHD based on clinical and genetic factors, we must:

(1) Estimate the RR for a one-unit change in GRS on the probability of CHD within 10 years controlled for CRS.

(2) For a given individual:

   a) Calculate the probability of CHD based on clinical factors via a FRS or Internal Score

   b) Calculate the GRS (based on equation 1) and standardize it using population mean and standard deviation (SD)

c) Multiply the probability from a) by the RR from [1] raised to the value of standardized GRS from b) (based on second line of equation 3)

3.3.7 Evaluation of performance of risk scores

We used 10-fold cross-validation to test both the CRS and GRS, dividing the cohort into a series of independent training and test sets. We created a series of updated risk scores:

(1) CRS based solely on the FRS (no genetic information considered)
(2) CRS based solely on the internal coefficients (no genetic information considered)
(3) CRS updated with a GRS constructed using all SNPs of interest that were either well genotyped or well imputed in ARIC.
(4) CRS updated with a GRS constructed using only "non–risk factor" SNPs among the SNPs in (3)
(5) CRS updated with a GRS constructed using only "risk factor" SNPs among the SNPs in (3)

The overall relative risk for a standardized one-unit change in GRS was estimated while incorporating the CRS (either FRS or internal). Within each of the 10 folds, the training (9/10) and test (1/10), we created a standardized score based on the mean and standard deviation from the training set. The models were estimated on the training split and applied to the test split. We used three forms of assessment. First, we calculated the c-statistic to assess discrimination of the various risk scores. Discrimination refers to a model's ability to separate subjects into distinct groups, in this case, those with CHD from those without. Secondly, we calculated the RR for a one standard deviation change in GRS. Finally, we calculated the calibration slope to assess each models overall calibration[128].The calibration of a model is the extent to which the predicted probability reflects the true underlying probability. The calibration slope is a more interpretable statistic than the more typical Hosmer-Lemeshow statistic, representing the degree of miscalibration[129]. A calibration slope of 1.0 indicates perfect calibration while values less than 1.0 suggest over-fitting and above 1.0 poorer calibration. For example a calibration slope of 2.0 indicates a 2-fold increase in miscalibration. We chose not to assess our models using the Net Reclassification Index (NRI) or the clinical NRI due to recent

concerns about the utility and validity of this metric combined with changing clinical guidelines for cardiovascular disease risk assessment[130-134].

In a sensitivity analysis, we repeated the above comparisons but restricted the cohort to those without prevalent diabetes. We also considered a risk prediction model using only a GRS adjusted for age and gender and no other clinical risk factors to provide a perspective on the overall impact of clinical risk factors compared to the genetic risk score. Finally, we assessed the potential for population stratification by performing a principal components analysis (PCA) with 741 ancestry informative markers (AIMs) using EIGENTRAT[135] followed by a regression of CHD status onto all significant components, adjusted for the clinical factors.

All analyses were performed in R 3.0.1[136].

3.3.8 Risk Reports

Using the generated information, we illustrate one means to provide a risk report about an individual's clinical and genetic risk of disease. Three key pieces of information are included:

(1) The number of risk alleles
(2) How the individual's GRS compares to the distribution of GRSs in a comparative population.
(3) The change in one's overall risk after accounting for genetic risk

The number of risk alleles represents a simple count of the number of alleles that have been associated with an increased risk of CHD. The GRS comparison to the general population is based on the individual's standardized GRS. Finally the updated risk is calculated from equation (3). A fourth piece of information that can be included in the risk report is a statement of how the individual's change in overall risk after accounting for genetic risk influences clinical management. This may be based on some well-accepted guidelines whose recommendations can be easily and reliably automated.

3.4 Results

3.4.1 ARIC cohort exclusions

Of the 12,771 from the ARIC cohort with phenotypic and genotypic data, 9,633 (75%) were white/European (see Figure 1). Among the remaining subjects, 721 (7.5%) had a history of CHD or CHF at baseline and were excluded from further analysis. Lastly, we excluded 380 people who were lost to follow-up or died of non-CHD related factors within 10 years and 41 people with missing covariate information, comprising a final cohort of 8,491.Table 1 shows the baseline characteristics for the ARIC subcohort used in our analyses. The predicted 10-year risk of developing CHD based on the FRS in this subcohort is 7.4% (interquartile range 4.3% to 12.3%). This predicted risk coincided very well with the observed proportion that developed CHD (7.3%).

3.4.2 Risk Scores

The 50 SNPs of interest for construction of the GRS are listed in supplemental Table 1 along with their relationship to risk factors, weights, high risk allele based on the 1000G reference + strand, imputation quality metrics, and genotype quality control metrics. Of the 50 SNPs, five had an estimated imputation accuracy $r^2 < 0.3$. These five SNPs, which included two SNPs in the *APOE* locus, were dropped from the GRS. The average r^2 of the remaining 45 SNPs was 0.857 (range: 0.361 to 0.999). The unstandardized mean value of the GRS was 3.17 (SD: 0.347) for all SNPs, 1.95 (0.307) for non-risk factor SNPs alone, and 1.22 (0.160) for risk factor SNPs alone. Interestingly, there was no difference in the unstandardized scores and standard deviations derived from the entire cohort compared to the scores derived from the subset of subjects without diabetes at baseline when considering up to three significant figures. After standardization, the mean and SD of all GRS was 0 and 1 as expected.

3.4.3 Performance of risk scores and sensitivity analyses

Table 2 summarizes the c-statistics for the 8 risk scores (as well as the age and sex only scores) and the associated RR for a 1-unit change in the risk score. Adding a GRS improves overall risk discrimination. As expected, the risk score using internal weights demonstrates the best discrimination and calibration. The calibration slope statistics

improved (i.e. they become smaller) with the addition of the GRS.A GRS restricted to SNPs that were not related to traditional risk factors performed essentially equally well to a GRS constructed from all SNPs combined, adding about 1 point to the c-statistic. This result suggests that the addition of CHD SNPs that are associated with CHD as well as risk factors will neither aid nor hurt risk assessment. Finally, creating a risk score only with age and sex performed worse than the risk scores with additional clinical factors. However, the improvement in both discrimination and calibration after adding the GRS is comparable to the scores with the full clinical factors.

Table 3 summarizes the same risk score comparisons presented in Table 2 after removing 626 ARIC participants (7.4%) who reported having diabetes at baseline. We found the general trend of results to be similar to the full cohort despite a smaller sample size. There was a modest improvement in discrimination by about 1 point in the c-statistic as well as improvement in calibration.

PCA revealed eight significant principal components. Only component 3 had a nominal associationwith CHD ($p = 0.023$, not corrected for number of components tested) suggesting that the addition of PCs into our model for this sample of self reported white/Europeanswould not materially influence our results (Supplemental Table 2).

3.4.4 Risk Reports

In figure 2, we illustrate a sample report for an individual to show how the addition of a GRS to the model can change the risk assessment that may be used for clinical decision-making. The goal of this report would be to facilitate a conversation around the risk of CHD due to genetics above beyond the known clinical risk factors. At baseline, the participant's estimated risk of CHD at 10 years is 5.5% based on traditional Framingham risk factors. The participant carries 49 of 90 potential risk alleles resulting in a weighted standardized GRS of 1.26 which places the individual in the 89th percentile of genetic risk (i.e. only 11% of the population has a higher risk based on alleles inherited at these 45 SNPs). Combining the participant's genetic risk with their clinical risk results in a final predicted risk of CHD of 7.6% given each SD increase in one's GRS leads to a 38% increase in risk of CHD (Table 2). This magnitude of increased risk may affect the

decision to treat this patient with statins[137]. Ultimately, this person did develop CHD suggesting that the upward adjustment of risk was appropriate.

3.5 Discussion

Genetic risk assessment will become an increasingly important component of overall clinical risk assessment. In this context, we ask the question: how can one most easily and effectively incorporate a GRS into an existing clinical risk assessment of a complex trait without compromising effectiveness? We present a straightforward means to combine genetic risk with clinical risk for a given disease where large-scale cohorts with prolonged follow up exist and can be used to evaluate novel biomarkers. Our approach requires knowing only three pieces of information: 1) an individual's GRS, 2) an individual's CRS, and 3) the RR associated with a 1-unit change in standardized GRS within the cohort. Recent studies demonstrate an increasing clinical utility of GRSs for CHD[115, 138-142]. Using our method, we were able to confirm this trend and demonstrate comparable or slightly improved discrimination even when comparing our results to the subset of studies that used a GRS constructed with a similar set of SNPs[115, 138, 139, 141-143]. We should stress that evidence in the form of a well-executed clinical trial that clearly demonstrates the value of a GRS in improving CHD outcomes does not yet exist[144]. Thus, we are not endorsing or negating the use of any specific GRS in the primary prevention of CHD on the basis of our results. Ongoing trials are examining the ability of information from GRS to improve outcomes[145, 146].

Our approach makes the simplifying assumption that the GRS is largely independent of the CRS. This assumption appears reasonable when one reliably restricts SNPs included in the GRS to those influencing risk independent of variables included in the CRS. We tested this assumption by creating two subset GRSs, one restricted to SNPs associated with risk factors and one restricted to SNPs that appear to influence risk of CHD independent of all established risk factors. The non-risk factor GRS performed noticeably better than the risk factor GRS confirming the consequence of grossly violating this assumption. However, we detected no notable difference between the non–risk factor GRS compared to the full GRS. Thus, our approach appears robust to small

violations of this assumption. This confirms others' and our experiences with GRSs that they are fairly robust to alternative constructions[114, 147].

An important consideration is the construction of the CRS. We suspect that the ability to derive and make use of such internal coefficients will be facilitated by the increasing availability of EHR with prolonged follow up of individuals receiving care as members of a large-scale health maintenance organization[148-152]. As expected, the use of internal coefficients led to a slightly more effective CRS compared to the FRS that was developed in a different cohort than ARIC. Despite this observation, we observed a negligible difference in the RR suggesting that perhaps under some circumstances one can develop a GRS using an internal CRS and apply it successfully in other cohorts (or vice-versa). We also note that while the GRS improves calibration, the risk scores overall are still poorly calibrated (> 1), particularly the one using the FRS. This reflects other work that has shown that the external coefficients applied to new populations can often lead to poorly calibrated models[131]. Finally, the risk score using only age and sex, not surprisingly, performed the worst. Moreover, the improvement in both discrimination (68.9 vs. 77.3) and calibration (11.22 vs. 4.34) after adding additional clinical factors is much greater than after the addition of a GRS highlighting the relative importance of clinical factors collectively at this point in time over the GRS in risk assessment for CHD. However, one should not automatically assume that the current GRS is not clinically useful given its ΔAUC as it is in the same range as that seen for the addition of any single modifiable traditional risk factor to a model that includes all other traditional risk factors.

Several steps need to be followed in reporting of a GRS for a trait using our method to facilitate its testing in additional populations or to easily disseminate its use. First, the cohort in whom the GRS was derived including the age range, sex distribution, risk factor profile, and the ethnicity of its members must be clearly described. The GRS we present here is most relevant to white/Europeans in the age range of 45 to 64 and free of CHD at the time of clinical risk assessment given the eligibility criteria of the ARIC study and the fact that the SNPs used in the GRS were derived from large-scale case-

control studies that included subjects in the same race/ethnic group and age range[119, 120]. A different sets of SNPs with different weights will likely be necessary for different race/ethnic groups and possibly different age ranges although we expect substantial overlap across race/ethnic groups in the genomic regions contributing at least one SNP to the GRS[145, 153]. Second, one must reliably identify and report which allele was coded as the high-risk allele as this allele is not necessarily the minor allele. Errors in this context due to inadvertent strand flipping either in the original study reporting the susceptibility variant or in the construction of the GRS may have a profound negative impact on the performance of the GRS. Third, the effect estimate for each SNP (generally a log odds ratio) used in the weighting of the GRS should be clearly presented. Lastly, the relative risk for a one-unit change in GRS should be calculated and clearly presented along with the mean and SD of the GRS to facilitate standardization of the score.

We suggest a means to communicate the effect on risk of someone's genetic data when combined with his or her clinical data. Our presentation includes both a contextualization relative to the general population and a statement on how one's inherited variants update one's clinical risk that is based strictly on traditional non-genetic risk factor data. In ongoing clinical investigation, we have applied a similar reporting system within a cardiology clinic[145]. Such a report can easily be automated and incorporated into an EHR. Moreover, it can also easily be updated as new susceptibility SNPs are discovered and/or weights refined. Given genome wide genotyping or sequencing is likely to become routine in the near future, more research is needed to identify the optimal way to communicate this information to subjects at risk and health care providers.

Risk scores are likely to evolve over time and practice guidelines may adopt different risk scores. For example, the FRS that we used here forms the basis of the Adult Treatment Panel III (ATPIII) guidelines[154]. Recently, ACC/AHA released new cardiovascular prevention guidelines, with new categories of risk, with a change in the relevant endpoints and in the risk calculation formulas[132, 137]. As of this writing, there is still large controversy about the accuracy of the new calculations and the validity of the

guidelines[130, 131, 134, 155]. Regardless, our proposed methods can be used to incorporate GRS in any sets of non-genetic predictive models.

In conclusion, we present a simple but effective means to combine a CRS with a GRS and illustrate one way to present such information to an individual interested in understanding how this genetic information influences their risk assessment and thus potentially their clinical management. Furthermore, we highlight information that should be included in all reports of GRSs to facilitate the timely assessment of a new GRS by other investigators in additional populations or, alternatively, to easily incorporate it into clinical practice if its efficacy is no longer in question. We expect the importance of such research to grow over time and hope that future studies will more clearly delineate the optimal way to implement a GRS and how to most effectively disseminate a well-established GRS to patients and their health care providers.

Funding Sources,

BAG is supported by an NIH career development award K25DK097279. JWK is supported by an American Heart Association, National Fellow to Faculty Award, 10FTF3360005. TLA is supported by an NIH career development award K23DK088942.

Disclosures

None

Figure Legends

Figure 1. Atherosclerosis Risk in Communities (ARIC) cohort inclusion and exclusion criteria applied to data obtained from the NCBI's database of genotypes and phenotypes (dbGAP).

Figure 2. A sample report on CHD risk for an individual in the ARIC study where the incorporation of genetic risk into the model of clinical risk potentially influences clinical management based on current guidelines.

Figure 1.

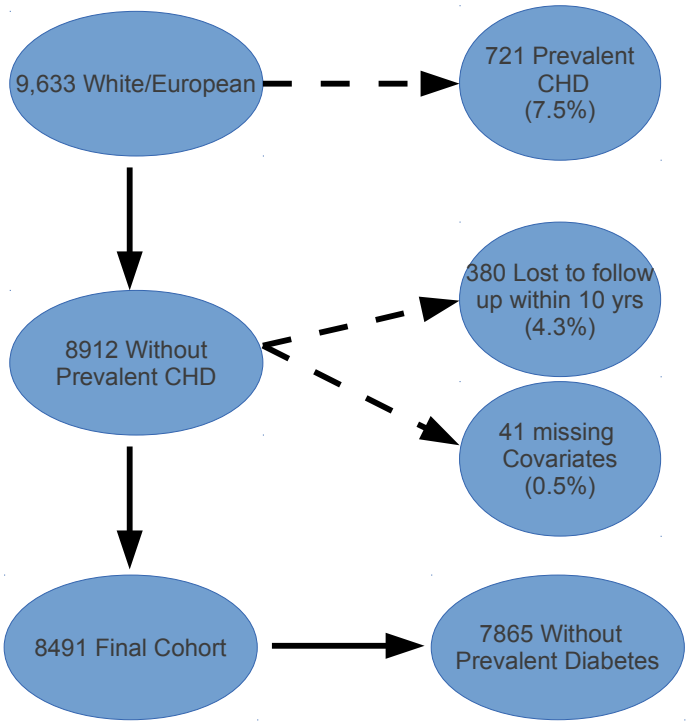
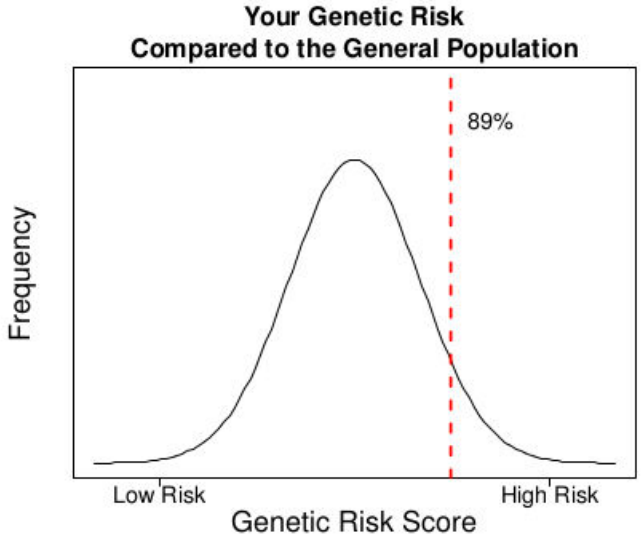

9,633 White/European
721 Prevalent CHD (7.5%)
8912 Without Prevalent CHD
380 Lost to follow up within 10 yrs (4.3%)
41 missing Covariates (0.5%)
8491 Final Cohort
7865 Without Prevalent Diabetes

Figure 2.

Your Risk Score

Based on the traditional Framingham risk score, your risk of coronary heart disease over the next 10 years is approximately 5.5%.

We tested for a total of 90 possible risk variants or alleles. Out of these 90, you carry 49 variants that are associated with higher risk. Your genetic profile puts you in the 89 percentile for risk. This means 89% of the general population have a genetic risk score more favorable than you and 11% have a genetic risk score less favorable than you.



Based on the traditional Framingham risk score plus the genetic risk score, your risk of coronary heart disease over the next 10 years is approximately 7.6%.

Your 10 year risk of coronary heart disease risk is ≥7.5% when considering your genetic risk. This information may be discussed with your physician in terms of what would be recommended as most appropriate management given your estimated risk.

Table 1: Characteristics of the ARIC subcohort used in analyses (n=8491)

	mean (IQR)
Age (years)	54 (49,59)
SBP (mm/Hg)	116 (106, 128)
DBP (mm/Hg)	71 (65, 78)
HDL (mg/dL)	48 (39, 61)
TC (mg/dL)	211 (187, 238)
	count (%)
white/European	8491 (100)
Male	3848 (45)
Diabetes	626 (7.4)
Smoking status	
Current	2010 (24)
Former	2914 (34)
Never	3567 (42)

IQR = inter-quartile range, SBP = Systolic Blood Pressure, DBP = Diastolic Blood Pressure, HDL = High-Density Lipoprotein Cholesterol, TC = Total Cholesterol

Table 2. Relative Risks and discrimination metrics for a genetic risk score derived from 50 genome wide significant susceptibility alleles for CHD in the full ARIC sample (n=8491) of white/Europeans subjects

	Relative Risk (95% CI)	C-statistic*	Calibration Slope
Using FRS for clinical risk score			
FRS alone	--	75.8	7.32
+ full GRS	1.29 (1.20, 1.40)	76.8	6.26
+ GRS restricted to non risk factor SNPs	1.29 (1.20, 1.40)	76.8	6.29
+ GRS restricted to risk factor SNPs	1.06 (0.98, 1.14)	75.8	7.22
Using internal coefficients for clinical risk score			
Internal coefficients alone	--	77.3	4.34
+ full GRS	1.28 (1.19,1.38)	78.3	4.17
+ GRS restricted to non risk factor SNPs	1.29 (1.20, 1.39)	78.3	4.18
+ GRS restricted to risk factor SNPs	1.05 (0.97, 1.13)	77.4	4.31
Using only age and sex			
Internal coefficients alone	--	68.9	11.22
+ full GRS	1.31 (1.22,1.41)	70.4	9.26
+ GRS restricted to non risk factor SNPs	1.29 (1.20,1.39)	70.1	9.69
+ GRS restricted to risk factor SNPs	1.11 (1.03, 1.20)	69.2	10.79

CHD = Coronary Heart Disease, ARIC = Atherosclerosis Risk in Communities, FRS = Framingham Risk score, SNPs = Single Nucleotide Polymorphism, GRS = genetic risk score, *performance of second model listed to first model listed

Table 3. Relative Risks and discrimination metrics for a genetic risk score derived from 50 genome wide significant susceptibility alleles for CHD in the ARIC subset of white/Europeans with no diabetes at baseline (n=7865)

	Relative Risk (95% CI)	C-statistic*	Calibration Slope
Using FRS for clinical risk score			
FRS alone	--	75.2	8.84
+ full GRS	1.28 (1.17, 1.39)	76.2	7.02
+ GRS restricted to non risk factor SNPs	1.30 (1.20, 1.41)	76.3	7.22
+ GRS restricted to risk factor SNPs	1.02 (0.94, 1.11)	75.1	8.67
Using internal coefficients for clinical risk score			
Internal coefficients alone	--	76.7	6.11
+ full GRS	1.28 (1.18, 1.39)	77.6	5.39
+ GRS restricted to non risk factor SNPs	1.30 (1.20, 1.42)	77.7	5.40
+ GRS restricted to risk factor SNPs	1.03 (0.95, 1.12)	76.6	6.00
Using only age and gender			
Internal coefficients alone	--	70.5	12.86
+ full GRS	1.30 (1.20,1.41)	71.8	10.49
+ GRS restricted to non risk factor SNPs	1.28 (1.18,1.39)	71.6	10.92
+ GRS restricted to risk factor SNPs	1.10 (1.01, 1.19)	70.7	12.44

CHD = Coronary Heart Disease, ARIC = Atherosclerosis Risk in Communities, FRS = Framingham Risk score, SNPs = Single Nucleotide Polymorphism, GRS = genetic risk score, *performance of second model listed to first model listed

Table S1. Single nucleotide polymorphisms considered for construction of Genetic Risk Score, summary of location, high risk allele, betas used as weights, relationship to Framingham Risk factor(s), imputation metrics, and genotype metrics

						1000G + strand reference		Information for Genotyped SNPs in ARIC															
Chr	Position	SNP	Nearest Transcript	Meta-Analysis Beta*	Associated Risk Factor(s)	HIGH RISK allele	Other allele	in ARIC?**	A1	A2	MAF	NCHROBS	N_MISS	N_GENO	F_MISS	Freq1	MAF	AvgCall	Rsq	LooRsq	EmpR	EmpRsq	High risk count in GRS?
1	55,496,039	rs11206510	PCSK9	0.055	TC,LDL	T	C	1	G	A	0.192	19256	5	9633	0.0005	0.829	0.171	0.860	**0.291**	0.291	0.000	0.0000	No
1	56,962,821	rs17114036	PPAP2B	0.106		A	G	0								0.911	0.089	0.994	0.946	-	-	-	Yes
1	109,821,511	rs602633	PSRC1/SORT1	0.116	LDL,TC,HDL	G	T	0								0.777	0.223	0.994	0.972	-	-	-	Yes
1	154,422,067	rs4845625	IL6R	0.049		T	C	0								0.580	0.420	0.995	0.979	-	-	-	Yes
2	19,942,473	rs16986953	AK097927	0.104		A	G	0								0.931	0.069	0.994	0.920	-	-	-	Yes
2	21,286,057	rs515135	APOB	0.075	LDL,TC	C	T	0								0.828	0.172	0.976	0.869	-	-	-	Yes
2	44,073,881	rs6544713	ABCG5/ABCG8	0.061	LDL,TC	T	C	0								0.738	0.262	0.850	0.478	-	-	-	Yes
2	85,809,989	rs1561198	GGCX/VAMP10/VAMP8	0.052		T	C	1	T	C	0.4579	19258	4	9633	0.0004	0.542	0.458	0.999	0.998	0.897	0.954	0.9106	Yes
2	145,801,461	rs2252641	ZEB2-AC074093.1	0.048		C	T	1	G	A	0.4545	19262	2	9633	0.0002	0.453	0.453	0.996	0.987	0.987	0.000	0.0000	Yes
2	203,880,992	rs2351524	WDR12/ALS2CR16	0.115		T	C	1	A	G	0.1319	19236	15	9633	0.0016	0.873	0.127	0.985	0.881	0.881	0.000	0.0000	Yes
3	138,119,952	rs2306374	MRAS	0.073		C	T	1	G	A	0.1567	19248	9	9633	0.0009	0.845	0.155	0.999	0.996	0.996	0.000	0.0000	Yes
4	148,288,067	rs1429141	EDNRA	0.066		T	C	0								0.835	0.165	0.999	0.996	-	-	-	Yes
4	156,635,309	rs7692387	GUCY1A3	0.065	DBP	G	A	0								0.814	0.186	0.994	0.962	-	-	-	Yes
5	131,667,353	rs273909	SLC22A4/SLC22A5	0.077		G	A	0								0.885	0.115	0.981	0.861	-	-	-	Yes
6	12,962,995	rs13211739	PHACTR1	0.058		G	A	0								0.774	0.226	0.982	0.915	-	-	-	Yes
6	34,898,455	rs12205331	ANKS1A	0.042		C	T	0								0.796	0.204	0.994	0.966	-	-	-	Yes
6	39,174,922	rs10947789	KCNK5	0.060		T	C	0								0.766	0.234	0.989	0.944	-	-	-	Yes
6	134,214,525	rs12190287	TCF21	0.072		C	G	2	C	G	0.3766	19248	9	9633	0.0009	0.641	0.359	0.908	0.691	-	-	-	Yes
6	160,863,532	rs2048327	SLC22A3/LPAL2/LPA	0.060	TC,LDL	C	T	1	C	T	0.3683	19264	1	9633	0.0001	0.632	0.368	1.000	1.000	0.987	0.996	0.9915	Yes
6	161,143,608	rs4252120	PLG	0.062		T	C	1	G	A	0.3019	19090	88	9633	0.0091	0.700	0.300	0.998	0.990	0.990	0.000	0.0000	Yes
7	19,036,775	rs2023938	HDAC9	0.073		C	T	0								0.906	0.094	0.994	0.939	-	-	-	Yes
7	129,663,496	rs11556924	ZC3HC1	0.083	DBP	C	T	1	T	C	0.3826	19166	50	9633	0.0052	0.618	0.382	0.999	0.996	0.311	0.902	0.8144	Yes
8	19,813,180	rs264	LPL	0.071	HDL,TG	G	A	0								0.866	0.134	0.979	0.863	-	-	-	Yes
8	126,490,972	rs2954029	TRIB1	0.048	TC,LDL ,HDL,	A	T	0								0.532	0.468	0.986	0.948	-	-	-	Yes
9	22,003,223	rs3217992	CDKN2BAS/MTAP/CDKN2B	0.145		T	C	2	A	G	0.3755	18850	208	9633	0.0216	0.641	0.359	0.973	0.897	-	-	-	Yes
9	22,125,503	rs1333049	CDKN2BAS	0.207		C	G	0	C	G	0.4781	19256	5	9633	0.0005	0.528	0.472	0.921	0.720	-	-	-	Yes
9	136,154,867	rs495828	ABO	0.066	LDL,TC	T	G	0								0.772	0.228	0.961	0.825	-	-	-	Yes
10	30,335,122	rs2505083	KIAA1462	0.061		C	T	1	C	T	0.4218	19116	75	9633	0.0078	0.581	0.420	0.998	0.992	0.640	0.880	0.7748	Yes
10	44,539,913	rs2047009	CXCL12/AX747950	0.053		G	T	1	A	C	0.4841	19260	3	9633	0.0003	0.485	0.485	0.999	0.997	0.997	0.000	0.0000	Yes
10	44,753,867	rs501120	CXCL12	0.067		T	C	0								0.868	0.132	1.000	0.999	-	-	-	Yes
10	90,989,109	rs11203042	LIPA	0.039		T	C	0								0.557	0.443	0.979	0.928	-	-	-	Yes
10	91,005,854	rs2246833	LIPA	0.055		T	C	1	T	C	0.3412	19146	60	9633	0.0062	0.659	0.341	0.999	0.997	0.959	0.982	0.9636	Yes
10	104,652,323	rs11191447	CYP17A1/CNNM2/NT5C2	0.087	SBP,DBP	C	T	0								0.912	0.088	0.999	0.991	-	-	-	Yes
11	103,660,567	rs974819	PDGFD	0.065		T	C	0								0.711	0.289	0.989	0.950	-	-	-	Yes
11	116,611,733	rs9326246	ZNF259/APO5A/APOA1	0.086	TC,HDL,LDL,	C	G	0								0.929	0.071	0.994	0.928	-	-	-	Yes
12	111,884,608	rs3184504	SH2B3	0.068	DBP,TC,SBP,LDL,HDL	T	C	0								0.518	0.482	0.762	0.361	-	-	-	Yes
13	28,973,621	rs9319428	FLT1	0.055		A	G	0								0.695	0.305	0.997	0.990	-	-	-	Yes
13	110,960,712	rs4773144	COL4A1/COL4A2	0.068		G	A	1	G	A	0.4432	18932	167	9633	0.0173	0.562	0.438	0.999	0.996	0.810	0.895	0.8008	Yes
13	111,049,623	rs9515203	COL4A1/COL4A2	0.079		T	C	1	C	T	0.2621	19244	11	9633	0.0011	0.737	0.263	0.998	0.991	0.797	0.926	0.8568	Yes
14	100,133,942	rs2895811	HHIPL1	0.056		C	T	0								0.617	0.383	0.824	0.478	-	-	-	Yes
15	79,006,582	rs11072794	ADAMTS7/DQ582071	0.066		T	C	1	A	G	0.254	19244	11	9633	0.0011	0.660	0.340	0.770	**0.231**	0.231	0.000	0.0000	No
15	79,141,784	rs7173743	ADAMTS7/MRG15	0.065		T	C	0								0.543	0.457	0.734	**0.266**	-	-	-	No
15	91,416,550	rs17514846	FES/FURIN	0.058	SBP	A	C	0								0.517	0.483	0.853	0.609	-	-	-	Yes
17	2,117,945	rs2281727	SMG6	0.050		G	A	0								0.651	0.349	0.952	0.812	-	-	-	Yes
17	17,543,722	rs12936587	RAI1/PEMT/RASD1	0.055		G	A	0								0.552	0.448	0.971	0.912	-	-	-	Yes
17	47,005,193	rs15563	UBE2Z	0.037		G	A	1	A	G	0.4634	19252	7	9633	0.0007	0.463	0.463	1.000	1.000	0.993	0.999	0.9982	Yes
19	11,163,601	rs1122608	LDLR/SMARCA4	0.092	LDL,TC	G	T	1	T	G	0.2484	19200	33	9633	0.0034	0.753	0.247	0.991	0.955	0.471	0.903	0.8161	Yes
19	45,395,619	rs2075650	APOE/APOC1/TOMM40	0.106	LDL,TC,TG,HDL,	G	A	0								0.868	0.132	0.868	**0.032**	-	-	-	No
19	45,415,640	rs445925	APOE/APOC1/TOMM40	0.119	LDL,TC,TG,HDL	G	A	0								0.869	0.131	0.870	**0.030**	-	-	-	No
21	35,599,128	rs9982601	KCNE2/C21orf82	0.119		T	C	0								0.866	0.134	0.988	0.926	-	-	-	Yes

*for the high risk allele compared to the other allele, **Genotyped in ARIC and pass QC for imputation (0 = no, 1 = yes, 2 = yes but does not pass QC for imputation)

LDL = low density cholesterol, HDL = high density cholesterol, TC = total cholesterol, TG = triglycerides, SBP = systolic blood pressure, DBP = diastolic blood pressure, A1 = first allele in ARIC, A2 = second allele in ARIC, MAF = minor allele frequency,

CHAPTER 4: GENETIC RISK ASSESSMENT OF CORONARY PLAQUE BURDEN

Preface to the Manuscript

This manuscript presents the last part of a study investigating the genetic basis of cardiovascular disease by examining a multi-locus genetic risk score of the strongest predictor of the major cause of death of cardiovascular disease.

This third and last study focuses on determining whether CAD associated loci collectively facilitate the formation of coronary plaque in a monotonic fashion throughout the life course. The second manuscript showed the benefit to combine genetic information at established susceptibility loci in any sets of non-genetic predictive models. Given these findings and the fact that genetic risk score improves overall risk discrimination, we felt that it was important to answer this question last because this genetic risk assessment could add incremental prognostic value to the most common type of cardiovascular disease above traditional risk scores across a range of ages.

We selected genotyped and imputed independent markers from the CARDIOGRAMplusC4D report that had reached genome-wide significance at any time during the GWAS era to construct the genetic risk score in White/European subjects from dbGaP genetic data (SEA, FHS, and MESA) as well as from the Stanford-Kaiser ADVANCE study. We stratified white/European subjects within each study into one of five age groups (≤30, 31-45, 46-60, 61-75, >75 years) and defined cases as subjects with either any raised lesions in their right coronary artery on autopsy (SEA study) or with an age and sex specific CAC score >75th percentile (all other studies, age > 30 years)

In this context, the general objective of this third study is to show that susceptibility alleles for clinical CAD uncovered through large-scale meta-analysis of GWAS predispose an individual to clinical complications of CAD from birth through the modulation of the rate of formation of coronary plaque.

The specific objective of this second study is:

1. To identify cohorts of subjects with subclinical atherosclerosis identified by either pathologic examination of the coronary arteries or by radiographic assessment of coronary artery calcification.

2. To test association between a multi-locus genetic risk score based on susceptibility variants for CAD with the presence of subclinical atherosclerosis among subjects with no previous history of clinical CAD, for each 15-year age categories.

3. To compare association between weighted and un-weighted genetic risk score with development of plaque.

4. To compare association between a genetic risk score including all susceptibility variants for CAD to a restricted genetic risk score to non-risk factor SNPs with the presence of subclinical atherosclerosis.

Susceptibility loci for clinical CAD and subclinical coronary atherosclerosis throughout the life course

Elias Salfati MS MPH[1,2], Steve Fortmann MD[1,3], Steve Sidney MD[4], Mark A. Hlatky MD[1], Thomas Quertermous MD[1], Alan S Go MD[4], Carlos Iribarren MD MPH PhD[4], Benjamin A. Goldstein, PhD[1] ,Themistocles L. Assimes MD, PhD[1*]

[1]Department of Medicine, Stanford University School of Humanities and Sciences, Stanford, CA USA 94305, [2]Ecole Doctorale B2T, IUH; Université Paris 7, Paris, France

[3]The Kaiser Permanente Center of Health Research, Portland, OR 97227

[4]Kaiser Permanente Division of Research, Oakland, CA 94612

4.1 Abstract

Recent genome wide association studies (GWAS) have identified 49 single nucleotide polymorphisms (SNPs) associated with clinically significant complications of CAD including myocardial infarction (MI), CABG, PCI, and/or angina. The mechanism by which these loci influence the risk of clinical CAD remains largely unclear. We hypothesized that variants at these loci collectively facilitate the formation of coronary plaque in a monotonic fashion throughout the life course. We used genetic data from dbGAP (SEA, FHS, and MESA) as well as from the Stanford-Kaiser ADVANCE study imputed to the 1000 genomes project to examine the association between a genetic risk score (GRS) of high-risk alleles at these 49 SNPs and the presence of subclinical atherosclerosis. Subclinical atherosclerosis was identified by either pathologic examination of the coronary arteries or by radiographic assessment of coronary artery calcification (CAC). We stratified white/European subjects within each study into one of five age groups (≤30, 31-45, 46-60, 61-75, >75 years) and defined cases as subjects with either any raised lesions in their right coronary artery on autopsy (SEA, 26.7% subjects aged 18 to 30 years at time of unexpected death) or with an age and sex specific CAC score >75th percentile (all other studies, age > 30 years). Among 1561 cases and 5068 controls, we found a one SD increase in the GRS was associated with a 28% increased risk of having advanced subclinical coronary atherosclerosis ($p = 1.43 \times 10^{-16}$). This increase in risk was significant in every age stratum ($.01 > p > 9.4 \times 10^{-7}$) and was remarkably similar across all age strata (p test of heterogeneity = 0.98). We obtained near identical results and levels of significance when we restricted the GRS to 32 SNPs not associated with traditional risk factors. Our findings strongly support the notion that susceptibility alleles for clinical CAD uncovered through large-scale meta-analysis of GWAS uniformly promote the development of coronary atherosclerosis from birth. This predisposition is sustained at a constant level throughout one's lifetime. Given it is observed at the earliest stage of plaque formation, it is unlikely to involve a concurrent predisposition to plaque rupture and/or thrombosis.

4.2 Introduction

Coronary artery disease (CAD) is a primary cause of death and disability worldwide[156]. Development of CAD is triggered by complex interactions of many environmental and inherited factors. Established traditional risk factors, including age, sex, high cholesterol, diabetes, smoking, and elevated blood pressure, explain only a fraction of the variability in disease risk and have limitations in their ability to discriminate individuals likely to experience CAD[157, 158]. More recently, coronary arterial calcification (CAC) as measured by computed tomography (CT) has been established as a major predictor for clinical complications of CAD including myocardial infarction (MI), independent of traditional risk factors[159-163]. CAC quantifies the degree of calcified plaque within the vessel wall of coronary arteries and has been shown to be highly correlated with the overall plaque burden of both calcified and non-calcified plaque[164-166].

The search for new predictors that may identify individuals who have an inherited predisposition to develop CAD has recently accelerated dramatically. Genome wide association studies (GWAS) over the last 7 years have identified 49 susceptibility variants robustly associated with clinically significant complications of CAD including myocardial infarction (MI), coronary artery bypass graft surgery (CABG), percutaneous coronary intervention (PCI), and/or angina[62-66, 68, 70, 167-169]; yet these SNPs account for small proportion of the overall genetic variance of CAD underscoring the polygenic nature of the disease. While the effect on risk of each of these variants is small, they are independent and additive. Accordingly, the relatively small effects of the high-risk alleles at these 49 identified loci can be aggregated into a single powerful predictor of clinical CAD through a multi-locus genetic risk score (GRS) [170]. Morrison et al. was among the first to illustrate the concept of GRS in the context of predicting CAD prior to the GWAS era, utilizing SNPs within candidate genes[80]. Over the years, several studies have examined the utility of combining multiple causal variants to improve the identification of subjects at increased risk of clinical CAD[115, 143, 170-172].

The mechanism by which CAD loci discovered through GWAS influence the risk of clinical CAD remains largely unclear. We hypothesized that variants at these loci collectively facilitate the formation of coronary plaque in a monotonic fashion throughout

the life course. To test this hypothesis, we investigated the association between a GRS based on the 49 susceptibility variants for CAD with the presence of subclinical atherosclerosis among subjects with no previous history of clinical CAD. Subclinical atherosclerosis was estimated in most subjects using CAC and in the youngest subjects through pathologic examination of the coronary arteries.

4.3 Material and Methods

4.3.1 Study Population

The Atherosclerotic Disease, VAscular functioN, and genetiC Epidemiology (ADVANCE) [173-175] study served as our initial cohort for this analysis. ADVANCE included a subset of 479 participants from the Coronary Artery Risk Development in Young Adults Study (CARDIA) [176], study originally recruited at the Oakland field center who attended the study's Year 15 examination in 2000–2001. We extended our analyses to several clinical and genetic datasets available through the NCBI's database of Genotypes and Phenotypes (dbGAP) including the SNPs and the Extent of Atherosclerosis (SEA) [177], the Multi Ethnic Study of Atherosclerosis (MESA) [178], the Framingham Heart Study (FHS) [179], and the Cardiovascular Health Study (CHS) [180, 181]. As all SNPs to be tested were identified in European and/or South Asian populations, we restricted our analyses to study participants who self-reported European ancestry. Detailed descriptions of the design and methods for the five studies have been published elsewhere[182-188]. The subset of participants from these studies all had either a post mortem pathologic determination of the degree of subclinical coronary atherosclerosis (SEA) or an assessment of subclinical coronary atherosclerosis through a CAC study at a point in time when they reported no history of clinical CAD. We stratified subjects within each study into one of five age groups at the time of assessment of subclinical coronary atherosclerosis (≤30, 31-45, 46-60, 61-75, >75 years) .

4.3.2 Case definition

In the majority of subjects with a measure of CAC, we defined cases within each study and age stratum as those subjects possessing an age and sex specific CAC score greater than the 75 percentile[189]. For some younger subgroups where the 75% percentile

CAC score was still 0, we defined cases as subjects with a CAC score > 0. For participants in SEA, we defined cases as subjects with any raised lesions in their right coronary artery on autopsy.

4.3.3 SNP Selection and Imputation

For the construction of the GRS, we selected independent SNPs (LD-pruned, $r^2 < 0.2$) from Supplementary Table 9 of the CARDIOGRAMplusC4D report that had reached genome-wide significance at any time during the GWAS era (referred to as either "known" or "novel" CAD GWAS SNPs). Of the 153 "low FDR" SNPs for CAD included in this table, 49 SNPs met these criteria (Supplemental Table). As expected, only a small fraction of the 49 SNPs were genotyped in participants of the five studies given the various arrays used. To minimize the need to search for proxies, we used the genotype data within each study to impute the SNPs in the latest release of the 1000 Genomes Project reference haplotypes. Each study was imputed separately using MaCH (v1.0.18.c) [103] and Minimac (2013-07-17) software[190]. A two-stage imputation procedure was followed. First, we used MaCH to phase individuals across chromosome to estimate the haplotypes. Phasing did not require reference panels as input. We excluded SNPs with minor allele frequency (MAF) < 0.01, Hardy-Weinberg equilibrium $P < 1 \times 10^{-6}$, call rate < 95% or large allele frequency discrepancies compared to the 1000 Genomes Project reference data using PLINK[191]. To ensure good quality of phased haplotypes, MaCH was run with the 20 rounds and 200 states for parameter estimation. Then, the phased haplotypes were compared to 1000 Genomes Project haplotypes (version 3 March 2012 release, 2184 haplotypes) Cosmopolitan panel (246 AFR + 181 AMR + 286 ASN + 379 EUR) for imputation using the OpenMP protocol based multi-threaded version of Minimac software with 20 rounds and 300 states for each chromosome.

4.3.4 Construction of the GRS

Next, we calculated both weighted and un-weighted GRS for each individual using their imputed genotype dosage of the number of high-risk alleles for each of the 40

SNPs. We then combined the information on the 49 SNPs using an allele count model. Our weighted multilocus GRSs (wGRSs) was calculated by taking the sum of the product for each SNP of the high risk allele dosage with the effect size observed (β) in the CARDIOGRAMplusC4D meta-analysis (PMID 23202125). The value of the weighted risk score on *n* SNPs is formulated as follow:

$$wGRS = \sum_{j}^{n} w_j \, x_{ij}$$

Where, x_{ij} is the dose of the coded allele at the *j*-th SNP in the *i*-th subject, and w is the effect of the *j*-th SNP. GRS scores were then standardized to a mean of zero and SD = 1.

We further calculated an un-weighted and weighted GRS restricted to the 32 SNPs that appears not be related to traditional risk factors as per the assessment of the CARDIoGRAM+C4D consortium which reported that 17 of the 49 SNPs also showed significant trends for association for either lipids (n = 12 SNPs) or blood pressure (n = 5).

4.3.5 Statistical Analysis

We used logistic regression to estimate the association between a GRS and case-control status. Despite the use of age and sex specific CAC score cutoffs to define cases, we further adjusted for age and sex to account for any residual confounding by these variables with the 15-year age categories. We also estimated the ORs for case-control status for subjects with the highest quintile of GRS compared to subjects within the lowest quintile of GRS to quantify the added risk in the extremes of the GRS.

A fixed-effect meta-regression model was used for estimating the overall effect within each age stratum and across all age-strata combined. The calculations yielded chi-squared statistic and is its degrees of freedom for testing the heterogeneity and the overall estimate for the fixed-effect model. To assess inconsistency across cohorts, a statistic describing the percentage of the variability in effect estimates due to heterogeneity rather than sampling error was calculated as follow:

$$\mathcal{I}^2 = \left(\frac{Q - df}{Q}\right) \times 100$$

Here Q is the chi-squared statistic and df is its degrees of freedom.
Calculations were performed with the metafor package 1.9-4 in R[136] (version 3.0.2) [192].

4.4 Results

Characteristics of the participants included in this study are summarized in Table 1, stratified by study and age-stratum. After quality control, we identified a total of 6 910 participants of white/European ancestry from all studies covering a very broad age range at the time of assessment of subclinical atherosclerosis (18 to >85). The largest cohort was Framingham Heart Study contributing a total of 3 131 individuals, and the two smallest cohorts contributing were CARDIA and CHS with 286 and 151 individuals (Table 1). A total of a 1 561 (23%) of subjects across all cohorts were defined as cases based on our case definitions.

Table 1 summarizes descriptive statistics for the 6 populations tested as well as quality of the imputation for the 49 SNPs used in the construction of the GRSs. The proportion of well-imputed SNPs was high (>93.9%) except SEA where only 34.69% of the SNPs had an imputation -R^2 score > 0.3. Imputation quality index ranged from 26.97% (SEA) to 96.87% (ADVANCE). None of the SNP showed significant departures from Hardy-Weinberg equilibrium in any population cohort. The overall mean 49 SNPs GRS was 48.9±4.1 risk alleles and the mean 32 SNPs GRS was 29.5±3.5 risk alleles (Table 1). Reassuringly, we found similar mean GRS across studies and across all age-strata.

Figure 1 summarizes our association results for our un-weighted GRS including all 49 SNPs. We found that this GRS was significantly associated with case-control status across all age groups and studies combined even after adjusting for age and sex, as well as in the meta-analysis. The meta-analysis results demonstrated a 28% increase in risk of being a case with each SD increase in the GRS (95% CI: 1.21-1.36, $p=1.43\times10^{-16}$). . The increase in risk was significant in every age stratum (meta-analysis $.01 > p > 9.4 \times 10^{-7}$)

and was remarkably similar across all age strata (p test of heterogeneity = 0.98). Association results for the GRS restricted to the 32 non-risk factor SNPs (GRS32NRF) did not differ substantially from our primary analysis that included all 49 SNPs (OR 1.29 per SD increase in the GRS, 95% CI 1.22-1.38; $p=6.6\times10^{-18}$) (Figure 2). Once again, the increase in risk was significant in every age stratum ($.014 > p > 2.73 \times 10^{-7}$) and was remarkably similar across all age strata. Effect estimates were consistent across all cohorts. Results were similar for our analyses using the weighted GRS (supplemental Figures 1, 2) but with evidence of increased heterogeneity in the ORs observed. The p value for the test of heterogeneity was still not significant (p = 0.55 and 0.23). The largest changes in the ORs occurred in the strata with the smallest number of cases.

Figure 3 summarizes the distribution of degree of raised lesions in the SEA participants. A large majority of subjects had only 1 to 30% of their RCA covered with raised lesions. Table 2 summarizes the association results for SEA using 3 additional imputation quality cutoffs including $r^2 > 0.3$, $r^2 > 0.5$, and $r^2 > 0.8$. The strongest association was observed for a cutoff of $r^2 > 0.3$ which allowed for 17 of the 49 SNPs to be used in the construction of the GRS.

Participants in the highest quintile of GRS had almost double the risk of advanced subclinical coronary atherosclerosis compared with those in the lowest quintile (OR 1.88, 95% CI 1.80 – 1.96, $p=5.5\times10^{-14}$). We obtained similar for the GRS32 (OR 1.80, 95% CI 1.72 – 1.88, $p=4.6\times10^{-12}$).

4.5 Discussion

We found that a multi-locus GRS derived from the high risk alleles of SNPs associated with clinical complications of CAD is strongly associated with the presence of advanced subclinical atherosclerosis as estimated by either the direct visualization of raised lesions within the right coronary artery or the degree of CAC within all three coronary arteries combined. Among subjects without a history of clinical CAD, the association is evident starting in young adulthood and persists throughout the life course to a degree that is remarkably homogenous. An

increase in the GRS by one standard deviation was associated with a ~30% increase in the risk of being in the quartile with the highest degree of subclinical coronary atherosclerosis.

Our findings have several important implications. First, they reinforce findings from several epidemiological studies demonstrating that a subclinical measure of plaque burden within the coronaries not only adds incremental prognostic value above traditional risk scores across a range of ages but also is the single strongest predictor of clinical complications of CAD in previously asymptomatic persons[193-202]. Second, they are consistent with a hypothesis that GWAS susceptibility loci for CAD discovered to date predispose an individual to clinical complications of CAD from birth through the modulation of the rate of formation of coronary plaque. This predisposition appears monotonic and continues unabated throughout one's lifespan with evidence of predisposition persistent into the 9th decade of life. Third, this predisposition is not likely to involve a specific predisposition to intra-plaque rupture, intraluminal plaque rupture, or thrombosis because it appears at a very young age when plaques are generally too small to be prone to any of these structural complications[203]. This hypothesis is best supported by the associations observed in SEA, which were largely driven by the presence of minimally raised lesions in a small fraction of the overall surface area of the right coronary artery.

We performed three sensitivity analyses. First, we repeated all analyses after removing 17 SNPs from the GRS that are most likely influencing clinical CAD through effects on traditional risk factors including dyslipidemia or elevated blood pressure[62]. Dyslipidemia in particular has already been unequivocally tightly linked to the rate of development of subclinical atherosclerosis both in human and animal studies[204-207]. We performed this sensitivity analysis to ensure that the effect of this subgroup of 17 SNPs was not driving our overall results. We found

that the strength of the associations overall and within each age stratum as well as the homogeneity of effects persisted even after excluding these SNPs from the GRS. These findings suggest that most, if not all, of the novel mechanisms predisposing to clinical complications of CAD uncovered by GWA studies to date involve physiological processes that facilitate the formation of coronary plaque.

Second, we tested whether weighting the high-risk alleles in the GRS by the effect size observed for those alleles in GWA studies influenced our results[208]. While this approach slightly increased the overall effect size and the statistical significance of the association between our GRS and subclinical CAD, it introduced some heterogeneity although our test of heterogeneity remained insignificant. Compared to a non-weighted GRS, a weighted GRS would be expected to improve an association when risk loci with a range of effects contribute to disease and when the GRS is being tested on the exact same phenotype that was used to identify the high risk alleles[208]. Conceivably, a weighted GRS may be less helpful and possibly even harmful in a situation where it is being tested on a phenotype that is different from the one used to identify the high-risk alleles as was done in this analysis.

Our third sensitivity analysis was focused on SEA. This study was unique not only because of the method of assessment of subclinical atherosclerosis but also because of the platform used for genome wide genotyping which was an early GWAS array by Perlegen that included only ~106000 SNPs. Furthermore, only ~2/3 of these SNPs passed our standard pre-imputation quality control. Consequently, the mean imputation quality for SEA was significantly lower than all other studies with only ~1/3 of SNPs having an imputation $r^2 > 0.3$. In the 4 models we tested, we invariable found the highest point estimates and the lowest p values for the analyses that restricted the GRS to the subset of SNPs with an imputation $r^2 > 0.3$ although these differences were not large when compared to

the models that used all 49 SNPs. Imposing an even higher imputation score threshold led to a further substantial restriction of SNPs as well as a noticeable degradation of the association signal. These circumstances suggest that the statistical significance of the association we observed in SEA with the GRS incorporating all 49 or all 33 non-risk factor SNPs likely represents a substantial underestimate of the true p value given the greater difficulty in accurately imputing genotypes.

Our study has a couple of important limitations. First, power was more limited in the extreme age categories where overall number of subjects available for study was lower. Nevertheless, we observed nominally significant associations in these age strata both in our main analysis and in our sensitivity analyses. Second, identification of the quartile of subjects with the highest degree of subclinical atherosclerosis was hampered in the younger age strata (30 to 45, 45 to 60 years) by a low prevalence of subjects with any CAC. Thus, the percentile of subjects with CAC > 0 was less than 25 and the size of the case group ranged from 11.9% to 24.1% of the stratum. We elected not to reclassify a random set of subjects with a CAC = 0 into the case group. Conceivably, the ORs of association in these strata may be biased towards the null because the control group includes some subjects with a burden of disease that is within the top quartile. We suspect this bias, if present, is minimal and would not change our conclusions. Of note, this issue was not a problem in SEA given the presence of any raised plaque was coincidently observed in about one quartile of the cohort.

Why would contemporary GWA studies of CAD involving predominantly (and often exclusively) subjects with clinical complications of CAD be identifying *only* loci predisposing to plaque formation? We propose two reasons. First, case-control GWA studies that have identified the 49 loci for CAD may not allow for the detection of other types of susceptibility loci. A key design principle in case-

control studies is the requirement for controls to be at risk of the outcome[209]. Risk factors for the outcome of interest cannot be identified if controls are not at risk. For example, the most appropriate controls for a study trying to identify whether using a cell phone while driving increases the risk of a car accident are drivers with a cell phone who did not get into an accident. Subjects who do not drive or do not own a cell phone are not appropriate controls as they are not at risk of suffering a car accident while using a cell phone. Similarly, loci that predispose to plaque rupture and/or thrombosis cannot be identified if controls are not at risk of these complications because they have minimal or no underlying atherosclerosis. A more effective design to identify such loci would be to compare subjects with a critical amount of CAD and one or more well-documented MIs to subjects with a similar amount of CAD but no history of MI. Such a design has been used in the recent past to identify one locus (ABO) that may predispose to MI[210]. However, the same locus was later identified in a more conventional GWA case-control study of CAD casting doubt on its specificity for susceptibility to plaque rupture or thrombosis[63]. Furthermore, such a design may also be substantially underpowered due to misclassification of controls as many ruptures are observed even in the absence of clinical symptoms[211-213].

In summary, we have shown that susceptibility loci for CAD discovered to date through GWAS appear to predispose to clinical CAD by exclusively facilitating the formation of coronary artery plaque and not by promoting plaque rupture or thrombosis. This susceptibility to plaque formation is life long, remarkably homogenous, and not driven by exposure to traditional risk factors. The identification of loci that predispose to plaque rupture or thrombosis is extremely challenging given non-invasive tools to reliably classify whether someone with CAD has suffered such an event do not exist. Investigators examining the mechanism of associations of established CAD loci should take these observations into consideration when designing their experiments.

Legends to Figures:

Figure 1: The forest plot show the meta-analysis of the association of the un-weighted genetic risk score of 49 SNPs with CAC. The horizontal axis indicates the odds ratio for CAC per SD unit increase in the standardized genetic risk score. SEA indicates SNPs and the Extent of Atherosclerosis; ADVANCE-CARDIA, Atherosclerotic Disease, VAscular functioN, and genetiC Epidemiology-Coronary Artery Risk Development in Young Adults Study; FHS, Framingham Heart Study; MESA, Multi Ethnic Study of Atherosclerosis; ADVANCE, Atherosclerotic Disease VAscular functioN and genetiC Epidemiology; CHS, Cardiovascular Health Study.

Figure 2: The forest plot show the meta-analysis of the association of the un-weighted genetic risk score of 32 non-risk factors SNPs with CAC. The horizontal axis indicates the odds ratio for CAC per SD unit increase in the standardized genetic risk score. SEA indicates SNPs and the Extent of Atherosclerosis; ADVANCE-CARDIA, Atherosclerotic Disease, VAscular functioN, and genetiC Epidemiology-Coronary Artery Risk Development in Young Adults Study; FHS, Framingham Heart Study; MESA, Multi Ethnic Study of Atherosclerosis; ADVANCE, Atherosclerotic Disease VAscular functioN and genetiC Epidemiology; CHS, Cardiovascular Health Study.

Figure 3: Distribution of the percentage of Raised Coronary Artery (RCA) lesions in SEA (SNPs and the Extent of Atherosclerosis) study.

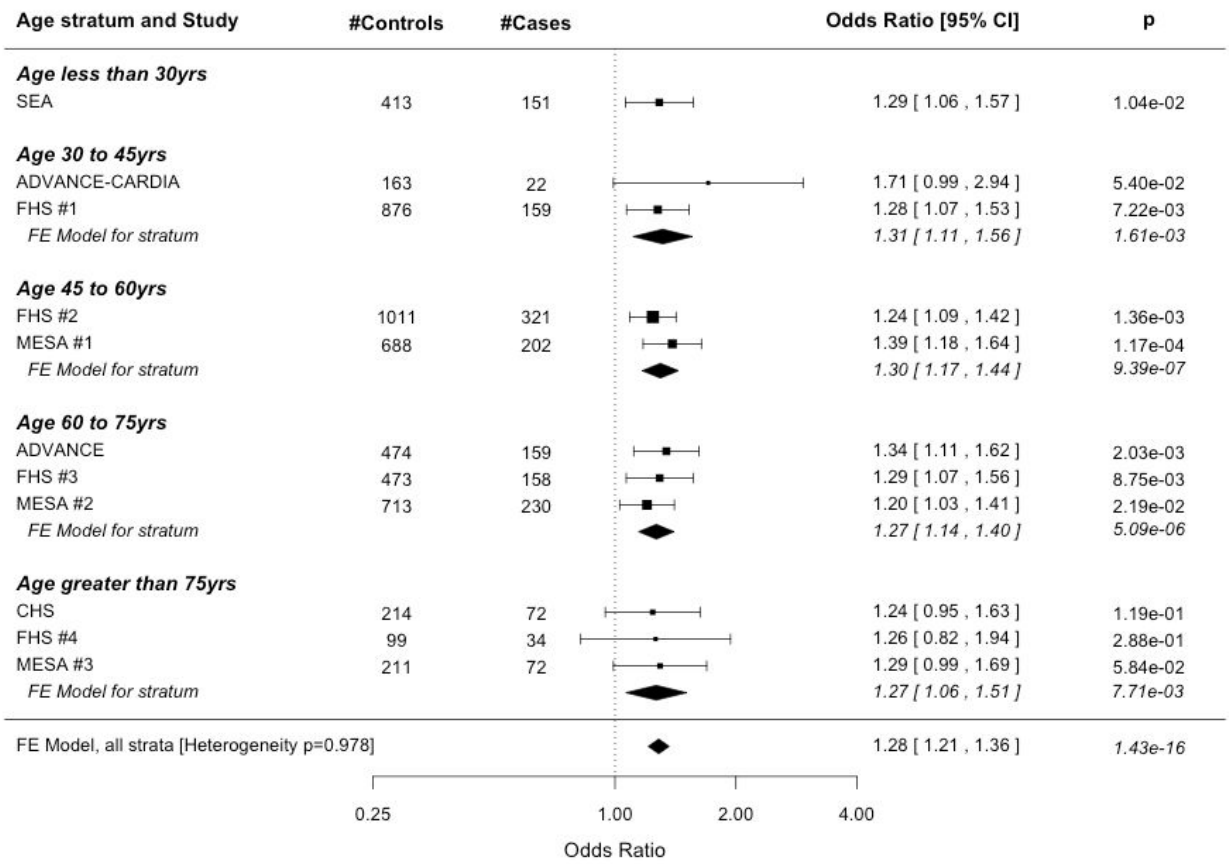

Age stratum and Study
#Controls
#Cases
Odds Ratio [95% CI]
p
Age less than 30yrs
SEA 413 151 1.29 [1.06 , 1.57] 1.04e-02
Age 30 to 45yrs
ADVANCE-CARDIA 163 22 1.71 [0.99 , 2.94] 5.40e-02
FHS #1 876 159 1.28 [1.07 , 1.53] 7.22e-03
FE Model for stratum 1.31 [1.11 , 1.56] 1.61e-03
Age 45 to 60yrs
FHS #2 1011 321 1.24 [1.09 , 1.42] 1.36e-03
MESA #1 688 202 1.39 [1.18 , 1.64] 1.17e-04
FE Model for stratum 1.30 [1.17 , 1.44] 9.39e-07
Age 60 to 75yrs
ADVANCE 474 159 1.34 [1.11 , 1.62] 2.03e-03
FHS #3 473 158 1.29 [1.07 , 1.56] 8.75e-03
MESA #2 713 230 1.20 [1.03 , 1.41] 2.19e-02
FE Model for stratum 1.27 [1.14 , 1.40] 5.09e-06
Age greater than 75yrs
CHS 214 72 1.24 [0.95 , 1.63] 1.19e-01
FHS #4 99 34 1.26 [0.82 , 1.94] 2.88e-01
MESA #3 211 72 1.29 [0.99 , 1.69] 5.84e-02
FE Model for stratum 1.27 [1.06 , 1.51] 7.71e-03
FE Model, all strata [Heterogeneity p=0.978] 1.28 [1.21 , 1.36] 1.43e-16
0.25
1.00
2.00
4.00
Odds Ratio

Figure 1.

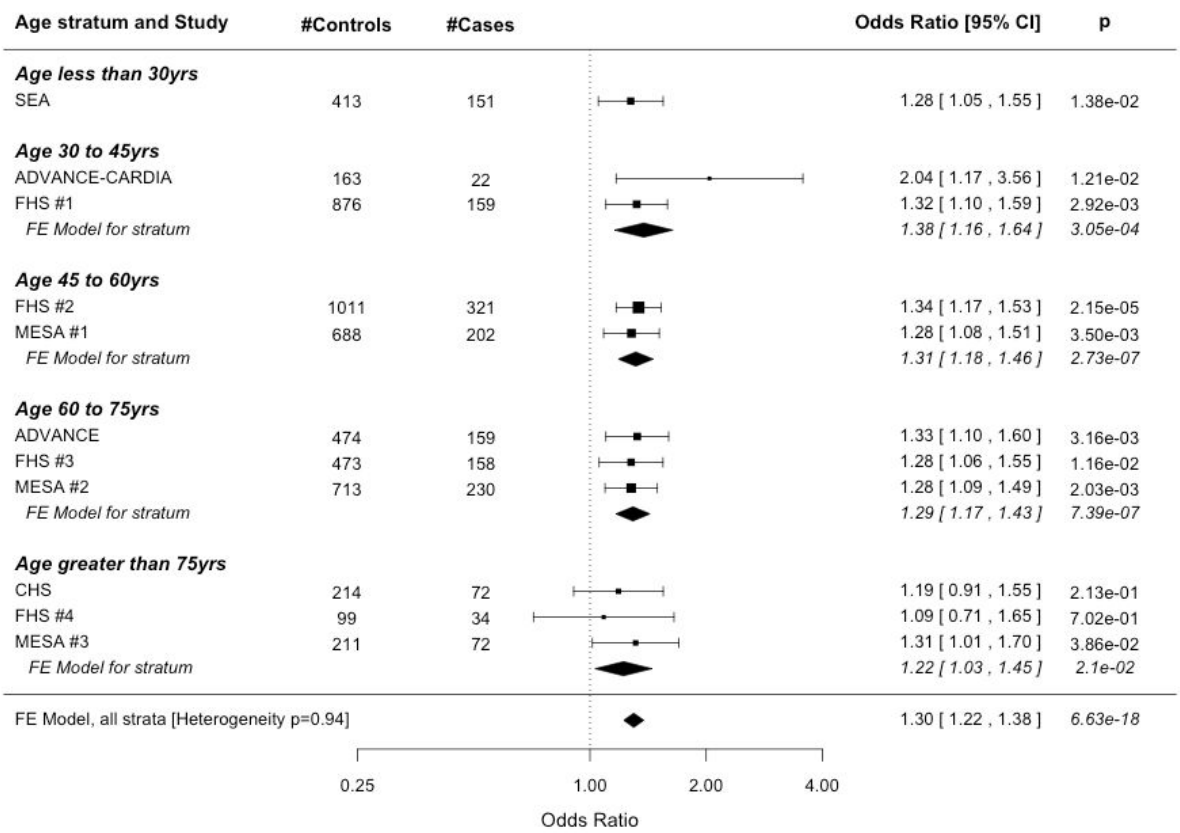

Age stratum and Study
#Controls
#Cases
Odds Ratio [95% CI]
p
Age less than 30yrs
SEA 413 151 1.28 [1.05 , 1.55] 1.38e-02
Age 30 to 45yrs
ADVANCE-CARDIA 163 22 2.04 [1.17 , 3.56] 1.21e-02
FHS #1 876 159 1.32 [1.10 , 1.59] 2.92e-03
FE Model for stratum 1.38 [1.16 , 1.64] 3.05e-04
Age 45 to 60yrs
FHS #2 1011 321 1.34 [1.17 , 1.53] 2.15e-05
MESA #1 688 202 1.28 [1.08 , 1.51] 3.50e-03
FE Model for stratum 1.31 [1.18 , 1.46] 2.73e-07
Age 60 to 75yrs
ADVANCE 474 159 1.33 [1.10 , 1.60] 3.16e-03
FHS #3 473 158 1.28 [1.06 , 1.55] 1.16e-02
MESA #2 713 230 1.28 [1.09 , 1.49] 2.03e-03
FE Model for stratum 1.29 [1.17 , 1.43] 7.39e-07
Age greater than 75yrs
CHS 214 72 1.19 [0.91 , 1.55] 2.13e-01
FHS #4 99 34 1.09 [0.71 , 1.65] 7.02e-01
MESA #3 211 72 1.31 [1.01 , 1.70] 3.86e-02
FE Model for stratum 1.22 [1.03 , 1.45] 2.1e-02
FE Model, all strata [Heterogeneity p=0.94] 1.30 [1.22 , 1.38] 6.63e-18
0.25
1.00
2.00
4.00
Odds Ratio

Figure 2.

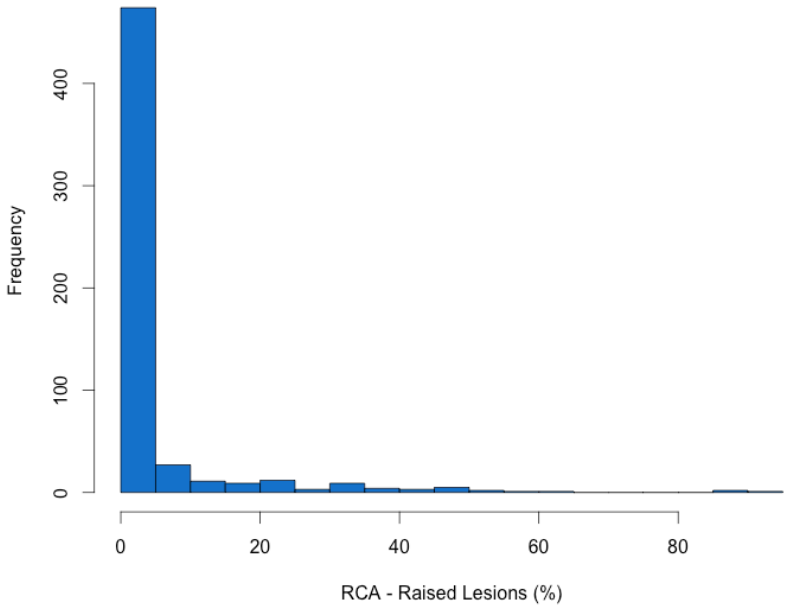


Figure 3.

Legends to Tables:

Table 1: Descriptive statistics for each study per 15-year age categories and quality of the imputation for the 49 SNPs used in the construction of the genetic risk score. SEA indicates SNPs and the Extent of Atherosclerosis; ADVANCE-CARDIA, Atherosclerotic Disease, VAscular functioN, and genetiC Epidemiology-Coronary Artery Risk Development in Young Adults Study; FHS, Framingham Heart Study; MESA, Multi Ethnic Study of Atherosclerosis; ADVANCE, Atherosclerotic Disease VAscular functioN and genetiC Epidemiology; CHS, Cardiovascular Health Study.

Table 2: Beta coefficient for the association between the genetic risk score of 49 and 33 SNPs respectively and CAC, using different cutoffs of imputation-R^2 for both weighted and un-weighted score.

AGE	< 30	30-45		45-60		60-75			> 75		
STUDY	***SEA***	***ADVANCE-CARDIA***	***FHS***	***FHS***	***MESA***	***ADVANCE***	***FHS***	***MESA***	***CHS***	***FHS***	***MESA***
N	564	151	1035	1332	904	633	631	956	286	133	285
Age (mean)	26.69	41.17	40.48	51.53	52.68	65.82	66.54	67.86	80.11	78.33	78.96
Age (s.d.)	5.01	2.98	3.01	4.27	4.506701	2.88	4.36	4.17	3.54	2.46	2.32
Female %	22.70	31.79	40.58	51.35	52	36.49	54.35	48.74	62.23	54.88	53.68
N_0 of Cases	151	18	159	321	202	159	158	230	72	34	72
% Cases	26.8%	11.9%	15.4%	24.1%	22.3%	25.1%	25.0%	24.1%	25.2%	25.6%	25.3%
CAC (mean)/ *Cases*	15.22%*	52.21	107.3	270.4	208.5	979.7	988.6	860.6	1742.8	1785	1406
CAC (mean)/ *Controls*	0.041%*	0	0	5.42	2.723	83.05	92.71	61.55	296.24	306.5	158.7
Platform	Perlegen	HumanHap550v1.1	Affymetrix 500K +50K	Affymetrix 500K +50K	Affymetrix 6.0	Metabochip	Affymetrix 500K +50K	Affymetrix 6.0	Human CNV370v1	Affymetrix 500K +50K	Affymetrix 6.0
N_0 of SNPs on Array	106,285	561,466	549,782	549,782	909,622	196,725	549,782	909,622	339,971	549,782	909,622
N_0 of 49 GRS SNPs on Array	0	21	11	11	18	45	11	18	12	11	18
N_0 of SNP used for Imputation	66,166	513,729	284,965	284,965	604,312	107,809	284,965	604,312	275,298	284,965	604,312
Average R^2 for imputed SNPs	0.2697	0.8803	0.853	0.853	0.8652	96.87	0.853	0.8652	0.7968	0.853	0.8652
Proportion of Imputed SNPs with R^2> 0.3	34.69	100	93.87	93.87	95.91	100	93.87	95.91	95.91	93.87	95.91
GRS (mean)	48.78	49.1	48.87	48.87	49.2	49.11	48.87	49.2	48.59	48.87	49.2
GRS (s.d.)	2.48	3.91	4.22	4.22	4.28	4.60	4.22	4.28	3.85	4.22	4.28

Table 1.

SEA	#SNPs	*Estimate*	*Std.Error*	*z_value*	*p*		*Imputed-R^2*
GRS49	***17***	0.253	0.099	2.562	1.04×10^{-2}	*	***0***
GRS49-weighted	***17***	0.202	0.098	2.06	3.94×10^{-2}	*	***0***
	13	0.236	0.098	2.393	1.67×10^{-2}	*	***0***
GRS32-weighted	***13***	0.182	0.098	1.856	6.34×10^{-2}	.	***0***
GRS49	***17***	0.306	0.099	3.075	2.11×10^{-3}	**	***0.3***
GRS49-weighted	***17***	0.249	0.098	2.528	1.15×10^{-2}	*	***0.3***
GRS32	***13***	0.304	0.099	3.078	2.09×10^{-3}	**	***0.3***
GRS32-weighted	***13***	0.245	0.098	2.49	1.28×10^{-2}	*	***0.3***
GRS49	***10***	0.272	0.099	2.753	5.9×10^{-3}	**	***0.5***
GRS49-weighted	***10***	0.212	0.098	2.168	3.01×10^{-2}	*	***0.5***
GRS32	***8***	0.237	0.098	2.411	1.59×10^{-2}	*	***0.5***
GRS32-weighted	***8***	0.184	0.097	1.89	5.88×10^{-2}	.	***0.5***
GRS49	***4***	0.271	0.098	2.751	5.94×10^{-3}	**	***0.8***
GRS49-weighted	***4***	0.244	0.098	2.498	1.25×10^{-2}	*	***0.8***
GRS32	***3***	0.198	0.097	2.046	4.08×10^{-2}	*	***0.8***

Table 2.

Legend to Supplementary Figures:

Figure 1: The forest plot show the meta-analysis of the association of the weighted genetic risk score of 49 SNPs with CAC. The horizontal axis indicates the odds ratio for CAC per SD unit increase in the standardized genetic risk score. SEA indicates SNPs and the Extent of Atherosclerosis; ADVANCE-CARDIA, Atherosclerotic Disease, VAscular functioN, and genetiC Epidemiology-Coronary Artery Risk Development in Young Adults Study; FHS, Framingham Heart Study; MESA, Multi Ethnic Study of Atherosclerosis; ADVANCE, Atherosclerotic Disease VAscular functioN and genetiC Epidemiology; CHS, Cardiovascular Health Study.

Figure 2: The forest plot show the meta-analysis of the association of the weighted genetic risk score of 32 non-risk factors SNPs with CAC. The horizontal axis indicates the odds ratio for CAC per SD unit increase in the standardized genetic risk score. SEA indicates SNPs and the Extent of Atherosclerosis; ADVANCE-CARDIA, Atherosclerotic Disease, VAscular functioN, and genetiC Epidemiology-Coronary Artery Risk Development in Young Adults Study; FHS, Framingham Heart Study; MESA, Multi Ethnic Study of Atherosclerosis; ADVANCE, Atherosclerotic Disease VAscular functioN and genetiC Epidemiology; CHS, Cardiovascular Health Study.

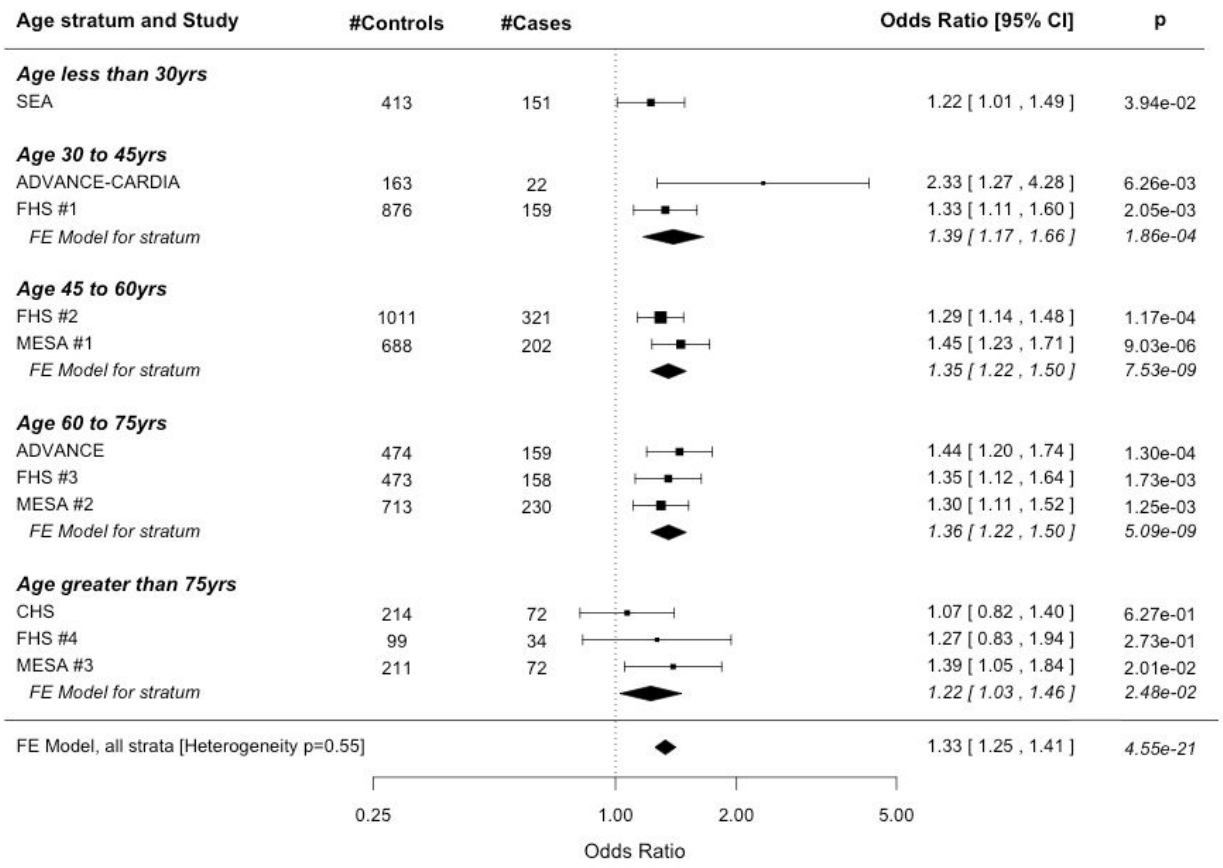

Age stratum and Study
#Controls
#Cases
Odds Ratio [95% CI]
p
Age less than 30yrs
SEA 413 151 1.22 [1.01 , 1.49] 3.94e-02
Age 30 to 45yrs
ADVANCE-CARDIA 163 22 2.33 [1.27 , 4.28] 6.26e-03
FHS #1 876 159 1.33 [1.11 , 1.60] 2.05e-03
FE Model for stratum 1.39 [1.17 , 1.66] 1.86e-04
Age 45 to 60yrs
FHS #2 1011 321 1.29 [1.14 , 1.48] 1.17e-04
MESA #1 688 202 1.45 [1.23 , 1.71] 9.03e-06
FE Model for stratum 1.35 [1.22 , 1.50] 7.53e-09
Age 60 to 75yrs
ADVANCE 474 159 1.44 [1.20 , 1.74] 1.30e-04
FHS #3 473 158 1.35 [1.12 , 1.64] 1.73e-03
MESA #2 713 230 1.30 [1.11 , 1.52] 1.25e-03
FE Model for stratum 1.36 [1.22 , 1.50] 5.09e-09
Age greater than 75yrs
CHS 214 72 1.07 [0.82 , 1.40] 6.27e-01
FHS #4 99 34 1.27 [0.83 , 1.94] 2.73e-01
MESA #3 211 72 1.39 [1.05 , 1.84] 2.01e-02
FE Model for stratum 1.22 [1.03 , 1.46] 2.48e-02
FE Model, all strata [Heterogeneity p=0.55] 1.33 [1.25 , 1.41] 4.55e-21
0.25
1.00
2.00
5.00
Odds Ratio

Figure S1.

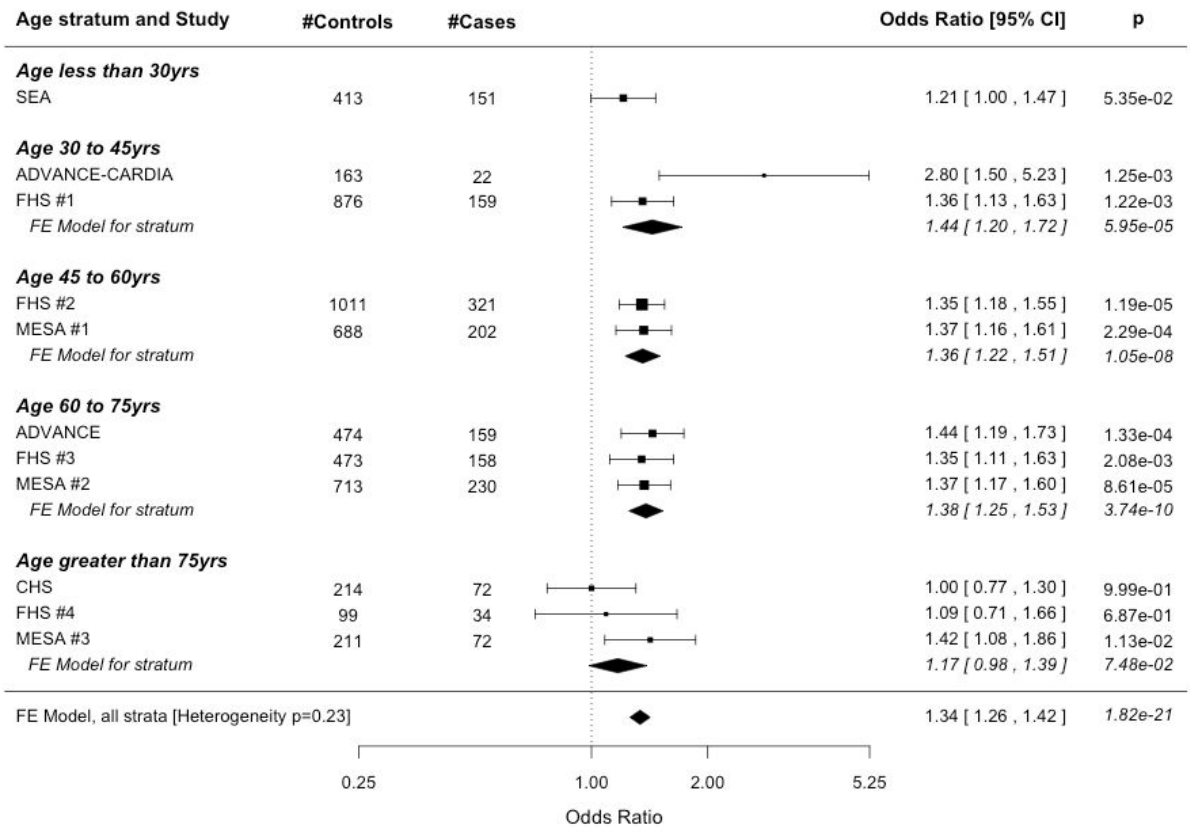

Age stratum and Study
#Controls
#Cases
Odds Ratio [95% CI]
p
Age less than 30yrs
SEA 413 151 1.21 [1.00 , 1.47] 5.35e-02
Age 30 to 45yrs
ADVANCE-CARDIA 163 22 2.80 [1.50 , 5.23] 1.25e-03
FHS #1 876 159 1.36 [1.13 , 1.63] 1.22e-03
FE Model for stratum 1.44 [1.20 , 1.72] 5.95e-05
Age 45 to 60yrs
FHS #2 1011 321 1.35 [1.18 , 1.55] 1.19e-05
MESA #1 688 202 1.37 [1.16 , 1.61] 2.29e-04
FE Model for stratum 1.36 [1.22 , 1.51] 1.05e-08
Age 60 to 75yrs
ADVANCE 474 159 1.44 [1.19 , 1.73] 1.33e-04
FHS #3 473 158 1.35 [1.11 , 1.63] 2.08e-03
MESA #2 713 230 1.37 [1.17 , 1.60] 8.61e-05
FE Model for stratum 1.38 [1.25 , 1.53] 3.74e-10
Age greater than 75yrs
CHS 214 72 1.00 [0.77 , 1.30] 9.99e-01
FHS #4 99 34 1.09 [0.71 , 1.66] 6.87e-01
MESA #3 211 72 1.42 [1.08 , 1.86] 1.13e-02
FE Model for stratum 1.17 [0.98 , 1.39] 7.48e-02
FE Model, all strata [Heterogeneity p=0.23] 1.34 [1.26 , 1.42] 1.82e-21
0.25
1.00
2.00
5.25
Odds Ratio

Figure S2.

CHAPTER 5: DISCUSSION

Since 2007, the GWAS approach has yielded more candidate regulators of cardiovascular traits and diseases than all of the genetics studies of the preceding era, yet the relative importance of these new loci to disease pathogenesis awaits future studies. We now have in hand a long list of genetic loci, harboring hundreds of genes, that enhanced our understanding of the genetic underpinnings of cardiovascular disease (CVD), but new approaches are needed to expose the missing variance[32, 88]. Focused studies investigating epistasis, gene-gene and/or gene-environment interactions, and rare variants in systematic and biologically plausible ways (such as through emphasis on genes in pathways) constitute novel alternative approaches[214]. For instance, most GWAS to date have been conducted in middle-aged and older adults so that the cumulative effects of multiple environmental effects or other gene-gene or gene-environment interactions in older age may have attenuated a modest but real genetic effect that may be more perceptible earlier in life. Such incomplete understanding of genetic and environmental causes and their interactions appeared to have puzzled those who attempted to identify a set of markers that could adequately explain or predict even a small fraction of complex diseases[215, 216]. Although, exhaustive epistasis examination imposes multiple comparisons, the examination between functionally related genes clustered in pathways would help reduce the multiples testing burden.

Next-generation re-sequencing approaches will be key to the discovery of rare and low frequency variants with potentially larger effects that influence CVD. With the expectation that many more yet undiscovered loci, possibly including variants in the rare allele spectrum that might have larger effect sizes[217], will contribute to explain the missing heritability for CVD. Another approach to discovering rare variants is to re-sequence genes harboring common variants[218] associated with CVD (as identified in GWAS). Since, most disease susceptibility loci identified by GWASs were found to be downstream targets of driver genes or were found in the boundary of gene regulatory networks[219], suggesting that major driver genes are often missed in traditional SNP analyses. Accordingly, genes harboring one trait-associated variant (of any allele

frequency) are more likely to contain additional variants altering their expression and/or function.

Additional insights into the lack of Blood Pressure (BP) variation may be explained through the study of ambulatory BP[220] or other approaches that provide more frequent measurements. Epigenetic modifications (including DNA methylation, histone modification, and alteration of microRNA expression) are also likely to contribute, as microRNAs have already been implicated in hypertension[221, 222] and could be key BP regulators by simultaneously influencing multiple genes. Epigenetic modifications constitute one hypothesized mechanism by which environmental factors interact with genes to influence BP. For instance, dietary factors cause epigenetic modifications. Therefore, increasing BMI through poor diet may influence BP through epigenetic modifications that alter expression patterns in the cell. Lastly, interrogating noncoding regions (such as regulatory elements) and structural variants by whole genome sequencing[223] may help reveal the "dark matter" of hypertension pathophysiology.

However, until there is direct molecular genetic evidence for these additional sources of genetic influence, missing heritability[88] is not clarified, and questions will remain about whether the heritability of BP have been overestimated by quantitative genetic studies.

The next principal challenges will be to define which of the susceptible markers are truly causal and to delineate the molecular mechanisms by which they influence atherosclerosis. Such study will ultimately yield important new genetic, epidemiological, and functional insights into the development of CVD.

One strategy to identify causal genes is to perform deep re-sequencing of positional candidate genes in the hopes of uncovering "smoking-gun" mutations (nonsense mutations that yield truncated protein products or missense mutations that alter amino acids critical to protein function) that are clearly linked to phenotype.

Future efforts along these lines might involve choosing the top and bottom strata of a prospective cohort study, or recruiting individuals who present to clinics with extreme phenotype (for qualitative traits: extreme presentations of disease or health), and sequencing all of the novel GWAS-nominated positional susceptible markers in these individuals, followed by genotype replication in a full prospective cohort study to confirm association. One important shortcoming of the re-sequencing strategy is that the failure to find smoking-gun mutations in a gene does not rule out its being a causal gene but may simply reflect that there are no naturally occurring mutations in the gene to be found in the study population. This could be because the gene is so important to normal development and function that a rare variant greatly perturbing the gene's function would not be tolerated in a viable organism. Another possibility is that variants do exist but in populations different from the study population (e.g., a different ethnic group). Multi-ethnic replications are useful in uncovering true susceptibility genes by identifying multiple significant top hits within a specific region, which is particularly valuable given allelic heterogeneity of the genetic effects[224] (different alleles may cause the disease in different populations).

Furthermore, causal variants likely exist in genes not identified by GWAS studies. This last point will soon be addressed by next-generation sequencing technology that will allow for whole-exome sequencing[225] (i.e., all exons of all genes in the genome) and, ultimately, whole-genome sequencing. Application of this approach to complex traits such as blood pressure and coronary disease is likely to expand the identification of genes and variants.

Undertaking functional validation in appropriate model systems is a parallel strategy to identify causal genes. For loci where there are obvious gene candidates, investigators can use mice as a model in which to study gene function and determine whether the genes influence atherosclerotic plaque formation or CAD risk factors (assuming that the genes have mouse orthologs) by using methods to reliably overexpress (e.g. viral vectors) and knock down (e.g. antisense/double-stranded RNAs) candidate genes. This somatic approach is preferable to generating transgenic or knockout mice for

each candidate gene (time-intensive). Furthermore, the ability to modulate specific traits in directions predicted to be favorable for MI risk by a somatic approach in mice forecasts a parallel strategy in humans that could be of therapeutic value.

Loci in which there are no clear gene candidates (e.g. the 9p21 locus for MI, for which there is no annotated coding gene within the ~58-kb span) may not be suitable for investigation in mice or other animal models. These loci most likely harbor non-genic regulatory elements, such as transcriptional enhancers, repressors, microRNAs that have long-range effects on distant genes; these elements may function quite differently in mice. It may ultimately prove necessary to study these loci in a human model system, such as human embryonic stem cells, to determine how they influence atherosclerotic disease.

In order to utilize the new genetic information for treatment and prevention of CAD, it will be necessary to understand the functions of the gene(s) near the disease-associated loci and the mechanisms through which they affect coronary risk. As mentioned before, most genes discovered so far do not fit into traditional risk mechanisms.

Modern genetics open up an entirely new sight on the biology of CAD. It appears that its genetically triggered pathogenesis is largely independent of that mediated through traditional risk factors. Nevertheless, it may be that genetic risk variants require a specific environment to come into effect. Indeed, it is likely that genetic factors are embedded in a network that also includes modifiable co-factors. A better knowledge of these interactions will be crucial to gain the greatest benefit from this emerging information on the genetic predisposition to CAD.

Clinicians may expect that addition of genetic information significantly will improve the predictive accuracy of risk scores. Nevertheless, data from future studies need to be awaited to learn which specific groups within our population benefit from determination of genetic risk. Any clinical application of genetic testing will require that

individuals receive not only precise estimations of risk but also profit from specific interventions that lower the overall risk of CAD.

The contribution of genes to the development and progression of CAD, and response to risk factor modification and lifestyle choices are evident. Individuals with a genetic susceptibility for CAD generally have a higher risk to develop disease at an earlier age. The family history is still the best method for initial identification and stratification of genetic risk for CAD, which can be polished through biochemical and DNA testing. Knowledge of genetic susceptibility to CAD has importance in providing risk information and can influence lifestyle choices and management options. Genetically susceptible individuals will benefit the most from treatment of established CAD risk factors. In addition, numerous emerging risk factors are modifiable, and targeting these risk factors with specific therapies may result in improved CAD prevention. Family-based prevention is the most effective for genetically predisposed individuals, since many established and emerging risk factors aggregate in families and most are open to lifestyle changes. Early detection of CAD may be appropriate for genetically susceptible individuals to guide decision-making about risk factor modification. Genetic evaluation, including pedigree analysis, genotyping, genetic counseling, and personalized recommendations for early detection, risk modification, and prevention strategies that are targeted to the genetic risk will result in improved health promotion and CAD prevention efforts. Translational research in CAD genomics will ultimately help to address a public health priority. Since CAD's genetic roots are diffuse, multifactorial, and nondeterministic because many variants scattered across the genome contribute to small risks for CAD. Thus, a polygenic risk score (summarize genetic effects among an ensemble of markers) may be useful. Lately there has been a growing interest in gathering multiple genetic markers into a single score for predicting disease risk. Even if many of the individual markers have no or small detected effect, the combined score could be a robust predictor of disease. This permitted researchers to validate that some diseases have a solid genetic basis, even if few actual genes have been discovered, and it has also revealed a common genetic basis for distinct diseases. Future studies are needed

that investigate the clinical utility of these approaches and the associated ethical, legal, and social issues.

References

1. Kaplan, N. M. Commentary on the sixth report of the Joint National Committee (JNC-6). *Am. J. Hypertens.* **11**, 134-136 (1998).

2. The 1984 Report of the Joint National Committee on Detection, Evaluation, and Treatment of High Blood Pressure. *Arch. Intern. Med.* **144**, 1045-1057 (1984).

3. Go, A. S. *et al.* Heart disease and stroke statistics--2014 update: a report from the American Heart Association. *Circulation* **129**, e28-e292 (2014).

4. Guyton, A. C. The relationship of cardiac output and arterial pressure control. *Circulation* **64**, 1079-1088 (1981).

5. Zanchetti, A. Platt versus Pickering: an episode in recent medical history. By J. D. Swales, editor. An essay review. *Med. Hist.* **30**, 94-96 (1986).

6. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928-933 (2001).

7. Lander, E. S. The new genomics: global views of biology. *Science* **274**, 536-539 (1996).

8. Chakravarti, A. Population genetics--making sense out of sequence. *Nat. Genet.* **21**, 56-60 (1999).

9. Reich, D. E. & Lander, E. S. On the allelic spectrum of human disease. *Trends Genet.* **17**, 502-510 (2001).

10. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* **22**, 231-238 (1999).

11. Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* **322**, 881-888 (2008).

12. 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**, 9362-9367 (2009).

13. Donnelly, P. Progress and challenges in genome-wide association studies in humans. *Nature* **456**, 728-731 (2008).

14. Duggal, P., Gillanders, E. M., Holmes, T. N. & Bailey-Wilson, J. E. Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome wide association studies. *BMC Genomics* **9**, 516-2164-9-516 (2008).

15. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299-1320 (2005).

16. 1000 Genomes Project Consortium *et al*. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073 (2010).

17. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-678 (2007).

18. International Consortium for Blood Pressure Genome-Wide Association Studies *et al*. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* **478**, 103-109 (2011).

19. Levy, D. *et al*. Genome-wide association study of blood pressure and hypertension. *Nat. Genet.* **41**, 677-687 (2009).

20. Newton-Cheh, C. *et al*. Genome-wide association study identifies eight loci associated with blood pressure. *Nat. Genet.* **41**, 666-676 (2009).

21. Newton-Cheh, C. *et al*. Association of common variants in NPPA and NPPB with circulating natriuretic peptides and blood pressure. *Nat. Genet.* **41**, 348-353 (2009).

22. Salvi, E. *et al*. Genomewide association study using a high-density single nucleotide polymorphism array and case-control design identifies a novel essential hypertension susceptibility locus in the promoter region of endothelial NO synthase. *Hypertension* **59**, 248-255 (2012).

23. Wain, L. V. *et al*. Genome-wide association study identifies six new loci influencing pulse pressure and mean arterial pressure. *Nat. Genet*. **43**, 1005-1011 (2011).

24. Padmanabhan, S. *et al*. Genome-wide association study of blood pressure extremes identifies variant near UMOD associated with hypertension. *PLoS Genet*. **6**, e1001177 (2010).

25. Johnson, A. D. *et al*. Association of hypertension drug target genes with blood pressure and hypertension in 86,588 individuals. *Hypertension* **57**, 903-910 (2011).

26. Ho, J. E. *et al*. Discovery and replication of novel blood pressure genetic loci in the Women's Genome Health Study. *J. Hypertens*. **29**, 62-69 (2011).

27. Fox, E. R. *et al*. Association of genetic variation with systolic and diastolic blood pressure among African Americans: the Candidate Gene Association Resource study. *Hum. Mol. Genet*. **20**, 2273-2284 (2011).

28. Zhu, X. *et al*. Combined admixture mapping and association analysis identifies a novel blood pressure genetic locus on 5p13: contributions from the CARe consortium. *Hum. Mol. Genet*. **20**, 2285-2295 (2011).

29. Stoltenberg, S. F. Coming to terms with heritability. *Genetica* **99**, 89-96 (1997).

30. Lynch, M. & Walsh, B. in ***Genetics and Analysis of Quantitative Traits*** 980 (Sinauer Associates, Sunderland, Massachusetts, 1998).

31. Manolio, T. A. *et al*. Finding the missing heritability of complex diseases. *Nature* **461**, 747-753 (2009).

32. Maher, B. Personal genomes: The case of the missing heritability. *Nature* **456**, 18-21 (2008).

33. Schork, N. J., Murray, S. S., Frazer, K. A. & Topol, E. J. Common vs. rare allele hypotheses for complex diseases. *Curr. Opin. Genet. Dev*. **19**, 212-219 (2009).

34. Iyengar, S. K. & Elston, R. C. The genetic basis of complex traits: rare variants or "common gene, common disease"? *Methods Mol. Biol.* **376**, 71-84 (2007).

35. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565-569 (2010).

36. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7-24 (2012).

37. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76-82 (2011).

38. Alexander, R. W. Theodore Cooper Memorial Lecture. Hypertension and the pathogenesis of atherosclerosis. Oxidative stress and the mediation of arterial inflammatory response: a new perspective. *Hypertension* **25**, 155-161 (1995).

39. Stamler, J., Neaton, J. D. & Wentworth, D. N. Blood pressure (systolic and diastolic) and risk of fatal coronary heart disease. *Hypertension* **13**, I2-12 (1989).

40. Kannel, W. B. *et al.* Overall and coronary heart disease mortality rates in relation to major risk factors in 325,348 men screened for the MRFIT. Multiple Risk Factor Intervention Trial. *Am. Heart J.* **112**, 825-836 (1986).

41. Lusis, A. J. Atherosclerosis. *Nature* **407**, 233-241 (2000).

42. Ross, R. The pathogenesis of atherosclerosis: a perspective for the 1990s. *Nature* **362**, 801-809 (1993).

43. Hansson, G. K. Inflammation, atherosclerosis, and coronary artery disease. *N. Engl. J. Med.* **352**, 1685-1695 (2005).

44. Moore, K. J. & Tabas, I. Macrophages in the pathogenesis of atherosclerosis. *Cell* **145**, 341-355 (2011).

45. Virmani, R. *et al.* Atherosclerotic plaque progression and vulnerability to rupture: angiogenesis as a source of intraplaque hemorrhage. *Arterioscler. Thromb. Vasc. Biol.* **25**, 2054-2061 (2005).

46. Libby, P., Ridker, P. M. & Hansson, G. K. Progress and challenges in translating the biology of atherosclerosis. *Nature* **473**, 317-325 (2011).

47. Zaman, A. G., Helft, G., Worthley, S. G. & Badimon, J. J. The role of plaque rupture and thrombosis in coronary artery disease. *Atherosclerosis* **149**, 251-266 (2000).

48. Davies, J. R., Rudd, J. H. & Weissberg, P. L. Molecular and metabolic imaging of atherosclerosis. *J. Nucl. Med.* **45**, 1898-1907 (2004).

49. Burke, A. P. *et al.* Healed plaque ruptures and sudden coronary death: evidence that subclinical rupture has a role in plaque progression. *Circulation* **103**, 934-940 (2001).

50. Strong, J. P. *et al.* Prevalence and extent of atherosclerosis in adolescents and young adults: implications for prevention from the Pathobiological Determinants of Atherosclerosis in Youth Study. *JAMA* **281**, 727-735 (1999).

51. Grundy, S. M. Coronary plaque as a replacement for age as a risk factor in global risk assessment. *Am. J. Cardiol.* **88**, 8E-11E (2001).

52. Schmermund, A., Mohlenkamp, S. & Erbel, R. Coronary artery calcium and its relationship to coronary artery disease. *Cardiol. Clin.* **21**, 521-534 (2003).

53. Wexler, L. *et al.* Coronary artery calcification: pathophysiology, epidemiology, imaging methods, and clinical implications. A statement for health professionals from the American Heart Association. Writing Group. *Circulation* **94**, 1175-1192 (1996).

54. Shaw, L. J., Raggi, P., Schisterman, E., Berman, D. S. & Callister, T. Q. Prognostic value of cardiac risk factors and coronary artery calcium screening for all-cause mortality. *Radiology* **228**, 826-833 (2003).

55. Samani, N. J. *et al.* Genomewide association analysis of coronary artery disease. *N. Engl. J. Med.* **357**, 443-453 (2007).

56. Helgadottir, A. *et al.* A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science* **316**, 1491-1493 (2007).

57. McPherson, R. *et al*. A common allele on chromosome 9 associated with coronary heart disease. *Science* **316**, 1488-1491 (2007).

58. Bown, M. J. *et al*. Association between the coronary artery disease risk locus on chromosome 9p21.3 and abdominal aortic aneurysm. *Circ. Cardiovasc. Genet.* **1**, 39-42 (2008).

59. Helgadottir, A. *et al*. The same sequence variant on 9p21 associates with myocardial infarction, abdominal aortic aneurysm and intracranial aneurysm. *Nat. Genet.* **40**, 217-224 (2008).

60. Horne, B. D., Carlquist, J. F., Muhlestein, J. B., Bair, T. L. & Anderson, J. L. Association of variation in the chromosome 9p21 locus with myocardial infarction versus chronic coronary artery disease. *Circ. Cardiovasc. Genet.* **1**, 85-92 (2008).

61. Pasmant, E. *et al*. Characterization of a germ-line deletion, including the entire INK4/ARF locus, in a melanoma-neural system tumor family: identification of ANRIL, an antisense noncoding RNA whose expression coclusters with ARF. *Cancer Res.* **67**, 3963-3969 (2007).

62. CARDIoGRAMplusC4D Consortium *et al*. Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat. Genet.* **45**, 25-33 (2013).

63. Schunkert, H. *et al*. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat. Genet.* **43**, 333-338 (2011).

64. Coronary Artery Disease (C4D) Genetics Consortium. A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. *Nat. Genet.* **43**, 339-344 (2011).

65. Clarke, R. *et al*. Genetic variants associated with Lp(a) lipoprotein level and coronary disease. *N. Engl. J. Med.* **361**, 2518-2528 (2009).

66. Erdmann, J. *et al*. New susceptibility locus for coronary artery disease on chromosome 3q22.3. *Nat. Genet.* **41**, 280-282 (2009).

67. Myocardial Infarction Genetics Consortium *et al*. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat. Genet.* **41**, 334-341 (2009).

68. Soranzo, N. *et al*. A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nat. Genet*. **41**, 1182-1190 (2009).

69. Wang, F. *et al*. Genome-wide association identifies a susceptibility locus for coronary artery disease in the Chinese Han population. *Nat. Genet*. **43**, 345-349 (2011).

70. IBC 50K CAD Consortium. Large-scale gene-centric analysis identifies novel variants for coronary artery disease. *PLoS Genet*. **7**, e1002260 (2011).

71. Janssens, A. C. Is the time right for translation research in genomics? *Eur. J. Epidemiol*. **23**, 707-710 (2008).

72. Khoury, M. J. *et al*. The Scientific Foundation for personal genomics: recommendations from a National Institutes of Health-Centers for Disease Control and Prevention multidisciplinary workshop. *Genet. Med*. **11**, 559-567 (2009).

73. 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**, 9362-9367 (2009).

74. Wray, N. R., Goddard, M. E. & Visscher, P. M. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res*. **17**, 1520-1528 (2007).

75. Yu, W., Gwinn, M., Clyne, M., Yesupriya, A. & Khoury, M. J. A navigator for human genome epidemiology. *Nat. Genet*. **40**, 124-125 (2008).

76. Welter, D. *et al*. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res*. **42**, D1001-6 (2014).

77. Khoury, M. J. *et al*. The genomic applications in practice and prevention network. *Genet. Med*. **11**, 488-494 (2009).

78. Wilson, P. W. *et al*. Prediction of coronary heart disease using risk factor categories. *Circulation* **97**, 1837-1847 (1998).

79. Hackam, D. G. & Anand, S. S. Emerging risk factors for atherosclerotic vascular disease: a critical review of the evidence. *JAMA* **290**, 932-940 (2003).

80. Morrison, A. C. *et al*. Prediction of coronary heart disease risk using a genetic risk score: the Atherosclerosis Risk in Communities Study. *Am. J. Epidemiol*. **166**, 28-35 (2007).

81. Plomin, R., Haworth, C. M. & Davis, O. S. Common disorders are quantitative traits. *Nat. Rev. Genet*. **10**, 872-878 (2009).

82. Fisher, R. The Correlation Between Relatives on the Supposition of Mendelian Inheritance. **Transactions of the Royal Society of Edinburgh, 52**, 399-433 (1918).

83. World Health Organization. World Health Statistics 2012. *WHO* **2012**, 35 (2012).

84. Lifton, R. P. & Jeunemaitre, X. Finding genes that cause human hypertension. *J. Hypertens*. **11**, 231-236 (1993).

85. Ji, W. *et al*. Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat. Genet*. **40**, 592-599 (2008).

86. Padmanabhan, S. *et al*. Genome-wide association study of blood pressure extremes identifies variant near UMOD associated with hypertension. *PLoS Genet*. **6**, e1001177 (2010).

87. Ehret, G. B. & Caulfield, M. J. Genes for blood pressure: an opportunity to understand hypertension. *Eur. Heart J*. **34**, 951-961 (2013).

88. Manolio, T. A. *et al*. Finding the missing heritability of complex diseases. *Nature* **461**, 747-753 (2009).

89. Maher, B. Personal genomes: The case of the missing heritability. *Nature* **456**, 18-21 (2008).

90. Yang, J. *et al*. Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet*. **43**, 519-525 (2011).

91. Lee, S. H. *et al*. Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat. Genet*. **44**, 247-250 (2012).

92. Lee, S. H. *et al*. Estimation and partitioning of polygenic variation captured by common SNPs for Alzheimer's disease, multiple sclerosis and endometriosis. *Hum. Mol. Genet*. **22**, 832-841 (2013).

93. Vinkhuyzen, A. A. *et al*. Common SNPs explain some of the variation in the personality dimensions of neuroticism and extraversion. *Transl. Psychiatry*. **2**, e102 (2012).

94. Ganesh, S. K. *et al*. Effects of long-term averaging of quantitative blood pressure traits on the detection of genetic associations. *Am. J. Hum. Genet*. **95**, 49-65 (2014).

95. Cook, N. R., Gillman, M. W., Rosner, B. A., Taylor, J. O. & Hennekens, C. H. Combining annual blood pressure measurements in childhood to improve prediction of young adult blood pressure. *Stat. Med*. **19**, 2625-2640 (2000).

96. Harvey, P. R., Holt, A., Nicholas, J. & Dasgupta, I. Is an average of routine postdialysis blood pressure a good indicator of blood pressure control and cardiovascular risk? *J. Nephrol*. **26**, 94-100 (2013).

97. Halushka, M. K. *et al*. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat. Genet*. **22**, 239-247 (1999).

98. Voight, B. F. *et al*. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet*. **8**, e1002793 (2012).

99. Levy, D. *et al*. Evidence for a gene influencing blood pressure on chromosome 17. Genome scan linkage results for longitudinal blood pressure phenotypes in subjects from the framingham heart study. *Hypertension* **36**, 477-483 (2000).

100. Zhang, S. *et al*. Genetic and environmental contributions to phenotypic components of metabolic syndrome: a population-based twin study. *Obesity (Silver Spring)* **17**, 1581-1587 (2009).

101. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators. *Am. J. Epidemiol*. **129**, 687-702 (1989).

102. ARIC Investigators. Sitting blood pressure and postural changes in blood pressure and heart rate in Atherosclerosis Risk in Communities Study Protocol. *Chapel Hill, NC: ARIC Coordinating Center, Department of Biostatistics, University of North Carolina* **Manual 11** (1988).

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

104. Dimmer, E. C. *et al.* The UniProt-GO Annotation database in 2011. *Nucleic Acids Res.* **40**, D565-70 (2012).

107. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001-6 (2014).

108. Johnson, A. D. *et al.* SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* **24**, 2938-2939 (2008).

109. De Jager, P. L. *et al.* Integration of genetic risk factors into a clinical algorithm for multiple sclerosis susceptibility: a weighted genetic risk score. *Lancet Neurol.* **8**, 1111-1119 (2009).

110. Gottesman, O. *et al.* The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet. Med.* **15**, 761-771 (2013).

111. Kho, A. N. *et al.* Practical challenges in integrating genomic data into the electronic health record. *Genet. Med.* **15**, 772-778 (2013).

112. Marsolo, K. & Spooner, S. A. Clinical genomics in the world of the electronic health record. *Genet. Med.* **15**, 786-791 (2013).

113. Ury, A. G. Storing and interpreting genomic information in widely deployed electronic health record systems. *Genet. Med.* **15**, 779-785 (2013).

114. International Schizophrenia Consortium *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748-752 (2009).

115. Thanassoulis, G. *et al.* A genetic risk score is associated with incident cardiovascular disease and coronary artery calcium: the Framingham Heart Study. *Circ. Cardiovasc. Genet.* **5**, 113-121 (2012).

116. Cook, N. R. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* **115**, 928-935 (2007).

117. Steyerberg, E. W. *et al*. Assessing the incremental value of diagnostic and prognostic markers: a review and illustration. *Eur. J. Clin. Invest.* **42**, 216-228 (2012).

118. Schrodi, S. J. *et al*. Genetic-based prediction of disease traits: prediction is very difficult, especially about the future. *Front. Genet.* **5**, 162 (2014).

119. CARDIoGRAMplusC4D Consortium *et al*. Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat. Genet.* **45**, 25-33 (2013).

120. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators. *Am. J. Epidemiol.* **129**, 687-702 (1989).

121. White, A. D. *et al*. Community surveillance of coronary heart disease in the Atherosclerosis Risk in Communities (ARIC) Study: methods and initial two years' experience. *J. Clin. Epidemiol.* **49**, 223-233 (1996).

122. Volcik, K. A. *et al*. P-selectin Thr715Pro polymorphism predicts P-selectin levels but not risk of incident coronary heart disease or ischemic stroke in a cohort of 14595 participants: the Atherosclerosis Risk in Communities Study. *Atherosclerosis* **186**, 74-79 (2006).

123. Wilson, P. W. *et al*. Prediction of coronary heart disease using risk factor categories. *Circulation* **97**, 1837-1847 (1998).

124. 1000 Genomes Project Consortium *et al*. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).

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

126. Zdravkovic, S. *et al*. Heritability of death from coronary heart disease: a 36-year follow-up of 20 966 Swedish twins. *J. Intern. Med.* **252**, 247-254 (2002).

127. Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.* **91**, 1011-1021 (2012).

128. Kramer, A. A. & Zimmerman, J. E. Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited. *Crit. Care Med.* **35**, 2052-2056 (2007).

129. Crowson, C. S., Atkinson, E. J. & Therneau, T. M. Assessing calibration of prognostic risk scores. *Stat. Methods Med. Res.* (2014).

130. Paynter, N. P. & Cook, N. R. A bias-corrected net reclassification improvement for clinical subgroups. *Med. Decis. Making* **33**, 154-162 (2013).

131. Ridker, P. M. & Cook, N. R. Statins: new American guidelines for prevention of cardiovascular disease. *Lancet* **382**, 1762-1765 (2013).

132. Goff, D. C.,Jr *et al.* 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation* **129**, S49-73 (2014).

133. Kerr, K. F. *et al.* Net reclassification indices for evaluating risk prediction instruments: a critical review. *Epidemiology* **25**, 114-121 (2014).

134. Muntner, P., Safford, M. M., Cushman, M. & Howard, G. Comment on the reports of over-estimation of ASCVD risk using the 2013 AHA/ACC risk equation. *Circulation* **129**, 266-267 (2014).

135. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904-909 (2006).

136. R Core Team. R: A Language and Environment for Statistical Computing
. **R 3.1.1 "Sock it to Me**, http://www.R-project.org-R Foundation for Statistical Computing (2014).

137. Stone, N. J. *et al.* 2013 ACC/AHA guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *J. Am. Coll. Cardiol.* **63**, 2889-2934 (2014).

138. Brautbar, A. *et al.* A genetic risk score based on direct associations with coronary heart disease improves coronary heart disease risk prediction in the Atherosclerosis Risk in

Communities (ARIC), but not in the Rotterdam and Framingham Offspring, Studies. *Atherosclerosis* **223**, 421-426 (2012).

139. Hughes, M. F. *et al*. Genetic markers enhance coronary risk prediction in men: the MORGAM prospective cohorts. *PLoS One* **7**, e40922 (2012).

140. Ganna, A. *et al*. Multilocus genetic risk scores for coronary heart disease prediction. *Arterioscler. Thromb. Vasc. Biol.* **33**, 2267-2272 (2013).

141. Thanassoulis, G., Peloso, G. M. & O'Donnell, C. J. Genomic medicine for improved prediction and primordial prevention of cardiovascular disease. *Arterioscler. Thromb. Vasc. Biol.* **33**, 2049-2050 (2013).

142. Tikkanen, E., Havulinna, A. S., Palotie, A., Salomaa, V. & Ripatti, S. Genetic risk prediction and a 2-stage risk screening strategy for coronary heart disease. *Arterioscler. Thromb. Vasc. Biol.* **33**, 2261-2266 (2013).

143. Ganna, A. *et al*. Multilocus genetic risk scores for coronary heart disease prediction. *Arterioscler. Thromb. Vasc. Biol.* **33**, 2267-2272 (2013).

144. Ioannidis, J. P. & Tzoulaki, I. What makes a good predictor?: the evidence applied to coronary artery calcium score. *JAMA* **303**, 1646-1647 (2010).

145. Knowles, J. W. *et al*. Randomized trial of personal genomics for preventive cardiology: design and challenges. *Circ. Cardiovasc. Genet.* **5**, 368-376 (2012).

146. Grant, R. W. *et al*. Personalized genetic risk counseling to motivate diabetes prevention: a randomized trial. *Diabetes Care* **36**, 13-19 (2013).

147. Simonson, M. A., Wills, A. G., Keller, M. C. & McQueen, M. B. Recent methods for polygenic analysis of genome-wide data implicate an important effect of common variants on cardiovascular disease risk. *BMC Med. Genet.* **12**, 146-2350-12-146 (2011).

148. Ollier, W., Sprosen, T. & Peakman, T. UK Biobank: from concept to reality. *Pharmacogenomics* **6**, 639-646 (2005).

149. Palmer, L. J. UK Biobank: bank on it. *Lancet* **369**, 1980-1982 (2007).

150. Hoffmann, T. J. *et al*. Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array. *Genomics* **98**, 79-89 (2011).

151. Hoffmann, T. J. *et al*. Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm. *Genomics* **98**, 422-430 (2011).

152. Kaufman, D., Bollinger, J., Dvoskin, R. & Scott, J. Preferences for opt-in and opt-out enrollment and consent models in biobank research: a national survey of Veterans Administration patients. *Genet. Med.* **14**, 787-794 (2012).

153. Ntzani, E. E., Liberopoulos, G., Manolio, T. A. & Ioannidis, J. P. Consistency of genome-wide associations across major ancestral groups. *Hum. Genet.* **131**, 1057-1071 (2012).

154. National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. *Circulation* **106**, 3143-3421 (2002).

155. Ioannidis, J. P. More than a billion people taking statins?: Potential implications of the new cardiovascular guidelines. *JAMA* **311**, 463-464 (2014).

156. Go, A. S. *et al*. Heart disease and stroke statistics--2014 update: a report from the American Heart Association. *Circulation* **129**, e28-e292 (2014).

157. Stamler, J., Wentworth, D. & Neaton, J. D. Is relationship between serum cholesterol and risk of premature death from coronary heart disease continuous and graded? Findings in 356,222 primary screenees of the Multiple Risk Factor Intervention Trial (MRFIT). *JAMA* **256**, 2823-2828 (1986).

158. Kannel, W. B. *et al*. Overall and coronary heart disease mortality rates in relation to major risk factors in 325,348 men screened for the MRFIT. Multiple Risk Factor Intervention Trial. *Am. Heart J.* **112**, 825-836 (1986).

159. Greenland, P. *et al*. ACCF/AHA 2007 clinical expert consensus document on coronary artery calcium scoring by computed tomography in global cardiovascular risk assessment and in evaluation of patients with chest pain: a report of the American College of Cardiology Foundation Clinical Expert Consensus Task Force (ACCF/AHA Writing Committee to Update the 2000 Expert Consensus Document on Electron Beam Computed Tomography) developed in collaboration with the Society of Atherosclerosis Imaging and Prevention and the Society of Cardiovascular Computed Tomography. *J. Am. Coll. Cardiol.* **49**, 378-402 (2007).

160. Arad, Y. *et al*. Predictive value of electron beam computed tomography of the coronary arteries. 19-month follow-up of 1173 asymptomatic subjects. *Circulation* **93**, 1951-1953 (1996).

161. Detrano, R. C. *et al*. Coronary calcium does not accurately predict near-term future coronary events in high-risk adults. *Circulation* **99**, 2633-2638 (1999).

162. Secci, A. *et al*. Electron beam computed tomographic coronary calcium as a predictor of coronary events: comparison of two protocols. *Circulation* **96**, 1122-1129 (1997).

163. Yeboah, J. *et al*. Comparison of novel risk markers for improvement in cardiovascular risk assessment in intermediate-risk individuals. *JAMA* **308**, 788-795 (2012).

164. Budoff, M. J. Atherosclerosis imaging and calcified plaque: coronary artery disease risk assessment. *Prog. Cardiovasc. Dis.* **46**, 135-148 (2003).

165. Rumberger, J. A., Simons, D. B., Fitzpatrick, L. A., Sheedy, P. F. & Schwartz, R. S. Coronary artery calcium area by electron-beam computed tomography and coronary atherosclerotic plaque area. A histopathologic correlative study. *Circulation* **92**, 2157-2162 (1995).

166. Sangiorgi, G. *et al*. Arterial calcification and not lumen stenosis is highly correlated with atherosclerotic plaque burden in humans: a histologic study of 723 coronary artery segments using nondecalcifying methodology. *J. Am. Coll. Cardiol.* **31**, 126-133 (1998).

167. Samani, N. J. *et al*. Genomewide association analysis of coronary artery disease. *N. Engl. J. Med.* **357**, 443-453 (2007).

168. Myocardial Infarction Genetics Consortium *et al*. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat. Genet*. **41**, 334-341 (2009).

169. Wang, F. *et al*. Genome-wide association identifies a susceptibility locus for coronary artery disease in the Chinese Han population. *Nat. Genet*. **43**, 345-349 (2011).

170. Goldstein, B. A., Knowles, J. W., Salfati, E., Ioannidis, J. P. & Assimes, T. L. Simple, standardized incorporation of genetic risk into non-genetic risk prediction tools for complex traits: coronary heart disease as an example. *Front. Genet*. **5**, 254 (2014).

171. Ripatti, S. *et al*. A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. *Lancet* **376**, 1393-1400 (2010).

172. Vaarhorst, A. A. *et al*. Literature-based genetic risk scores for coronary heart disease: the Cardiovascular Registry Maastricht (CAREMA) prospective cohort study. *Circ. Cardiovasc. Genet*. **5**, 202-209 (2012).

173. Go, A. S. *et al*. Statin and beta-blocker therapy and the initial presentation of coronary heart disease. *Ann. Intern. Med*. **144**, 229-238 (2006).

174. Taylor-Piliae, R. E. *et al*. Validation of a new brief physical activity survey among men and women aged 60-69 years. *Am. J. Epidemiol*. **164**, 598-606 (2006).

175. Iribarren, C. *et al*. Metabolic syndrome and early-onset coronary artery disease: is the whole greater than its parts? *J. Am. Coll. Cardiol*. **48**, 1800-1807 (2006).

176. Hughes, G. H. *et al*. Recruitment in the Coronary Artery Disease Risk Development in Young Adults (Cardia) Study. *Control. Clin. Trials* **8**, 68S-73S (1987).

177. Strong, J. P. *et al*. Prevalence and extent of atherosclerosis in adolescents and young adults: implications for prevention from the Pathobiological Determinants of Atherosclerosis in Youth Study. *JAMA* **281**, 727-735 (1999).

178. Bild, D. E. *et al*. Multi-ethnic study of atherosclerosis: objectives and design. *Am. J. Epidemiol*. **156**, 871-881 (2002).

179. DAWBER, T. R., MEADORS, G. F. & MOORE, F. E.,Jr. Epidemiological approaches to heart disease: the Framingham Study. *Am. J. Public Health Nations Health* **41**, 279-281 (1951).

180. Tell, G. S. *et al*. Recruitment of adults 65 years and older as participants in the Cardiovascular Health Study. *Ann. Epidemiol*. **3**, 358-366 (1993).

181. Fried, L. P. *et al*. The Cardiovascular Health Study: design and rationale. *Ann. Epidemiol*. **1**, 263-276 (1991).

182. Assimes, T. L. *et al*. Susceptibility locus for clinical and subclinical coronary artery disease at chromosome 9p21 in the multi-ethnic ADVANCE study. *Hum. Mol. Genet*. **17**, 2320-2328 (2008).

183. Fair, J. M. *et al*. Ethnic differences in coronary artery calcium in a healthy cohort aged 60 to 69 years. *Am. J. Cardiol*. **100**, 981-985 (2007).

184. Kronmal, R. A. *et al*. Risk factors for the progression of coronary artery calcification in asymptomatic subjects: results from the Multi-Ethnic Study of Atherosclerosis (MESA). *Circulation* **115**, 2722-2730 (2007).

185. Kannel, W. B., Feinleib, M., McNamara, P. M., Garrison, R. J. & Castelli, W. P. An investigation of coronary heart disease in families. The Framingham offspring study. *Am. J. Epidemiol*. **110**, 281-290 (1979).

186. Newman, A. B. *et al*. Coronary artery calcification in older adults with minimal clinical or subclinical cardiovascular disease. *J. Am. Geriatr. Soc*. **48**, 256-263 (2000).

187. Chuang, M. L. *et al*. Prevalence and distribution of abdominal aortic calcium by gender and age group in a community-based cohort (from the Framingham Heart Study). *Am. J. Cardiol*. **110**, 891-896 (2012).

188. Hoffmann, U., Massaro, J. M., Fox, C. S., Manders, E. & O'Donnell, C. J. Defining normal distributions of coronary artery calcium in women and men (from the Framingham Heart Study). *Am. J. Cardiol*. **102**, 1136-41, 1141.e1 (2008).

189. Raggi, P. *et al*. Identification of patients at increased risk of first unheralded acute myocardial infarction by electron-beam computed tomography. *Circulation* **101**, 850-855 (2000).

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

191. Purcell, S. *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559-575 (2007).

192. Viechtbauer, W. Conducting meta-analyses in R with the metafor package. Journal of Statistical Software, URL http://www.jstatsoft.org/v36/i03/. **36** (2010).

193. Arad, Y., Goodman, K. J., Roth, M., Newstein, D. & Guerci, A. D. Coronary calcification, coronary disease risk factors, C-reactive protein, and atherosclerotic cardiovascular disease events: the St. Francis Heart Study. *J. Am. Coll. Cardiol.* **46**, 158-165 (2005).

194. Budoff, M. J. *et al*. Long-term prognosis associated with coronary calcification: observations from a registry of 25,253 patients. *J. Am. Coll. Cardiol.* **49**, 1860-1870 (2007).

195. Elias-Smale, S. E. *et al*. Coronary calcium score improves classification of coronary heart disease risk in the elderly: the Rotterdam study. *J. Am. Coll. Cardiol.* **56**, 1407-1414 (2010).

196. Erbel, R. *et al*. Coronary risk stratification, discrimination, and reclassification improvement based on quantification of subclinical coronary atherosclerosis: the Heinz Nixdorf Recall study. *J. Am. Coll. Cardiol.* **56**, 1397-1406 (2010).

197. Hopkins, P. N. *et al*. Association of coronary artery calcified plaque with clinical coronary heart disease in the National Heart, Lung, and Blood Institute's Family Heart Study. *Am. J. Cardiol.* **97**, 1564-1569 (2006).

198. Nasir, K., Michos, E. D., Blumenthal, R. S. & Raggi, P. Detection of high-risk young adults and women by coronary calcium and National Cholesterol Education Program Panel III guidelines. *J. Am. Coll. Cardiol.* **46**, 1931-1936 (2005).

199. Polonsky, T. S. *et al*. Coronary artery calcium score and risk classification for coronary heart disease prediction. *JAMA* **303**, 1610-1616 (2010).

200. Shaw, L. J., Raggi, P., Schisterman, E., Berman, D. S. & Callister, T. Q. Prognostic value of cardiac risk factors and coronary artery calcium screening for all-cause mortality. *Radiology* **228**, 826-833 (2003).

201. Taylor, A. J. *et al*. Coronary calcium independently predicts incident premature coronary heart disease over measured cardiovascular risk factors: mean three-year outcomes in the Prospective Army Coronary Calcium (PACC) project. *J. Am. Coll. Cardiol*. **46**, 807-814 (2005).

202. Greenland, P., LaBree, L., Azen, S. P., Doherty, T. M. & Detrano, R. C. Coronary artery calcium score combined with Framingham score for risk prediction in asymptomatic individuals. *JAMA* **291**, 210-215 (2004).

203. Simionescu, M. & Sima, A. V. in ***Inflammation and Atherosclerosis*** (ed Georg Wick, C. G.) 19 (Springer-Verlag, Wien, 2012).

204. Romm, P. A., Green, C. E., Reagan, K. & Rackley, C. E. Relation of serum lipoprotein cholesterol levels to presence and severity of angiographic coronary artery disease. *Am. J. Cardiol*. **67**, 479-483 (1991).

205. Report of the National Cholesterol Education Program Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. The Expert Panel. *Arch. Intern. Med*. **148**, 36-69 (1988).

206. Steinberg, D. & Witztum, J. L. Oxidized low-density lipoprotein and atherosclerosis. *Arterioscler. Thromb. Vasc. Biol*. **30**, 2311-2316 (2010).

207. Grundtman, C. in *Inflammation and Atherosclerosis* (ed Georg Wick, C. G.) 133 (Springer-Verlag, Wien, 2012).

208. Evans, D. M., Visscher, P. M. & Wray, N. R. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Hum. Mol. Genet*. **18**, 3525-3531 (2009).

209. Wacholder, S., McLaughlin, J. K., Silverman, D. T. & Mandel, J. S. Selection of controls in case-control studies. I. Principles. *Am. J. Epidemiol.* **135**, 1019-1028 (1992).

210. Reilly, M. P. *et al.* Identification of ADAMTS7 as a novel locus for coronary atherosclerosis and association of ABO with myocardial infarction in the presence of coronary atherosclerosis: two genome-wide association studies. *Lancet* **377**, 383-392 (2011).

211. Burke, A. P. *et al.* Healed plaque ruptures and sudden coronary death: evidence that subclinical rupture has a role in plaque progression. *Circulation* **103**, 934-940 (2001).

212. Hong, M. K. *et al.* Comparison of coronary plaque rupture between stable angina and acute myocardial infarction: a three-vessel intravascular ultrasound study in 235 patients. *Circulation* **110**, 928-933 (2004).

213. Hong, M. K. *et al.* Plaque ruptures in stable angina pectoris compared with acute coronary syndrome. *Int. J. Cardiol.* **114**, 78-82 (2007).

214. Brown, A. A. *et al.* Genetic interactions affecting human gene expression identified by variance association mapping. *Elife* **3**, e01381 (2014).

215. van der Net, J. B., Janssens, A. C., Sijbrands, E. J. & Steyerberg, E. W. Value of genetic profiling for the prediction of coronary heart disease. *Am. Heart J.* **158**, 105-110 (2009).

216. Yang, Q., Khoury, M. J., Friedman, J., Little, J. & Flanders, W. D. How many genes underlie the occurrence of common complex diseases in the population? *Int. J. Epidemiol.* **34**, 1129-1137 (2005).

217. Robinson, M. R., Wray, N. R. & Visscher, P. M. Explaining additional genetic variation in complex traits. *Trends Genet.* **30**, 124-132 (2014).

218. Service, S. K. *et al.* Re-sequencing expands our understanding of the phenotypic impact of variants at GWAS loci. *PLoS Genet.* **10**, e1004147 (2014).

219. Makinen, V. P. *et al.* Integrative genomics reveals novel molecular pathways and gene networks for coronary artery disease. *PLoS Genet.* **10**, e1004502 (2014).

220. Grossman, E. Ambulatory blood pressure monitoring in the diagnosis and management of hypertension. *Diabetes Care* **36 Suppl 2**, S307-11 (2013).

221. Batkai, S. & Thum, T. MicroRNAs in hypertension: mechanisms and therapeutic targets. *Curr. Hypertens. Rep.* **14**, 79-87 (2012).

222. Kontaraki, J. E., Marketou, M. E., Zacharis, E. A., Parthenakis, F. I. & Vardas, P. E. Differential expression of vascular smooth muscle-modulating microRNAs in human peripheral blood mononuclear cells: novel targets in essential hypertension. *J. Hum. Hypertens.* **28**, 510-516 (2014).

223. Ward, L. D. & Kellis, M. Interpreting noncoding genetic variation in complex traits and human disease. *Nat. Biotechnol.* **30**, 1095-1106 (2012).

224. Wain, L. V., Armour, J. A. & Tobin, M. D. Genomic copy number variation, human health, and disease. *Lancet* **374**, 340-350 (2009).

225. Rabbani, B., Tekin, M. & Mahdieh, N. The promise of whole-exome sequencing in medical genetics. *J. Hum. Genet.* **59**, 5-15 (2014).