EXPLORING THE ROLE OF BIOMARKER GENETICS IN CARDIOVASCULAR DISEASE

EXPLORING THE ROLE OF BIOMARKER GENETICS IN CARDIOVASCULAR DISEASE

By JENNIFER SJAARDA, B.Sc. (Hons.)

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

DOCTOR OF PHILOSOPHY (2018) McMaster University
(Department of Medical Sciences) Hamilton, Ontario

TITLE: Exploring the role of biomarker genetics in cardiovascular disease

AUTHOR: Jennifer Sjaarda
B.Sc. (Honours) Biomedical Science
(University of Waterloo)

SUPERVISOR: Dr. Guillaume Paré, MD, M.Sc., FRCPC

NUMBER OF PAGES: iii - 213

ABSTRACT

Biomarkers provide the opportunity to identify subclinical disease states before development of the disease and apply preventative measures, facilitate research and understanding of disease mechanisms, and allow for the assessment of therapeutic measures. However, a major challenge in the field of biomarker research is to discern cause and effect, and as a result, association has often been mistaken for causation. Additionally, the pathogenesis behind chronic diseases is extremely complex, resulting from several modifiable and unmodifiable risk factors and often caused by an interaction of many biomarkers simultaneously. Furthermore, biomarker levels show marked differences across ethnicities and it is difficult to distinguish whether this is a result of environmental or genetic factors. Through longitudinal, genetic, multi-biomarker studies, these barriers can be partially overcome. This thesis addresses how advancements in genetic and biomarker research may help to gain novel insights into both known and novel biomarkers of cardiovascular disease, inform and guide clinical decision-making and validate potential disease target pathways.

Using a variety of statistical approaches, we analyzed 4,147 participants of the Outcome Reduction with Initial Glargine Intervention (ORIGIN) trial measured for 237 biomarkers and followed for an average of 6.2 years, to investigate the genetic effects on biomarkers and their relation to cardiovascular diseases. Specifically, we applied Mendelian randomization (MR) analyses to identify novel, causal mediators of coronary artery disease (CAD) and chronic kidney disease (CKD). Additionally, we used admixture mapping to explore the impact of ancestry on serum biomarker levels in the Native Latin ORIGIN population and identified genes conferring differential risk across ancestries. We identified macrophage colony-stimulating factor 1 (CSF1) and stromal cell-derived factor (CXCL12) as novel, causal mediators of CAD using MR. Similarly, MR analysis also revealed uromodulin (UMOD) and human EGF receptor 2 (HER2) as new mediators of CKD.

Through admixture mapping, we have demonstrated the importance of ethnicity across a comprehensive panel of biomarkers and shown a novel method for inferring the contribution of ethnicity to phenotypic traits in admixed individuals.

By applying two major statistical methods employed in genetic epidemiology we have revealed important insights into the role of biomarkers in health and disease. Taken together, this thesis implicates new biomarkers for CAD and CKD which are potential therapeutic interventions for prevention and treatment. Furthermore, this work indicates the importance of ancestry in disease, and paves the way for clinical treatment which is tailored to ethnicity. Future studies may adopt the novel approaches presented here to identify additional causal markers of disease and biological pathways and processes which are influenced by ancestry.

ACKNOWLEDGEMENTS

First, and foremost I would like to express my sincere gratitude towards my supervisor Dr. Guillaume Paré, without whom none of this work would be possible. Dr. Paré provided me with tremendous support, encouragement and opportunity throughout my PhD. His continued guidance and mentorship is deeply appreciated. His confidence in my ability have in turn built my confidence and taught me to be a critical researcher. I consider it a great privilege and honor to have had the opportunity to be under his tutelage for the past five years.

Second of all, I would like to express my deepest thanks to my PhD committee members, Dr. Hertzel Gerstein and Dr. David Meyre. Their valuable input and feedback throughout my PhD are very much appreciated. Thank-you both for making committee meetings enjoyable rather than nerve-wracking! Your enthusiasm and love of science is contagious, and I'm confident my PhD experience would not have been as positive without either of your support and insights.

I would also like to thank the past and present member of the Genetic Molecular Epidemiology Lab for their collaboration and friendship. Finally, I would like to thank all of my friends and family for their unwavering support. A special thanks to my husband, Matthew Sjaarda, for his endless encouragement during my PhD.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 GENERAL INTRODUCTION

1.1 Background

1.1.1 Burden of Cardiovascular Disease

Cardiovascular disease (CVD) represents a class of disorders that involve the heart and blood vessels, and includes a wide array of diseases such as atherosclerosis, hypertrophy, heart failure, myocardial infarction (MI), stroke and coronary artery disease (CAD)[1]. CVD is caused by a complex interaction of genetic, environmental and behavioral risk factors that is not fully understood. The underlying pathological mechanisms vary depending on the disease in question, however atherosclerosis is a common feature among many cases[2]. Much of CVD is thought to be preventable through risk management of modifiable factors such as smoking, physical inactivity, obesity, high blood pressure and diabetes[3–5]. Despite this, CVD remains the leading cause of chronic disease morbidity and mortality in developed countries and the prevalence is steadily increasing around the globe[1]. Canadians are at a particularly high risk of CVD, with over 80% carrying at least one major risk factor. As a consequence, CVD represents the underlying cause for 1 in 3 deaths in Canada[6]. Furthermore, the increasing prevalence of modifiable risk factors is expected to reduce life expectancy in the coming years[7]. Diabetes is emerging as one of the most important of these risk factors for CVD[8–11]. Additionally, CVD is the most common cause of mortality in those with diabetes, accounting for over 60% of all deaths[12]. While epidemiological studies have shown that this relationship is independent of common risk factors such as hypertension, obesity and chronic kidney disease (CKD)[13], the pathophysiology underlying this observation remains unclear. Therefore there is a need to further characterize this relationship in an effort to reduce the overall burden of CVD.

1.1.2 Cardiovascular Biomarkers

With the prevalence of CVD on the rise, prevention and risk stratification are major public health priorities[14]. Accurate and early identification of individuals at an increased risk for CVD allows clinicians to apply preventative measures before the progression of the disease[15]. Substantial data suggests that CVD is a life-long disease, occurring through an evolution of risk factors and resulting in subclinical disease states that often go undetected in routine appointments[16,17]. Biomarker measurement and evaluation is one tool to combat this ongoing issue and offers tremendous potential for early diagnosis, better treatment and improved management of patients at a relatively low-cost and in a non-invasive manner[18]. A biomarker is defined as a characteristic that is objectively measured in a patient and used as an indicator of normal biological processes or pharmacological responses to interventions[19]. Identification of such biological mediators has proven to be an extremely valuable research endeavor. For instance, the discovery of risk factors of CVD, such as hypertension, hypercholesterolemia and BMI, have not only led to an increased understanding of the disease biology, but also many clinical advances[20–22]. Many biomarkers, such as C-reactive protein (CRP), interleukin-6, and troponin, have been studied in the context of CVD[23–25]. However, despite much research we have a relatively poor understanding of the exact role these biomarkers play in the disease. Results among studies are often inconsistent and lack conclusive findings likely due to the fact that CVD is a complex disorder resulting from an interaction of genetic and environmental factors and many biomarkers in parallel[26]. Therefore, longitudinal studies in large sample sizes are often needed to elucidate conflicting findings.

1.2 Genetics for Biomarker Discovery and Understanding

1.2.1 Mendelian Randomization

Overview of MR in cardiovascular disease

Understanding disease aetiology and identifying novel opportunities for treatment and prevention remains the sole objective and purpose for scientific and clinical research. Motivated by these central goals, epidemiological studies have investigated numerous traits and biomarkers for associations with a vast array of cardiometabolic outcomes[27–29]. Although numerous markers have been robustly identified, even the strongest epidemiological associations preclude drawing conclusions about the causality or effect direction in the underlying relationship due to inherent limitations and biases, including residual confounding and reverse causation[30]. As a result, many of the associations found in epidemiological analyses are mere markers of disease risk, rather than causal factors directly involved in disease progression, and are therefore poor choices for therapeutic targets. Indeed, many putative causal biomarkers have been subsequently tested in randomized control trials (RCTs) where blockade through pharmaceutical intervention conferred no benefit, in contrast to the findings predicted by epidemiological models[31]. Such discordant findings between RCTs and observational studies are commonplace and suggest a non-causal link between the modifiable factor and outcome under study. While well designed and conducted RCTs remain the gold-standard for causal inference, they are exceedingly expensive, time-consuming, may not be feasible or ethical, and have high failure rates[32,33]. Therefore, given the complexity of cardiometabolic disorders, the number of biomarkers at large, and the cost of drug development programmes, it is essential to employ cost-effective methods which provide preliminary evidence on promising therapeutic targets[34].

In genetic epidemiology, Mendelian randomization (MR) studies are powerful and useful tools which are able to provide information on the causality of known and novel relationships in the absence of trial data, shedding light on potential biomarkers for future drug development[35–39]. MR studies harness the fundamental principles of genetic inheritance to infer whether a biomarker is causally related to a disease. This is made possible since genetic variants are inherited independently

of each other, randomly determined, and thus unrelated to other confounding factors. In this regard, MR studies act as "natural" RCTs which make use of the unmodifiable, random allocation of genetic variants at meiosis, and are therefore able to mitigate traditional sources of bias due to confounding and reverse causation. Specifically, in MR studies, genetic variants are used as instrumental variables (IVs), or proxies, for a modifiable intermediate phenotype to investigate its risk on a disease. In other words, a valid set of IVs, that mimic the effect of an exposure causally linked to a disease, should also be associated with risk of disease, proportional to the effect of the IVs on the exposure[40]. This paradigm is analogous to that of a RCT, where stronger doses of drugs have a greater effect on the levels of the causal biomarker and the resultant effect on the outcome is also greater[41] (Figure 1-1). For instance, many independent trials have shown statins to reduce low-density lipoprotein cholesterol (LDL-C) levels and risk of coronary artery disease (CAD), proportional to the dose of the statin, owing to the causal link between LDL-C and CAD[42].

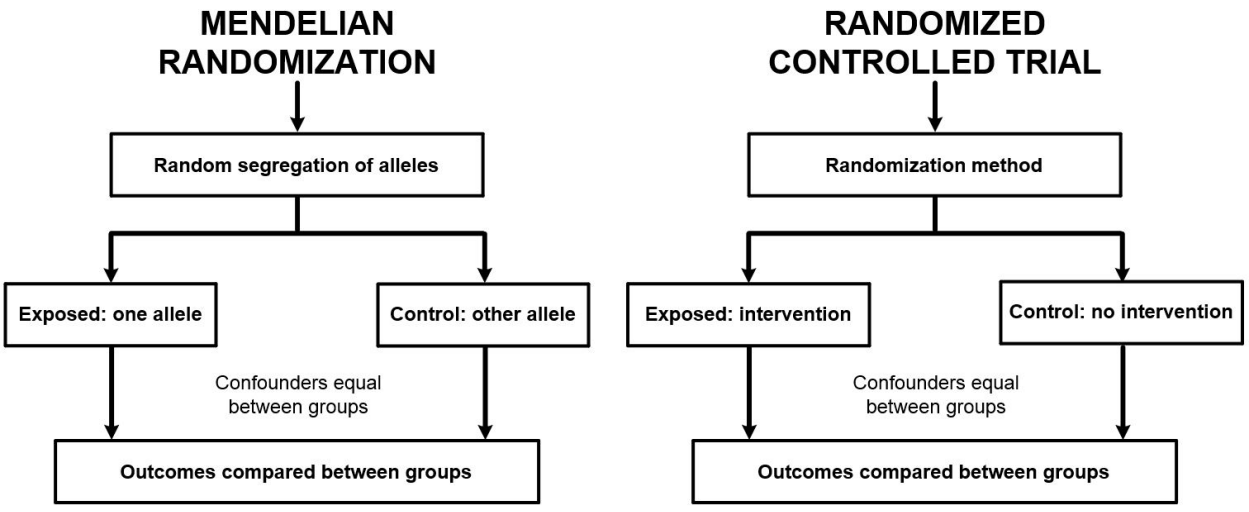


Figure 1-1: Comparison of Mendelian randomization studies to randomized controlled trials.

The use of MR has become an increasingly common technique to infer causal relationships in health-related research which has led to several major discoveries

and validations in the field of cardiometabolic disorders. MR analyses have been particularly important in providing insights into the role of LDL-C on CAD development. Using all known genetic variants associated with LDL-C as instruments, the causal role of LDL-C was confirmed through MR[43]. Furthermore, MR models have also had success predicting the effect of specific LDL-C lowering drugs by restricting the analysis to the gene target of the drug in question[44–46]. For example, the cardiovascular benefit of ezetimibe, which lowers LDL-C through inhibition of Niemann-Pick C1-like 1 (NPC1L1), was once a question of debate which was in part resolved through MR studies. By restricting the analysis to genetic variants at the *NPC1L1*, to mimic the effect of ezetimibe, researchers provided strong evidence for inhibition of NPC1L1 to decrease risk of CAD[47]. Subsequently, this question was investigated in a randomized trial where consistent and conclusive evidence confirmed the longstanding hypothesis that ezetimibe, indeed, prevents future cardiovascular events[48]. Moreover, MR studies have also encouraged the development of novel drugs, such as PCSK9 inhibitors, which have been recently shown to reduce cardiovascular events in phase III clinical trials[49,50]. Applications of MR extend beyond that of assessing the intended effect of a drug, to revealing potential adverse or favorable side effects. For instance, such studies were again successful in replicating trial data showing an on-target effect of statins on increased risk of type-2 diabetes (T2D)[51,52]. Similarly, MR studies have found that other LDL-C lowering drugs are likely to exert a similar effect on T2D risk[53–56]. In addition to elucidating the complex relationship of LDL-C and CAD, MR studies have also revealed many other notable associations in CAD including a causal effect of blood pressure, diabetes, adiposity and alcohol[57–60], and have convincingly excluded a causal role for C-reactive protein (CRP) and high-density lipoprotein cholesterol[61–63]. These findings, among many others, have increased our understanding of the aetiology and pathophysiology of cardiometabolic disorders and paved the way for improved treatments and prevention.

Methodology and study design

The wealth of information provided by MR studies has been made possible by both the increased precision and decreased price of genotyping platforms interrogating single nucleotide polymorphisms (SNPs) across the entire genome. Historically, this data has been used in genome-wide association studies (GWAS) to systematically assess the relationship between millions of common genetic variants (SNPs) and a single trait[64–67]. While the MR framework relies on this same genetic data, the fundamental aim is qualitatively different with far greater implications. However, the linchpin of a MR analysis, to correctly infer a causal relationship, depends on the use of genetic variants as IVs for the exposure of interest. To be used as a valid IV, the genetic variant(s) must follow three important assumptions: (1) the instrument must be associated with the exposure of interest, (2) the instrument must not be associated with confounding factors in the exposure-outcome association, and (3) the genetic variant must only affect the outcome through the exposure variable[68].

The first assumption can be easily tested by evaluating the strength of association between the instrument and the exposure. Although the second assumption cannot be proven for all possible confounders, instruments should be tested against likely and measured confounders. On the other hand, the second assumption is valid even in the absence of such tests according to Mendel's second law of random assortment. The third assumption, however, is likely the most problematic for MR studies. Because it is extremely difficult to distinguish the exact biological effect of a single genetic variant, it is nearly impossible to prove for certain that the genetic instrument affects the outcome only through the exposure variable. A violation of this assumption, whereby the genetic variant(s) has affects beyond those on the exposure, is known as pleiotropy[69]. Pleiotropy can occur in two forms: horizontal and vertical. Horizontal pleiotropy refers to situations where the genetic variant(s) has effects through another pathway or trait independent to the one under

investigation. Vertical pleiotropy, on the other hand, occurs when the genetic variant(s) affect multiple exposures on the same pathway as interest, and does not bias results[70]. A number of promising strategies have been developed to mitigate bias due to horizontal pleiotropy. If the exposure of interest is a protein then the optimal instrument(s) would be genetic variants lying within or near the gene coding for the protein itself. In this way, the presumed effect of the instrument on the outcome can only possibly be through the exposure. Often, however, there are no valid genetic instruments near a relevant gene or the exposure under study is not a protein with a gene directly responsible for its expression (e.g. alcohol consumption and BMI). In this case, an alternative solution is to use multiple genetic variants across the genome and look for a homogenous affect across all instruments[71].

Provided the core assumptions are satisfied and adequate statistical power, estimating the causal relationship between and an exposure and an outcome is straightforward. The conventional approach is to compare the effect from the SNP-outcome relationship ($\beta_{outcome}$) to the effect from the SNP-exposure relationship ($\beta_{exposure)}$ to derive a causal estimate which corresponds to a unit increase in the exposure variable[72]. This is known as the ratio method and can be used for a single SNP or multiple SNPs in combination. More precisely, the causal estimate is given by the estimate obtained by regressing $\beta_{outcome}$ onto $\beta_{exposure}$, with an intercept forced through the origin. A common variation of this method is the inverse-variance weighted (IVW) estimate to account for the strength of the $\beta_{outcome}$ relationship[73]. The IVW method, as the name suggests, weights each instrument according to the inverse variance of gene-outcome association. Other adaptations and extensions to the ratio method have also been developed to deal with pleiotropic bias. The Egger method, for example, allows for relaxation of the IV assumption requiring of a direct effect of the genetic instrument on the outcome[74]. Specifically, MR-Egger takes a similar approach as Egger regression, designed to mitigate small study bias in clinical trials, and allows the y-intercept to float rather

than be fixed to zero[75]. In this case, a deviation from the origin would suggest the presence of pleiotropy. Additionally, because pleiotropic effects are absorbed, into the y-intercept, the resulting slope is a reliable causal estimate even in the presence of pleiotropy, under certain assumptions[74]. Alternative MR models to mitigate bias caused by invalid instruments include the simple median and weighted median estimates[76]. The simple median estimate is the median ratio estimate when ratios from each instrument are arranged from the smallest to largest. This method suffers from limitations by excluding all instruments except one (or two if the number of genetic variants is odd) and is said to be inefficient. Conversely, the weighted median estimate retains all variants, weighting each ratio by its proximity to the median location, such that ratios falling in the middle of the distribution receive the highest weight. However, all these MR methods rely on comparing the association of the IV(s) with both the exposure and the outcome of interest. Assuming no bias caused by pleiotropy, the IVW is the best choice due to highest statistical power. In situations where the assumptions may be violated, other techniques should be employed as a sensitivity analysis to ensure results are consistent between models.

Ordinarily, the exposure, outcome and genetic instrument(s) have all been measured in a single sample. However, recently many studies have adapted a two-sample MR design where $\beta_{outcome}$ and $\beta_{exposure}$ are ascertained in independent samples[77–80]. This design offers many advantages over the typical one-sample design, as obtaining all the necessary data for a MR analysis in a single, large sample can be both expensive and time-consuming[81]. Additionally, this design is free of the weak IV bias plaguing one-sample designs, which states that weak instruments lead to biased estimates towards the causal association[82]. Conversely, when genetic effects are obtained from independent samples, weak instruments lead to a bias towards the null hypothesis[83]. This two-sample MR framework, has another important implication as summary-level data from large, publicly-available collaborations can be used for either the instrument-exposure or the instrument-

outcome relationship. In fact, depending on the question of interest, both estimates can be obtained from published GWAS, requiring no data to be physically collected to conduct the MR analysis. As genetic researchers continue to pool resources across GWAS, increasing sample sizes, statistical power and the number of genome-wide significant loci, our ability to detect causal associations through MR will undoubtedly increase as well. These efforts allow researchers to interrogate numerous causal relationships at a very low cost with relatively limited resources. Therefore, MR promises to provide an extremely valuable and efficient resource to identify novel causal markers of cardiometabolic traits, which will in turn help prioritize future drug targets and ultimately result in improved treatment and prevention of disease.

Considerations

Several potential caveats should be considered when conducting a MR analysis. First, and arguably the most important limitation to MR studies, is the issue of pleiotropy, whereby a genetic instrument has effects beyond its effect on the exposure of interest. In the presence of pleiotropy, MR studies can be easily misinterpreted, giving rise to both false positive and false negative results. Specifically, pleiotropic effects can counteract an effect of the variant on the disease acting via the causal biomarker, and thus lead to a null association, when in fact there is a causal relationship. On the other hand, pleiotropic effects can lead to a false positive association between a variant and disease that can result in a mistaken causal interpretation between a biomarker and disease. For example, the known causal effect of LDL-cholesterol on CAD can lead false-positive MR findings showing a protective effect of HDL-cholesterol and CAD if the genetic instrument(s) used has pleiotropic effects by both decreasing HDL-cholesterol and increasing LDL-cholesterol levels. In this case, the causal factor is known through functional experiments and RCTs, however often it is unclear which biomarker is responsible for the observed effects. Confounding due to pleiotropy is least likely when genetic

instruments are used that lie near the gene for the exposure under study. In situations where there is no possible instrument or there is no single gene responsible for the exposure of interest (e.g. BMI), then many instruments across the genome can be used to evaluate the relationship in an MR analysis. Pleiotropy is unlikely to be confounding the relationship if all instruments show a strong, consistent, directional effect on the exposure and outcome of interest.

Genetic loci in close proximity on a given chromosome tend to be inherited together. This observation is known as linkage disequilibrium (LD), and can result in misinterpretations of MR results, similar to pleiotropic bias. For example, a SNP affecting the expression of gene A may be in LD with a SNP affecting the expression of gene B. If biomarker B, encoded by gene B, exerts a causal effect on the disease, then a MR analysis investigating the effect of the biomarker produced by gene A on a disease could result in false positive findings. In this scenario, biomarker B is a causal factor, while biomarker A is merely a bystander with no causal effect. To properly evaluate the effect of biomarker A, genetic instruments should be used that show no LD with variants in gene B that may circumvent the relationship. This phenomenon is especially problematic in gene clusters, as it is often impossible to disentangle the causal biomarker unless all biomarkers for each gene in the cluster are measured.

Limited statistical power can influence the ability of a MR study to inform on a causal relationship between an exposure and an outcome. Specifically, the effect of SNP on a phenotype (both exposure and outcome) can be difficult to ascertain. First, there are usually multiple genetic and environmental factors influencing the variability of a trait and consequently, the effect of a single SNP can be very small. Secondly, risk factors often act together to exert their effect on a disease such that a causal biomarker may only be responsible for a portion of the resulting outcome. For example, while LDL-cholesterol is a known causal risk factor for CAD, rarely are elevated cholesterol levels sufficient to lead to CAD. Indeed, most CAD cases

display a combination of risk factors, such as obesity, smoking, older age, among others. Therefore, to properly evaluate causal associations through MR, a sufficiently large sample size is needed, particularly to detect relationships with small effects.

1.2.2 Admixture Mapping

Classic epidemiological studies have demonstrated differential risk between ethnic groups for numerous diseases, but it is difficult to infer whether this is a consequence of genetic or environmental factors. Admixture mapping is a powerful tool used in genetic epidemiological studies that overcomes this challenge and allows mapping of genes conferring differential risk. Genetic admixture occurs when two or more previously independent populations interbreed, resulting in the introduction of new genetic lineages. Admixture mapping is a method applied to recently admixed populations used to localize disease causing genetic variants that differ in frequency across ancestral groups[84]. Most genetic variation is shared between populations, but allele frequencies can vary substantially. For instance, the null Duffy antigen has frequencies of ~100% in West African populations and ~0% in populations outside of Africa[85]. The approach is based on the assumption that increased proportion of ancestry from the population with a greater risk of the disease will be observed in patients near a disease-causing gene. In this way, differential risk across ancestral groups can be observed at specific genetic loci[86]. Traditional association mapping techniques indirectly measure recombination across many generations, as far back as the most recent common ancestor for the entire sample. Similarly, in admixture mapping studies, recombination is assessed to localize genetic signals, however, most admixture studies involve recent admixture (<20 generations). Therefore, the resolution of admixture mapping is inferior to genome-wide association studies (GWAS), but superior to linkage analysis. Admixture mapping has been an effective technique, particularly in African Americans, in identifying novel loci involved in disease, namely, 6q21 for

hypertension[87], 8q24 for prostate cancer[88] and MYH for focal segmental glomerulosclerosis[89].

Admixture mapping studies offer the unique advantage over single-ancestry studies at identifying genes conferring differential risk between populations. Additionally, such studies are able to inform on the genetic factors that contribute to observed ancestry-level differences versus differences due to socio-economic and lifestyle factors[90]. Indeed, this is a worthwhile investigation as clinical diagnostic and prognostic measures are often determined from European populations and applied broadly among other ethnic groups. However, specific cut-off values and prognostic measures likely apply to various populations, and should therefore be implemented accordingly. Indeed, little is known about the generalizability of these markers among other ethnic groups. While differences in biomarkers levels have been observed between ethnic groups, the reasons for these differences has not been elucidated. Admixture mapping studies may shed light on these observations and pave the way for better informed, ethnic specific defined cut-offs[91].

1.3 Study Population

1.3.1 ORIGIN-Trial

The ORIGIN (Outcome Reduction with an Initial Glargine Intervention) trial was an international outcomes multicenter, two-arm, randomized control trial designed and led by researchers at McMaster University. The study population included 12,537 participants from 40 countries with either prediabetes or early type 2 diabetes. The goal of the study was to determine whether cardiovascular events and other clinical outcomes, could be reduced in a population with dysglycemia by: (1) normalizing fasting glucose levels with basal insulin glargine (versus standard glycemic control), and/or (2) a 1 gram omega-3 fatty acid supplement (versus placebo)[92]. The results of the trial are published in two articles in the *New England Journal of*

Medicine, in which it was reported that both interventions had a neutral effect on CVD outcomes[93,94]. The ORIGIN data is a well characterized global population that was followed for approximately 6 years for the development of a number of health outcomes. Three blood samples were taken and stored in a subset of participants at baseline, follow-up and end of study for future analysis. All clinical data collected in this trial is currently being stored at the ORIGIN Project Office at the Population Heath Research Institute (an organization jointly connected with McMaster University and Hamilton Health Sciences) in Hamilton, Ontario. A subset of 8,494 ORIGIN participants consented to the collection and storage of blood samples for future analyses. After completion of the trial, baseline samples for these participants were analyzed for a comprehensive panel of biomarkers. The results from this project have been published in *Circulation* where a novel panel of biomarkers was identified for an independent association with cardiovascular outcomes beyond classical risk factors[95]. Additionally, genotyping has been performed on 5,433 ORIGIN participants who consented to genetic analysis and provided a sample suitable for DNA extraction. The ORIGIN genetic population will serve as the study population for the current project where we propose to investigate the relationship between biomarkers, genetic variants, and cardiovascular events.

1.3.2 Quality Control and Processing of Genetic Data

Genome-wide association studies (GWAS) are used to identify common single nucleotide polymorphisms (SNPs) that influence human traits. However, the ability of a GWAS to detect true genetic associations directly depends on the quality of the data[96]. Improper data cleaning can compromise the results of simple association tests, leading to both false-positive and false-negative associations[97]. Typically, GWAS involve large sample sizes to detect extremely small effects in hundreds of thousands of polymorphisms; therefore, even small artifactual differences can result in false-positives. Likewise, false-negatives can be

increased due to failure to control various experimental factors, such as low quality samples, poorly-performed SNP assays, and sample ID error; these factors lead to "noise" and reduce the power of the association test[96].

Quality control (QC) measures have been discussed and reported on in a number of manuscripts, and are usually completed using a combination of procedures[96–100]. Seven main steps were followed: (1): removal of non-informative SNPs, (2) examination of individual samples and SNPs for genotyping completeness, (3) sex check, (4) ethnicity check, (5) relatedness, (6) Hard-Weinberg deviation and (7) minor allele frequency. Table 1-1 illustrates a summary of the QC steps taken in the ORIGIN study. PLINK (version 1.07)[101] was used for the majority of the quality control procedures, R (version 3.5)was used for simple data manipulation and formatting, and Genome-wide Complex Trait Analysis (GCTA, version 1.91.3)[102] was used for ethnicity check and relatedness. Other pre-processing steps included principal component calculation for population adjustment, imputation of non-typed genotypes using IMPUTE2[103] and computation of local and global ancestry for admixture mapping.

Table 1-1: Summary of quality control steps in the ORIGIN genetic data.

No.	Step	Number of Samples Removed	Number of SNPs Removed	Samples Remaining	SNPs Remaining
1	Initial Population	NA	NA	5,078	545,555
2	MAF = 0	-	3,011	5,078	542,544
3	> 10% missingness/SNP	-	13,903	5,078	528,641
4	>10% missingness/sample	109	-	4,969	528,641
5	> 5% missingness/SNP	-	10,707	4,969	517,934
6	> 5% missingness/sample	16	-	4,953	517,934
7	> 1% missingness/SNP	-	41,641	4,953	476,293
8	> 1% missingness/sample	97	-	4,856	476,293

9	Sex check	21	-	4,835	476,293
10	Centers > 20% sex error rate	40	-	4,795	476,293
11	Samples without reported ethnicity	1	-	4,794	476,293
12	Ethnicity check	129	-	4,665	476,293
13	Estimated relatedness > 0.2	88	-	4,577	476,293
14	Ethnicities n < 250	187	-	4,147	476,293
15	Hardy Weinberg (in any ethnicity)	-	851	4,147	475,442
16	MAF < 0.01 (in all ethnicities)	-	191,418	**4,147**	**284,024**

Removal of Non-informative SNPs

Firstly, all SNPs with a minor allele frequency (MAF) of 0 were removed, as they have no ability to provide any additional information on explaining the variation of an outcome of event. A SNP with a MAF of 0 is known as a monomorphic SNP and signifies that all individuals in the study sample have the same genotype at that location. In ORIGN, 3,011 SNPs had a MAF of 0 and were removed.

Genotyping Completeness and Accuracy

Current genotyping technology is very reliable and accurate, producing data with high call rates and high accuracy[97]. However, it is still essential to consider both these measures, as reagents and instruments may vary between labs. Individual samples and individual SNPs were checked for completeness in parallel. Missing call rate is a measure of data completeness, but more importantly it is also a measure of data quality because genotype missingness is often non-random[96]. Samples with high missing rates often imply poor DNA quality and are therefore removed. Similarly, SNPs with high missing rates are often a result of poor primer design and non-specific DNA binding to a SNP probe, such SNPs must be removed[96]. SNPs with greater than 10% missing genotypes were removed first, followed by samples with greater than 10% missing genotypes. This process was

then repeated for a threshold of 5% and 1% (steps 2-8). Following this procedure, all SNPs and samples remaining exceeded a 99% genotype completion rate, which is the standard for GWAS. Originally, this process was reversed, such that low-genotyped samples were removed first, followed by SNPs, and repeated at each threshold. However, this order resulted in removing 399 samples (7.9%) versus only 222 (4.4%) with the current order. In total, 66,251 SNPs were removed over the three thresholds.

Sex Check

A sex check was performed which compares the genetic sex of individuals to their self-reported sex. Sex checks are important in identifying cases of individual misrepresentation, careless error or possible sample mix-ups[96]. The sex check procedure consists of two steps. Firstly, individuals whose genetic sex did not match their reported sex were removed. PLINK assigns a genetic sex to each individual using the heterozygosity rates on the X-chromosome. If the X-chromosome homozygosity rate is greater than 0.8, a male call is made; conversely, if the homozygosity rate is less than 0.2, a female call is made. If the homozygosity rate falls between 0.2 and 0.8, the genetic sex is said to be ambiguous. Samples who had an ambiguous genetic sex given by PLINK, were also compared to the sex assigned by the Illumina chip in the lab. If the Illumina-assigned sex was inconsistent with the self-reported sex, they were removed. In total, 21 samples were removed due to inconsistences between self-reported and genetic sex. Secondly, the rates of these inconsistencies were examined for each data collection center. Data from centers with greater than 20% sex error rates in either sex were discarded. This was done to combat the likelihood of an individual reporting error or mix-up not being detected, by using sex alone as a metric. An additional 40 samples were removed due to one center having unexplainably high sex error rates.

Ethnicity Check

A similar concept was followed for the investigation of population structure, which involved comparing the genetic ancestry with the self-reported ancestry. To do this, principal component analysis (PCA) was performed using the software GCTA[102]. PCA is a complex mathematical procedure used on multi-dimensional data which reveals the internal structure of the data through dimensionality reduction[104]. An algorithm is followed in such a way to preserve as much of the variance and information in the data as possible[105]. This is accomplished by identifying directions in the data where the variation is maximal, these directions are known as principal components[106]. The number of principal components is always less than or equal to the number of original variables or dimensions, in genotype data this is equal to the number of SNPs. The principal components, or axes of variation, are by definition orthogonal, and therefore uncorrelated[107]. Through the use of orthogonal transformation, a set of possibly correlated M samples can be converted into a set of values of K uncorrelated variables. The first principal component (PC1) accounts for as much of the variability in the data as possible[105]. Each succeeding component must be uncorrelated to all preceding components while still explaining as much variance as possible. Each sample is mathematically projected onto each principal component resulting in a new data set of M samples and K projections. When investigating population structure, it is often sufficient to only consider the first two principal components because they have captured a significant portion of the variability in the data set[106]. This conveniently reduces the dimensionality considerably such that a large amount of the variability in the data can be viewed on an x-y axis by plotting PC1 against PC2, where each point is a sample in the data set. Individuals with similar inferred genetic ancestry will cluster together on the plot. By color-coding the plots based on self-reported ethnicity, those whose self-reported ethnicity differs from their inferred genetic ancestry can be identified[107].

Prior to performing PCA on the ORIGIN data, one sample was removed because it had no associated reported ethnicity. Following this, the data was prepared for PCA. When investigating population structure, reference samples of known ethnicities should be selected and used as a control to ensure that PCA is computed correctly and that the study sample follows a reasonable structure[108]. This is accomplished by comparing the results of the two independent samples, such that the graphs (PC1 versus PC2) should exhibit similar patterns. The publicly available International HapMap Project data[109] was selected as the reference for ORIGIN. Only SNPs common between HapMap and ORIGIN were chosen for the analysis to guarantee an accurate comparison. PCA was then executed using GCTA[102] on 179,842 SNPs in 1,184 HapMap samples and 4,794 ORIGIN samples. As anticipated, the results of the HapMap PCA displayed clear, distinct clusters for each ethnicity (Figure 1-2). Likewise, the ORIGIN PCA displayed clusters of ethnicities, although not nearly as distinct (Figure 1-3). These results suggest that samples exist whose self-reported ethnicity was inconsistent with their genetic ethnicity. Therefore, samples that fell three standard deviations away from the mean of principal component one or two of their self-reported ethnicity were considered to have ancestral inconsistencies between genetics and self-identification and were consequently removed; 129 samples were removed at this step. The clusters became more distinct after removing these individuals (Figure 1-4).

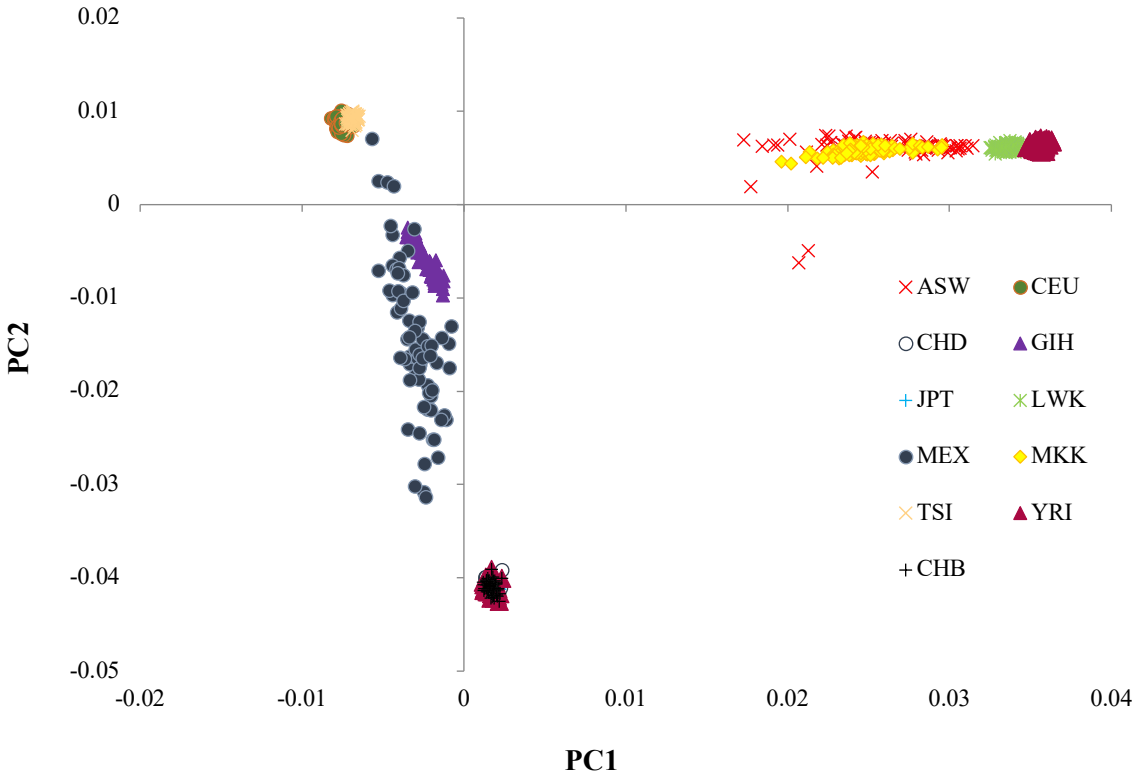


Figure 1-2: Hap Map PCA.

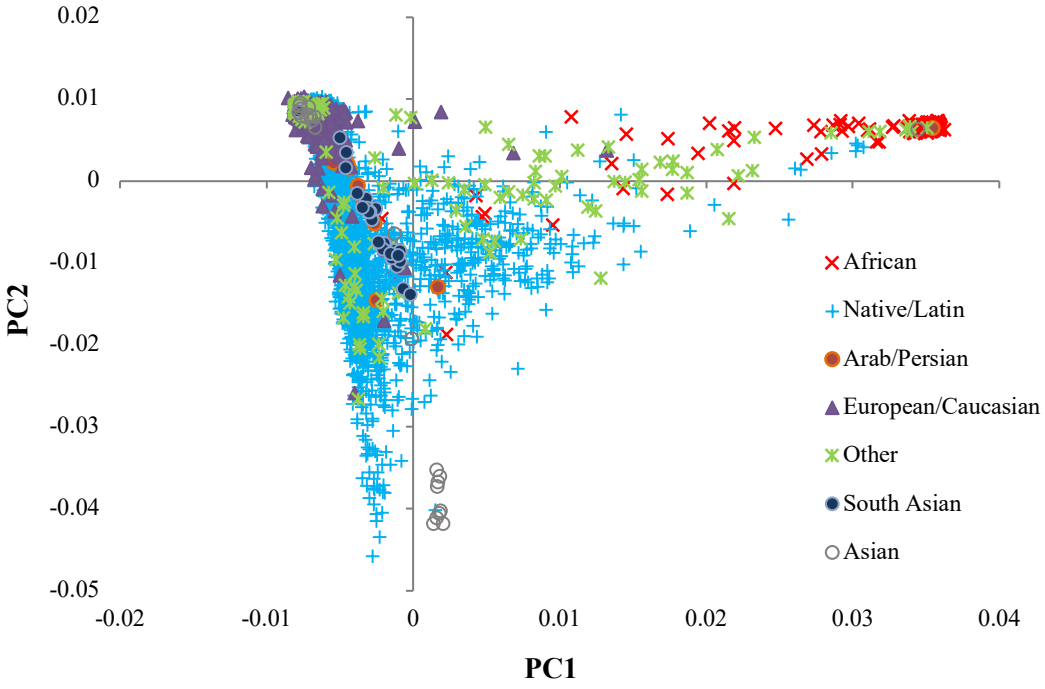


Figure 1-3: Origin PCA pre-filtering

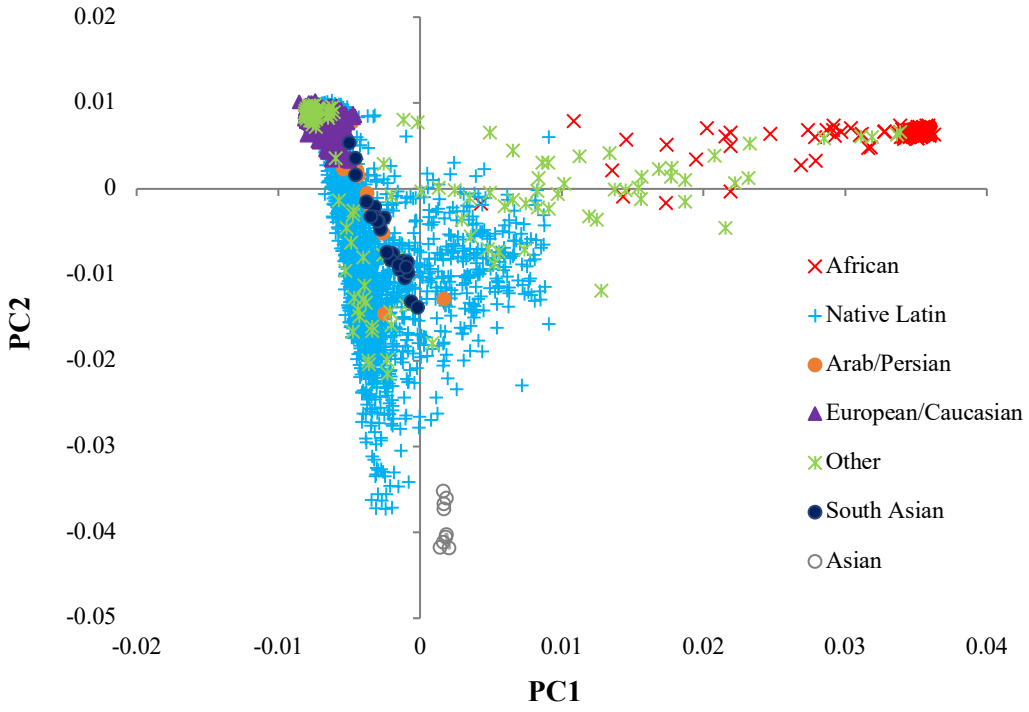


Figure 1-4: Origin PCA post-filtering.

Hardy-Weinberg

Before completing the final pre-processing steps for the GWAS, ethnicities with less than 250 individuals were removed for accurate results and appropriate statistical power when meta-analyzing the association results across ethnicities. The following five ethnic groups were discarded for the current GWAS: African, Asian, Arab/Persian, South Asian and Others; a total of 187 individuals were removed. European/Caucasian and Native Latin ethnic groups remained in the study sample.

SNPs were screened for deviation from Hardy-Weinberg (HW) equilibrium. Briefly, the HW theorem states that allele and genotype frequencies in a population will remain constant across generations given a specific set of conditions. These frequencies can be estimated using the HW equation: $p^2 + 2pq + q^2 = 1$, where p

and q are the frequencies of the two alleles (denoted A and a), respectively, and p^2 and q^2 represent the expected genotype frequencies for the homozygotes, AA and aa, and 2pq is the expected genotype frequency for the heterozygotes, Aa[110]. Departure from this equilibrium can be indicative of selection, mixture of heterogeneous populations, unexplained relatedness (such as an inbreeding population), population stratification, or genotyping errors (either misclassification or discriminatory drop of a given allele)[97,111]. There is no universally recognized HW p-value threshold, but typically they range from 10^{-5} to 10^{-7}, where a p-value smaller than the selected threshold indicates significant deviation from HW equilibrium. In ORIGIN, a threshold of 10^{-7} was used, and 851 SNPs significantly deviated from HW equilibrium and were consequently removed.

Minor Allele Frequency

For the final quality control step, SNPs were filtered based on minor allele frequency. This is a crucial step to ensure adequate statistical power when performing the association analyses[37]. For rare SNPs, alternative statistical methods must be utilized as power to detect a genome-wide association is extremely low even for large effect sizes (OR between 1.3 and 1.7) and with a reasonably sized dataset (n=10,000)[97]. (Analysis of rare SNPs will be explored as an independent project.) The threshold for the MAF filter varies based on sample size and expected effect size[96]. In ORIGIN, a threshold of 1% was used. Because MAF can vary depending on ancestral background, SNPs were only removed if they had a MAF less than 1% in all three ethnicities, rather than just a single ethnic group. This is done to preserve as many of the SNPs as possible. However, it is important to mention that association signals at SNPs with MAFs less than the specified threshold (1%) in any ethnic group will be interpreted with caution. The MAF filter removed 191,418 SNPs. This completed the quality control steps. The final dataset for association analyses contained 4,390 individuals and 284,024 SNPs.

Final PCA Calculation for Adjustment

Principal components were recalculated for all remaining study participants to use as adjustment covariates in the logistic and linear regression models. Before this computation, the number of SNPs and the choice of which SNPs to include must be determined. This decision is not straightforward. It is possible, albeit conceptually demanding, to include all SNPs in the database and would use all available information to infer genetic ancestry, as was the case with ethnicity check. However, this is not necessarily the best option at this step. A database which has been constructed using a whole-genome array will contain clusters of highly correlated SNPs and as a result may have a very strong influence on certain principal components, limiting their ability to detect and control for population structure[96]. This has been noted previously in a number of studies[112,113]. For this calculation, precise derivation of the individual principal components is crucial, as they are used to adjust for differences within ethnic groups[96]. Therefore, SNPs were pruned for linkage disequilibrium with a pairwise threshold of r^2=0.5. Conversely, with ethnicity check, PCA is merely used to gain an understanding of the ethnic groups in the study population and confirm that genetic ancestry is relatively consistent with the reported ethnicity, therefore pruning is not necessary.

Imputation

Imputation is a method to statistically infer untyped genotypes to expand the set of SNPs available for association testing. It is accomplished through the use of a reference panel of individuals genotyped at a dense set of SNPs to predict unobserved genotypes in the study population, which was genotyped at only a subset of these SNPs[28]. The 1000 Genomes Project[114] was used as the reference panel for ORIGIN. Imputation was performed using the software IMPUTE2[115]. Over 10 million SNPs were imputed, allowing for comprehensive coverage of known genetic variants and the opportunity to meta-analyze with external studies

1.4 References

1. Mendis, S., Puska, P. & Norrving, B. Global atlas on cardiovascular disease prevention and control. *World Heal. Organ.* 2–14 (2011).

2. Nabel, E. G. Cardiovascular disease. *N. Engl. J. Med.* **349,** 60–72 (2003).

3. Organization, W. H. Prevention of Cardiovascular Disease Prevention of Cardiovascular Disease. *World Heal. Organ.* 1–30 (2007). doi:10.1093/innovait/inr119

4. Yusuf, S., Reddy, S., Ounpuu, S. & Anand, S. Global burden of cardiovascular diseases: part I: general considerations, the epidemiologic transition, risk factors, and impact of urbanization. *Circulation* **104,** 2746–53 (2001).

5. McGill, H. C., McMahan, C. A. & Gidding, S. S. Preventing heart disease in the 21st century: Implications of the pathobiological determinants of atherosclerosis in youth (PDAY) study. *Circulation* **117,** 1216–1227 (2008).

6. Public Health Agency of Canada. The Growing Burden of Heart Disease and Stroke in Canada. (2003). Available at: http://publications.gc.ca/collections/Collection/H1-10-2003E.pdf.

7. Olshansky, S. J. *et al.* A potential decline in life expectancy in the United States in the 21st century. *N. Engl. J. Med.* **352,** 1138–1145 (2005).

8. ADVANCE Collaborative Group *et al.* Intensive blood glucose control and vascular outcomes in patients with type 2 diabetes. *N. Engl. J. Med.* **358,** 2560–72 (2008).

9. Grundy, S. M. *et al.* Diabetes and cardiovascular disease: a statement for healthcare professionals from the American Heart Association. *Circulation*

100, 1134–46 (1999).

10. Emerging Risk Factors Collaboration *et al.* Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. *Lancet* **375,** 2215–22 (2010).

11. Saydah, S. H., Fradkin, J. & Cowie, C. C. Poor control of risk factors for vascular disease among adults with previously diagnosed diabetes. *JAMA* **291,** 335–42 (2004).

12. Harris, M. I. Diabetes in America: Epidemiology and scope of the problem. *Diabetes Care* **21,** (1998).

13. Stamler, J., Vaccaro, O., Neaton, J. D. & Wentworth, D. Diabetes, other risk factors, and 12-yr cardiovascular mortality for men screened in the Multiple Risk Factor Intervention Trial. *Diabetes Care* **16,** 434–44 (1993).

14. Pearson, T. A. *et al.* AHA Guidelines for Primary Prevention of Cardiovascular Disease and Stroke: 2002 Update: Consensus panel guide to comprehensive risk reduction for adult patients without coronary or other atherosclerotic vascular diseases. *Circulation* **106,** 388–391 (2002).

15. Stoner, L. *et al.* Inflammatory biomarkers for predicting cardiovascular disease. *Clin. Biochem.* **46,** 1353–1371 (2013).

16. Raitakari, O. T. *et al.* Cardiovascular risk factors in childhood and carotid artery intima-media thickness in adulthood: the Cardiovascular Risk in Young Finns Study. *JAMA* **290,** 2277–83 (2003).

17. Berenson, G. S. *et al.* Association between multiple cardiovascular risk factors and atherosclerosis in children and young adults. The Bogalusa Heart Study. *N. Engl. J. Med.* **338,** 1650–6 (1998).

18. Vasan, R. S. Biomarkers of cardiovascular disease: Molecular basis and

practical considerations. *Circulation* **113,** 2335–2362 (2006).

19. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin. Pharmacol. Ther.* **69,** 89–95 (2001).

20. Ridker, P. M. LDL cholesterol: Controversies and future therapeutic directions. *Lancet* **384,** 607–617 (2014).

21. Lavie, C. J., Milani, R. V. & Ventura, H. O. Obesity and Cardiovascular Disease. Risk Factor, Paradox, and Impact of Weight Loss. *J. Am. Coll. Cardiol.* **53,** 1925–1932 (2009).

22. Sowers, J. R., Epstein, M. & Frohlich, E. D. Diabetes, hypertension, and cardiovascular disease: an update. *Hypertension* **37,** 1053–9 (2001).

23. Patil, H., Vaidya, O. & Bogart, D. A review of causes and systemic approach to cardiac troponin elevation. *Clin. Cardiol.* **34,** 723–728 (2011).

24. Kanda, T. & Takahashi, T. Interleukin-6 and cardiovascular diseases. *Jpn. Heart J.* **45,** 183–193 (2004).

25. Casas, J. P. *et al.* Insight into the nature of the CRP-coronary event association using Mendelian randomization. *Int. J. Epidemiol.* **35,** 922–931 (2006).

26. Averna, M. & Noto, D. Clinical utility of novel biomarkers for cardiovascular disease risk stratification. *Intern. Emerg. Med.* **7,** 263–270 (2012).

27. Zakynthinos, E. & Pappa, N. Inflammatory biomarkers in coronary artery disease. *J. Cardiol.* **53,** 317–33 (2009).

28. Sun, X., Jia, Z. & Sun, X. A brief review of biomarkers for preventing and treating cardiovascular diseases. *J. Cardiovasc. Dis. Res.* **3,** 251–254

(2012).

29. Upadhyay, R. K. Emerging risk biomarkers in cardiovascular diseases and disorders. *J. Lipids* **2015,** 971453 (2015).

30. Fewell, Z., Davey Smith, G. & Sterne, J. A. C. The impact of residual and unmeasured confounding in epidemiologic studies: A simulation study. *Am. J. Epidemiol.* **166,** 646–655 (2007).

31. Barter, P. J. *et al.* Effects of torcetrapib in patients at high risk for coronary events. *N. Engl. J. Med.* **357,** 2109–22 (2007).

32. Harrison, R. K. Phase II and phase III failures: 2013–2015. *Nat. Rev. Drug Discov.* **15,** 1–2 (2016).

33. Fordyce, C. B. *et al.* Cardiovascular drug development: Is it dead or just hibernating? *J. Am. Coll. Cardiol.* **65,** 1567–1582 (2015).

34. Jansen, H., Samani, N. J. & Schunkert, H. Mendelian randomization studies in coronary artery disease. *Eur. Heart J.* **35,** 1917–1924 (2014).

35. Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.* **23,** R89-98 (2014).

36. Smith, G. D. & Ebrahim, S. 'Mendelian randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* **32,** 1–22 (2003).

37. Swerdlow, D. I. *et al.* Selecting instruments for Mendelian randomization in the wake of genome-wide association studies. *Int. J. Epidemiol.* dyw088 (2016). doi:10.1093/ije/dyw088

38. Evans, D. M. & Davey Smith, G. Mendelian Randomization: New

Applications in the Coming Age of Hypothesis-Free Causality. *Annu. Rev. Genomics Hum. Genet.* 1–24 (2015). doi:10.1146/annurev-genom-090314-050016

39. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47,** 856–860 (2015).

40. Smith, G. D., Timpson, N. & Ebrahim, S. Strengthening causal inference in cardiovascular epidemiology through Mendelian randomization. *Ann. Med.* **40,** 524–41 (2008).

41. Hingorani, A. & Humphries, S. Nature's randomised trials. *Lancet (London, England)* **366,** 1906–8 (2005).

42. Cannon, C. P. *et al.* Intensive versus moderate lipid lowering with statins after acute coronary syndromes. *N. Engl. J. Med.* **350,** 1495–504 (2004).

43. Ference, B. A. *et al.* Effect of long-term exposure to lower low-density lipoprotein cholesterol beginning early in life on the risk of coronary heart disease: A mendelian randomization analysis. *J. Am. Coll. Cardiol.* **60,** 2631–2639 (2012).

44. Holmes, M. V. *et al.* Secretory Phospholipase A 2 -IIA and Cardiovascular Disease. *J. Am. Coll. Cardiol.* **62,** 1966–1976 (2013).

45. Würtz, P. *et al.* Metabolomic Profiling of Statin Use and Genetic Inhibition of HMG-CoA Reductase. *J. Am. Coll. Cardiol.* **67,** 1200–1210 (2016).

46. Ference, B. A., Majeed, F., Penumetcha, R., Flack, J. M. & Brook, R. D. Effect of naturally random allocation to lower low-density lipoprotein cholesterol on the risk of coronary heart disease mediated by polymorphisms in NPC1L1, HMGCR, or both: a 2 × 2 factorial Mendelian randomization study. *J. Am. Coll. Cardiol.* **65,** 1552–61 (2015).

47. Myocardial Infarction Genetics Consortium Investigators *et al.* Inactivating mutations in NPC1L1 and protection from coronary heart disease. *N. Engl. J. Med.* **371,** 2072–82 (2014).

48. Cannon, C. P. *et al.* Ezetimibe Added to Statin Therapy after Acute Coronary Syndromes. *N. Engl. J. Med.* **372,** 2387–97 (2015).

49. Giugliano, R. P. & Sabatine, M. S. Are PCSK9 Inhibitors the Next Breakthrough in the Cardiovascular Field? *J. Am. Coll. Cardiol.* **65,** 2638–2651 (2015).

50. Sabatine, M. S. *et al.* Evolocumab and Clinical Outcomes in Patients with Cardiovascular Disease. *N. Engl. J. Med.* **376,** 1713–1722 (2017).

51. Collins, R. *et al.* Interpretation of the evidence for the efficacy and safety of statin therapy. *Lancet* **388,** 2532–2561 (2016).

52. Swerdlow, D. I. *et al.* HMG-coenzyme A reductase inhibition, type 2 diabetes, and bodyweight: evidence from genetic analysis and randomised trials. *Lancet (London, England)* **385,** 351–61 (2015).

53. Lotta, L. A. *et al.* Association Between Low-Density Lipoprotein Cholesterol-Lowering Genetic Variants and Risk of Type 2 Diabetes: A Meta-analysis. *JAMA* **316,** 1383–1391 (2016).

54. Schmidt, A. F. *et al.* PCSK9 genetic variants and risk of type 2 diabetes: a mendelian randomisation study. *Lancet Diabetes Endocrinol.* **8587,** 1–9 (2016).

55. Ference, B. A. *et al.* Variation in PCSK9 and HMGCR and Risk of Cardiovascular Disease and Diabetes. *N. Engl. J. Med.* **375,** 2144–2153 (2016).

56. White, J. *et al.* Association of Lipid Fractions With Risks for Coronary Artery

Disease and Diabetes. *JAMA Cardiol.* **366,** 1108–1118 (2016).

57. Lieb, W. *et al.* Genetic Predisposition to Higher Blood Pressure Increases Coronary Artery Disease Risk. *Hypertension* **61,** 995–1001 (2013).

58. Ahmad, O. S. *et al.* A Mendelian randomization study of the effect of type-2 diabetes on coronary heart disease. *Nat. Commun.* **6,** 7060 (2015).

59. Hägg, S. *et al.* Adiposity as a cause of cardiovascular disease: A Mendelian randomization study. *Int. J. Epidemiol.* **44,** 578–586 (2015).

60. Holmes, M. V. *et al.* Association between alcohol and cardiovascular disease: Mendelian randomisation analysis based on individual participant data. *BMJ* **349,** g4164 (2014).

61. C Reactive Protein Coronary Heart Disease Genetics Collaboration (CCGC) *et al.* Association between C reactive protein and coronary heart disease: mendelian randomisation analysis based on individual participant data. *BMJ* **342,** d548 (2011).

62. Elliott, P. *et al.* Genetic Loci associated with C-reactive protein levels and risk of coronary heart disease. *JAMA* **302,** 37–48 (2009).

63. Voight, B. F. *et al.* Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. *Lancet (London, England)* **380,** 572–80 (2012).

64. Nikpay, M. *et al.* A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* **47,** 1121–30 (2015).

65. Teslovich, T. Musunuru, K. Smith, a. E. Al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466,** 707–713 (2010).

66. Morris, A., Voight, B. & Teslovich, T. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **44,** 981–990 (2012).

67. Ehret, G. B. *et al.* Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* **478,** 103–109 (2011).

68. Bennett, D. A. & Holmes, M. V. Mendelian randomisation in cardiovascular research: an introduction for clinicians. *Heart* heartjnl-2016-310605 (2017). doi:10.1136/heartjnl-2016-310605

69. Hemani, G., Bowden, J. & Davey Smith, G. Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Hum. Mol. Genet.* **0,** ddy163 (2018).

70. Holmes, M. V., Ala-Korpela, M. & Smith, G. D. Mendelian randomization in cardiometabolic disease: challenges in evaluating causality. *Nat. Rev. Cardiol.* (2017). doi:10.1038/nrcardio.2017.78

71. Palmer, T. M. *et al.* Using multiple genetic variants as instrumental variables for modifiable risk factors. *Stat. Methods Med. Res.* **21,** 223–242 (2011).

72. Lawlor, D. A. *et al.* Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat. Med.* **27,** 1133–1163 (2008).

73. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet. Epidemiol.* **37,** 658–665 (2013).

74. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44,** 512–25 (2015).

75. Egger, M., Davey Smith, G., Schneider, M. & Minder, C. Bias in meta-

analysis detected by a simple, graphical test. *BMJ* **315,** 629–34 (1997).

76. Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet. Epidemiol.* **40,** 304–14 (2016).

77. Gill, D. *et al.* The Effect of Iron Status on Risk of Coronary Artery Disease: A Mendelian Randomization Study-Brief Report. *Arterioscler. Thromb. Vasc. Biol.* **37,** 1788–1792 (2017).

78. Mack, S. *et al.* Evaluating the Causal Relation of ApoA-IV with Disease-Related Traits - A Bidirectional Two-sample Mendelian Randomization Study. *Sci. Rep.* **7,** 8734 (2017).

79. Au Yeung, S. L., Lam, H. S. H. S. & Schooling, C. M. Vascular Endothelial Growth Factor and Ischemic Heart Disease Risk: A Mendelian Randomization Study. *J. Am. Heart Assoc.* **6,** e005619 (2017).

80. Lanktree, M. B., Thériault, S., Walsh, M. & Paré, G. HDL Cholesterol, LDL Cholesterol, and Triglycerides as Risk Factors for CKD: A Mendelian Randomization Study. *Am. J. Kidney Dis.* **71,** 166–172 (2018).

81. Pierce, B. L. & Burgess, S. Efficient design for mendelian randomization studies: Subsample and 2-sample instrumental variable estimators. *Am. J. Epidemiol.* **178,** 1177–1184 (2013).

82. Burgess, S. & Thompson, S. G. Bias in causal estimates from Mendelian randomization studies with weak instruments. *Stat. Med.* **30,** 1312–1323 (2011).

83. Inoue, A. & Solon, G. Two-Sample Instrumental variables estimators. *Rev. Econ. Stat.* **92,** 557–561 (2010).

84. Sankararaman, S., Sridhar, S., Kimmel, G. & Halperin, E. Estimating local

ancestry in admixed populations. *Am. J. Hum. Genet.* **82,** 290–303 (2008).

85. Shriner, D. Overview of Admixture Mapping. *Curr. Protoc. Hum. Genet.* **94,** 1.23.1-1.23.8 (2017).

86. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *Am. J. Hum. Genet.* **93,** 278–288 (2013).

87. Zhu, X. *et al.* Admixture mapping for hypertension loci with genome-scan markers. *Nat. Genet.* **37,** 177–181 (2005).

88. Freedman, M. L. *et al.* Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 14068–14073 (2006).

89. Kopp, J. B. *et al.* MYH9 is a major-effect risk gene for focal segmental glomerulosclerosis. *Nat. Genet.* **40,** 1175–1184 (2008).

90. Shriner, D. *et al.* Phenotypic variance explained by local ancestry in admixed African Americans. *Front. Genet.* **6,** 1–8 (2015).

91. Gijsberts, C. M. *et al.* Biomarkers of Coronary Artery Disease Differ Between Asians and Caucasians in the General Population. *Glob. Heart* **10,** 301–311.e11 (2015).

92. Gerstein, H. C., Yusuf, S., Riddle, M. C., Ryden, L. & Bosch, J. Rationale, design, and baseline characteristics for a large international trial of cardiovascular disease prevention in people with dysglycemia: The ORIGIN Trial (Outcome Reduction with an Initial Glargine Intervention). *Am. Heart J.* **155,** 26. e1-26. e13 (2008).

93. Bosch, J. *et al.* n-3 fatty acids and cardiovascular outcomes in patients with dysglycemia. *N. Engl. J. Med.* **367,** 309–18 (2012).

94. Origin, T. *et al.* Basal insulin and cardiovascular and other outcomes in dysglycemia. *N. Engl. J. Med.* **367,** 319–328 (2012).

95. Gerstein, H. C. *et al.* Identifying novel biomarkers for cardiovascular events or death in people with dysglycemia. *Circulation* **132,** 2297–2304 (2015).

96. Laurie, C. C. *et al.* Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet. Epidemiol.* **34,** 591–602 (2010).

97. Turner, S. *et al.* Quality control procedures for genome wide association studies. *Curr. Proc. Hum. Genet.* **68,** 1–24 (2011).

98. Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461,** 272–276 (2009).

99. Weale, M. E. in *Methods in molecular biology (Clifton, N.J.)* **628,** 341–372 (2010).

100. Teo, Y. Y. Common statistical issues in genome-wide association studies: a review on power, data quality control, genotype calling and population structure. *Curr. Opin. Lipidol.* **19,** 133–143 (2008).

101. 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).

102. 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).

103. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5,** e1000529 (2009).

104. Abdi, H. & Williams, L. J. Principal component analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **2,** 433–459 (2010).

105. Reich, D., Price, A. L. & Patterson, N. Principal component analysis of genetic data. *Nat. Genet.* **40,** 491–492 (2008).

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

107. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2,** 2074–2093 (2006).

108. Cooper, R. S., Tayo, B. & Zhu, X. Genome-wide association studies: implications for multiethnic samples. *Hum. Mol. Genet.* **17,** R151-5 (2008).

109. International HapMap Consortium. The International HapMap Project. *Nature* **426,** 789–96 (2003).

110. Relethford, J. H. Hardy–Weinberg Equilibrium. *Hum. Popul. Genet.* 23–48

111. Cox, D. G. & Kraft, P. Quantification of the power of Hardy-Weinberg equilibrium testing to detect genotyping error. *Hum. Hered.* **61,** 10–14 (2006).

112. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456,** 98–101 (2008).

113. Tian, C. *et al.* Analysis and application of European genetic substructure using 300 K SNP information. *PLoS Genet.* **4,** e4 (2008).

114. Siva, N. 1000 Genomes project. *Nat. Biotechnol.* **26,** 256 (2008).

115. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3 (Bethesda).* **1,** 457–70 (2011).

2 GENERAL HYPOTHESIS, OBJECTIVE, & APPROACH

2.1 General Hypothesis

We hypothesize that analysis of the three-way relationship between genetic variants, biomarkers and clinical events will identify novel biological pathways involved in cardiovascular disease and other health outcomes and posit that genetic analysis of an admixture population will provide new insights into the involvement of ancestry in health and disease.

2.2 General Objective

The overall objective of this PhD thesis is to identify novel biomarkers for cardiovascular disease and determine the impact of ancestry on biomarkers, in general.

2.3 Rationale and Approach

By combining information on genetic markers, plasma biomarkers, and clinical events there is an unprecedented opportunity to identify both novel and causal determinants of cardiovascular outcomes. Furthermore, there is opportunity to identify biomarker interactions and pathways given the extensive number of biomarkers analyzed. Currently, there is a relatively poor understanding of the biological role of many biomarkers, and there is much insight to be gained by studying biomarkers in parallel rather than individually. The proposed PhD project will therefore investigate the relationship between genetic variants, plasma biomarkers and clinical events using the ORIGIN data. We will use Mendelian randomization to identify causal mediators of disease (Chapter 3 and 4). Significant findings will be further investigated to understand the biology underlying the identified relationships and will also be verified where possible using independent cohorts and epidemiological analyses. Finally, we will employ admixture mapping

techniques to identify the role ancestry plays in determining biomarker levels and investigate the clinical implications of these findings (Chapter 5).

3 IDENTIFICATION OF BLOOD CSF1 AND CXCL12 AS CAUSAL MEDIATORS OF CORONARY ARTERY DISEASE USING MENDELIAN RANDOMIZATION IN THE ORIGIN TRIAL

Jennifer Sjaarda BSci [1,2,3], Hertzel Gerstein MD MSc [1,4], Michael Chong MSc [1,5], Salim Yusuf DPhil [1], David Meyre PhD [4,5], Sonia S. Anand MD PhD [1,4], Sibylle Hess PhD [6], Guillaume Paré MD MSc [1,2,3,4,5]

[1] Population Health Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, 237 Barton Street East, Hamilton, ON L8L 2X2, Canada

[2] Thrombosis and Atherosclerosis Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, 237 Barton Street East, Hamilton, ON L8L 2X2, Canada

[3] Department of Pathology and Molecular Medicine, McMaster University, Michael G. DeGroote School of Medicine, 1280 Main Street West, Hamilton ON L8S 4K1, Canada

[4] Department of Clinical Epidemiology & Biostatistics, McMaster University, 1280 Main Street West, Hamilton ON L8S 4K1, Canada

[5] Department of Biochemistry, McMaster University, 1280 Main Street West, Hamilton ON L8S 4K1, Canada

[6] Sanofi Aventis Deutschland GmbH, Research and Development Division, Translational Medicine and Early Development, Biomarkers and Clinical Bioanalyses, Frankfurt 65926, Germany

3.1 Forward

Identification of causal markers of coronary artery disease (CAD) has led to tremendous advances in both prevention and treatment. While epidemiological studies have identified numerous biomarkers associated with CAD, these associations are limited by reverse causation and confounding such that it is often impossible to distinguish true causal biomarkers. Mendelian randomization (MR) is able to overcome these challenges by taking advantage of the random allocation of alleles from parents to offspring and is often used to confirm causality of known biomarkers. We adopted an innovative approach to identify novel, causal biomarkers of CAD by screening a comprehensive panel of 237 biomarkers in 4,197 ORIGIN participants using MR.

Our MR analysis identified six biomarkers to be associated with CAD. While four of these biomarkers have been previously linked to CAD, we provide the first report establishing blood colony stimulating factor 1 (CSF1) as a causal mediator of CAD and the first MR analysis of stromal cell derived factor 1 (CXCL12) as a novel biomarker for CAD. The MR results were corroborated through epidemiological association of CSF1 and CXCL12 levels with prospective MACE in ORIGIN (n=8,197), and analysis in the large UKBiobank cohort (n=343,735), whereby genetically elevated CSF1 and CXCL12 were both associated with increased risk of CAD. Both biomarkers are linked to inflammatory processes characteristic of atherosclerosis and consistent with previous reports, including results from the CANTOS study, showing that an intervention aimed at decreasing inflammation through interleukin-1 beta (IL-1β) inhibition can lead to lower rates of recurrent cardiovascular events. Notably, IL-1β is known to be an important upregulator of CSF1 levels, which we also confirmed using MR. Finally, we also showed a causal effect of CSF1 on CRP levels, again consistent with results of the CANTOS trial. Together these results suggest that canakinumab may work in part by reducing CSF1 levels.

This manuscript has been recently published to the *Journal of American College of Cardiology* was published in July 2018, Volume 72, Issue 3 (PMID: 30012324). Hertzel Gerstein and Guillaume Paré conceptualized and designed the study. Jennifer Sjaarda designed the analysis plan, conducted all statistical analysis, and wrote the manuscript. All authors contributed to the interpretation of findings and to the critical reading and revision of the manuscript.

3.2 Abstract

BACKGROUND: Identification of biomarkers that cause coronary artery disease (CAD) has led to important advances in prevention and treatment. Epidemiologic analyses have identified many biomarker-CAD relationships, however, these associations may arise from reverse causation and/or confounding and therefore may not represent true causal associations. Mendelian randomization (MR) analyses overcome these limitations.

OBJECTIVE: We sought to identify causal mediators of CAD through a comprehensive screen of 237 biomarkers using MR.

METHODS: MR was performed by identifying genetic determinants of 227 biomarkers in ORIGIN (Outcome Reduction with Initial Glargine Intervention) trial participants (N=4,147) and combining these with genetic effects on CAD from the CARDIoGRAM consortium (60,801 cases and 123,504 controls). Blood concentrations of novel biomarkers identified by MR were then tested for association with incident major adverse cardiovascular events (MACE) in ORIGIN.

RESULTS: Six biomarkers were found to be causally linked to CAD after adjustment for multiple hypothesis testing. The causal role of four of these is well documented, whereas macrophage colony-stimulating factor 1 (CSF1) and stromal cell-derived factor 1 (CXCL12) have not previously been reported, to the best of our knowledge. MR analysis predicted an 18% higher risk of CAD per SD increase in CSF1 (OR=1.18, 95% CI 1.08 to 1.30; p=2.1x10-4) and epidemiologic analysis identified a 16% higher risk of MACE per SD (HR=1.16, 95% CI 1.09 to 1.23; p<0.001). Elevated CXCL12 levels were also identified as a causal risk factor for CAD with consistent epidemiological results. Furthermore, genetically predicted CSF1 and CXCL12 levels were associated with CAD in the UK Biobank (n=343,735).

CONCLUSIONS: We identified CSF1 and CXCL12 as causal mediators of CAD in humans. Understanding the mechanism by which these markers mediate CAD will provide novel insights into CAD and could lead to new approaches to prevention. Our results support targeting inflammatory processes and macrophages, in particular, to prevent CAD, consistent with the recent CANTOS (Canakinumab Antiinflammatory Thrombosis Outcome Study) trial.

3.3 Condensed Abstract

Identification of causal markers of CAD has led to tremendous advances in both prevention and treatment. Using Mendelian randomization analyses, we identified CSF1 and CXCL12 as two new, causal biomarkers of CAD. These markers point to inflammation and macrophages, in particular, as important actors for CAD. Our results are consistent with previous reports, including results from the CANTOS study, showing that an intervention aimed at decreasing inflammation leads to lower rates of recurrent cardiovascular events. Our study supports a role for CXCL12 and CSF1 for both risk stratification and as therapeutic targets.

FUNDING: Sanofi, Canadian Institutes of Health Research.

KEYWORDS: Mendelian randomization; biomarker; coronary artery disease; genetics; CSF1; CXCL12

clinicaltrials.gov: NCT00069784

3.4 Introduction

Identification of the biological mediators of a disease can both increase our understanding of its pathogenesis and suggest novel ways to prevent and/or treat it. For example, recognition that both blood pressure and LDL-cholesterol levels are causally related to coronary artery disease (CAD) has led to major advances in prevention.(1, 2) Epidemiological studies have identified numerous biomarkers associated with CAD. However, even strong associations obtained from observational studies are not guaranteed to be causally related to CAD because of the risk of reverse-causation and confounding.(3)

Mendelian randomization is a powerful genetic methodology that can be used to identify causal risk factors for CAD.(4) It is based on the principle that genetic variants are inherited randomly and independently of other risk factors for disease. If the levels of a particular risk factor (e.g. a biomarker) are affected by the presence of a genetic variant and if that variant also affects the incidence of a disease, the variant may be causing the disease by modulating the levels of the risk factor. The random distribution of the genetic variant at birth minimizes the possibility of confounding or reverse causation as explanations for the link between the biomarker and disease in the same way that the random allocation of a therapy in a randomized controlled trial minimizes this possibility.(5) These principles have allowed MR techniques to confirm LDL-cholesterol, interleukin-6 receptor, and lipoprotein(a) as causal biomarkers of CAD.(6–8)

To date, MR methodology has been used to determine whether or not a particular biomarker is causally related to a clinical outcome. However, when combined with a large panel of biomarkers measured in a prospective study which accrued many clinical outcomes, it can also be used to identify new, unsuspected but causally related biomarkers. We therefore sought to identify such novel mediators of CAD by applying MR techniques to a comprehensive panel of 227 serum biomarkers covering cardiovascular, metabolic and inflammatory processes within the recently

completed Outcomes Reduction with an Initial Glargine Intervention (ORIGIN) trial.(9)

3.5 Methods

3.5.1 Study Group - ORIGIN

The design and findings of the ORIGIN trial have been described in detail. Briefly, 12,537 people with established cardiovascular risk factors who also had diabetes, impaired glucose tolerance, or impaired fasting glucose were studied. After random allocation to 2 therapies using a factorial design (basal insulin glargine versus standard care and omega 3 fatty acid supplements versus placebo) they were followed for a median of 6.2 years for cardiovascular events and other health outcomes. The ethics committee at each participating site approved the trial, and all participants provided written informed consent. As previously described(10) biomarker levels were analyzed in the serum of 8,401 people that was drawn at the beginning of the study. The analysis was done using a customized human discovery multi-analyte profile (MAP) on the Luminex 100/200 platform and the biomarkers were selected based on their implication in physiologic processes related to cardiovascular diseases.

A subset of 5,078 ORIGIN individuals who consented to genetic analyses were genotyped on Illumina's HumanCore Exome chip. Standard quality control measures were assessed. After quality control, the sample consisted of 4,147 participants and 284,024 SNPs from two ethnic groups (European and Latin American). Imputation was then performed on the post QC data through to predict unobserved genotypes in the study group. Over 30 million SNPs were imputed (using 1000Genomes data), allowing for comprehensive coverage of known genetic variants (a detailed description of quality control and imputation procedures is in the supplement). The summary level data, analytic methods, and study materials have been made available to other researchers for purposes of

reproducing the results or replicating the procedure. Key clinical characteristics of the study populations are shown in Supplementary Tables 1 to 3.

3.5.2 CARDIoGRAM Consortium Data

Genetic data on coronary artery disease and myocardial infarction (MI) was obtained and contributed by CARDIoGRAMplusC4D investigators and have been downloaded from www.CARDIOGRAMPLUS -C4D.org. Specifically, we used the most recent CARDIoGRAM genome-wide association data (released October 2015), a meta-analysis of 48 genome-wide association studies with 60,801 cases and 123,504 controls, primarily of European descent (approximately 77%).(11) CAD outcome was defined by an inclusive CAD diagnosis, including MI, coronary stenosis >50%, chronic stable angina or acute coronary syndrome.(11)

3.5.3 Statistical Analysis

SNP association with biomarkers and CAD

The analysis was restricted to biomarkers that are directly encoded by a gene(s) on the autosomal chromosome (i.e. chromosome 1-22). Five biomarkers were removed because they are products of genes on the X chromosome. An additional five biomarkers from our panel were hormones which are not a direct gene product (e.g. cortisol) and were also removed, leaving 227 biomarkers for analysis. SNP selection was then carried out in four steps. First, for each of our 227 biomarkers, we restricted our analysis to SNPs within 300 Kb of the gene(s) encoding the corresponding protein (or a protein component), thereafter referred to as *cis* associations (Supplementary Figure 1). The corresponding gene names for each biomarker in the ORIGIN panel were identified (a list of all biomarkers and their genes is found in Supplementary Table 4) and subsequently used to filter to only *cis* genotypes (a detailed description of the process is found in the supplement). After filtering, there were 1,067,955 SNP/biomarker *cis* pairs.

Second, association of *cis* SNPs with biomarker and CAD were determined. After removing SNPs not found in the CARDIoGRAM database or with minor allele frequency below 0.05, 397,140 SNPs were tested for association with their corresponding *cis* biomarker in ORIGIN. Linear regression of each SNP with each biomarker was performed, with biomarker concentration as the dependent variable. The regression models were first computed in each ethnic group separately, adjusting for age, sex, and the first five principal components, using SNPtest.(12) The ethnic specific models were meta-analyzed across the two ethnicities using fixed effects models to minimize the risk of confounding caused by population stratification.

Third, *cis* SNPs with biomarker association $p<0.01$ were selected. Finally, SNPs were pruned for linkage disequilibrium at a stringent threshold of $r^2<0.1$ using the 1000 Genomes data (Europeans) to ensure associations retained for Mendelian randomization analysis were non-redundant (using PLINK(13)). SNPs were selectively prioritized based on the significance of the association with their biomarker. For each biomarker, the *cis* SNP with the most significant association with the biomarker was first retained and all SNPs in linkage ($r^2>0.1$) with that SNP removed. This process was then repeated for any remaining SNPs. 1,880 *cis* SNP/biomarker associations remained after pruning (see Supplementary Figure 2 for schematic of SNP and biomarker selection).

Identification of blood mediators of CAD using Mendelian randomization

A two sample MR was performed on 205 of the remaining 227 biomarkers, retaining only those biomarkers which had at least one significant *cis* SNP ($p<0.01$) and was also found in the CARDIoGRAM data. This threshold of $p<0.01$ for SNP associations with biomarkers was chosen in an effort to include all possible valid instruments. According to our two-sample MR, a less stringent threshold can be applied without increasing the risk of type-1 error.(14) In other words, this threshold was selected to balance potential false-positive SNP-biomarker associations which will bias our results towards the null in a two-sample MR design,

versus including as much genetic variance as possible, which will increase power. For each biomarker, we used (1) the effect of the SNPs on their *cis* biomarker (calculated in ORIGIN), and (2) the effect of the SNPs on CAD from the CARDIoGRAM consortium as input variables for the MR analysis (Figure 3-1, Central Figure). MR associations were performed using the inverse-variance weighted (IVW) fixed effects method by regressing genetic effect estimates for CAD (dependent variable) on genetic effect estimates of biomarkers. To determine significance, a bootstrap method was used under the null hypothesis that the ratio of genetic effect estimates for CAD on genetic effect estimates for biomarkers is 0 for all SNPs. Predicted effects on CAD were sampled from a normal distribution with mean and standard deviations as determined from CARDIoGRAM. A two-tailed p-value was calculated using a z-test from 100,000 random simulations. Biomarkers were deemed significant after adjusting for multiple testing hypothesis (p<0.05/205). The MR association with MI was also assessed using MI specific estimates from CARDIoGRAM.

Replication of MR findings in the UKBiobank

Novel MR findings were replicated using a weighted genetic risk score (GRS) in the UKBiobank (UKB)(15) using 343,735 unrelated British individuals. Specifically, the allelic dosage at each variant site was weighted by the predicted change in biomarker level conferred per additional effect allele. Subsequently, the weighted contribution at each variant site was summed to create an overall score. Using these biomarker's GRS, the association between CAD and genetically elevated biomarkers was tested using a logistic regression model with age and sex as covariates (see supplementary material for more details).

Association of biomarker levels with MACE in ORIGIN

Once significant biomarkers were identified by the MR analysis, we tested whether biomarker levels showed a consistent association with major adverse cardiovascular events (MACE) in ORIGIN, defined as nonfatal MI, nonfatal stroke

or cardiovascular death. This was assessed using Cox proportional-hazards models in all ORIGIN participants with biomarker data (i.e. individuals with and without genetic data were included). Incident MACE was used as the dependent variable and biomarker concentration as the independent variable of interest. Models were adjusted for age, sex and ethnicity to remain consistent with MR model adjustment. We also tested models further adjusting for traditional CAD risk factors as a sensitivity analysis, namely prior type 2 diabetes, body mass index (BMI), serum cystatin-c, current smoker, diagnosis of hypertension, and LDL. We removed individuals without specific ethnicity information resulting in a final sample of N=8,197. Subgroup analyses were performed to test heterogeneity between groups using logistic models adjusted for age, sex and ethnicity (where appropriate).

Mendelian randomization association of biomarkers with other endpoints

To explore the relationship between the biomarkers that were identified and established clinical risk factors for CAD we conducted additional MR analyses. Using publicly available consortia (Supplementary Table 5) we obtained genetic estimates for 11 CAD risk factors including: BMI, chronic kidney disease (CKD), diastolic blood pressure (DBP), systolic blood pressure (SBP), type 2 diabetes, fasting glucose, glycated hemoglobin (HbA_{1c}), HDL-cholesterol, LDL-cholesterol, and triglycerides. MR associations were obtained using the same method as for CAD (described above). Additionally, CSF1 and CXCL12 SNPs were tested for their effect on the 236 remaining biomarkers using MR. In other words, for both CSF1 and CXCL12, 236 MR models were used to assess their causal effect on the 236 remaining ORIGIN biomarkers. Specifically, the input variables for each MR were (1) the effect of SNPs on their cis biomarker (CSF1 and CXCL12) as the independent variable and (2) the effect of the same set of SNPs on an additional biomarker as the dependent variable. Statistical analyses were performed using R (version 3.0.1), unless stated otherwise.

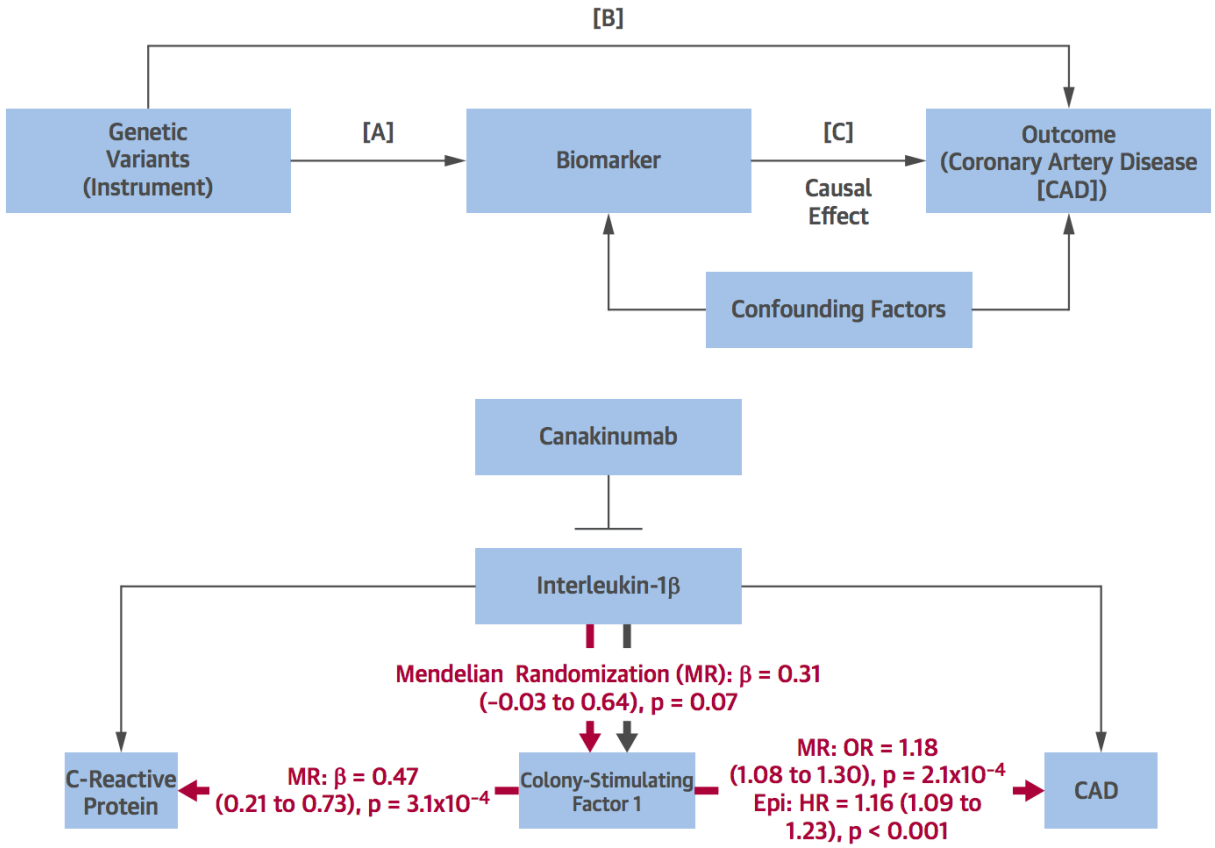


Figure 3-1: Central Figure - CSF1 and CXCL12 as Causal Mediators of CAD Using MR.

(Top) Mendelian randomization (MR) design. This study used MR to perform a comprehensive screen of 237 blood biomarkers for a causal effect on coronary artery disease (CAD). MR takes advantage of the random allocation of alleles from parents to offspring to inform on the causality of exposures on disease by using genetic variants as instrumental variables. It is robust to reverse causation and confounding, which often affects epidemiological associations. Specifically, the associations of [A] and [B] were used to estimate the causal effect of each biomarker coronary artery disease, represented by association [C]. (Bottom) Proposed mechanism for the beneficial effect of canakinumab on CAD. Using MR, our study identified blood colony-stimulating factor 1 (CSF1) and stromal cell–derived factor 1 (CXCL12) as new, causal mediators of CAD, findings that were corroborated through epidemiological association of CSF1 and stromal cell–derived factor 1 levels with prospective cardiovascular events in ORIGIN (Outcome Reduction With Initial Glargine Intervention). Results from the recent CANTOS (Canakinumab Antiinflammatory Thrombosis Outcome Study) trial have shown interleukin-1 beta inhibition with canakinumab leads to a reduction in both

cardiovascular events and C-reactive protein levels. In light of these findings, we confirmed the known up-regulating effect of interleukin-1 beta on CSF1 levels by using MR. Subsequent MR analyses also revealed CSF1 to be an up-regulator of C-reactive protein levels, again consistent with trial results. Taken together, these results point to CSF1 as a potential mediator of canakinumab's beneficial effect. The figure shows a tentative mechanism for the beneficial effect of canakinumab, mediated in part through CSF1, where the red arrows represent novel associations and the black lines represent known relationships. Epi = epidemiological association.

3.6 Results

3.6.1 Identification of CAD biomarkers using Mendelian randomization

Two-hundred twenty-seven serum biomarkers were tested for an association with CAD using a Mendelian randomization approach. After removing biomarkers without any significant *cis* SNP association, 205 biomarkers were retained for downstream analysis. In other words, for each of these 205 biomarkers there was at least one SNP within 300 Kb of the gene(s) encoding that same biomarker that was significantly associated (p<0.01) with biomarker levels. When the relationship between each of these SNPs and CAD was assessed using the CARDIoGRAM database, six biomarkers were significantly associated with CAD after Bonferroni correction for multiple hypothesis testing (p<0.05/205) (Table 3-1). Four of these six (lipoprotein(a), apolipoprotein E, apolipoprotein C3 and interleukin-6 receptor) have been previously linked to CAD in many studies (6, 8, 16, 17), consistent with our findings. However, to the best of our knowledge, we report the first CAD MR study performed on stromal cell-derived factor 1 (CXCL12) and identified macrophage colony-stimulating factor (CSF1) as a novel mediator of CAD. As noted in Figure 3-2 and Supplementary Figure 3, the MR analysis suggested a deleterious effect of CSF1 (OR=1.18 per SD; 95% CI 1.08 to 1.30; p=2.1x10^{-4}) and CXCL12 (OR=1.69 per SD; 95% CI 1.40 to 2.05; p=6.2x10^{-8}) on CAD. Consistent estimates were also obtained using the IVW fixed-effects model using the 'MR-base' package in R (see supplementary material). The CSF1 and CXCL12 MR

models included seven and two independent SNPs as instrumental variables (IVs) (p<0.01 and FDR<0.1), respectively. Although nominal to weak associations with each individual SNP were observed in CARDIoGRAM (see Supplementary Tables 6 and 7), by pooling multiple, independent signals together in an IVW-MR design, significant estimates were obtained analogous to the pooling of effects across studies in a meta-analysis. MR associations of CSF1 and CXCL12 with MI were consistent with CAD estimates (Table 3-2). Regional plots of SNP associations with serum CSF1 and CXCL12 at the *CSF1* and *CXCL12* loci, respectively, are depicted in Supplementary Figure 4. Results remained significant when tested in each ethnic group separately and at a MAF threshold of 0.01 (see supplementary material).

Table 3-1: Summary of top Mendelian Randomization results with CAD (p<0.05/205).

Biomarker	Number of SNPs	OR (95% CI)	P-value
Lipoprotein(a) (LPA)	18	1.22 (1.20, 1.25)	<1.00E-50
Apolipoprotein E(APOE)	24	0.86 (0.84, 0.88)	1.92E-32
Interleukin-6 receptor (IL6R)	29	0.94 (0.92, 0.95)	5.83E-18
Stromal cell-derived factor 1 (CXCL12)	2	1.69 (1.40, 2.05)	6.16E-08
Apolipoprotein C3 (APOC3)	6	1.17 (1.08, 1.26)	9.35E-05
Macrophage colony-stimulating factor 1 (CSF1)	7	1.18 (1.08, 1.30)	2.07E-04

OR per 1 SD increase in biomarker.

Table 3-2: MR Association of CSF1 and CXCL12 serum levels with CAD risk factors and related endpoints.

Variable	CSF1 (per SD)		CXCL12 (per SD)	
	OR (95% CI)	**P-value**	**OR (95% CI)**	**P-value**
MI (yes/no)	1.21 (1.10, 1.34)	0.0001	1.59 (1.29, 1.95)	<0.0001
CKD (yes/no)	1.05 (0.89, 1.24)	0.56	0.72 (0.55, 0.96)	0.02
Diabetes (yes/no)	1.08 (0.96, 1.21)	0.21	1.00 (0.80, 1.26)	0.99
	β (95% CI)	**P-value**	**β (95% CI)**	**P-value**
Fasting Glucose (mmol/L)	0.04 (-0.001, 0.08)	0.06	-0.005 (-0.07, 0.06)	0.88
HbA1c (%)	0.002 (-0.03, 0.04)	0.90	0.01 (-0.06, 0.07)	0.80
HDL-cholesterol (SD)	-0.07 (-0.12, -0.02)	0.007	0.02 (-0.06, 0.12)	0.58
LDL-cholesterol (SD)	0.03 (-0.03, 0.08)	0.35	-0.04 (-0.14, 0.05)	0.31
Triglycerides (SD)	0.01 (-0.04, 0.06)	0.78	0.01 (-0.07, 0.10)	0.73
SBP (mmHg)	-0.08 (-1.11, 0.96)	0.88	-1.65 (-3.47, 0.16)	0.07
DBP (mmHg)	-0.18 (-0.83, 0.48)	0.60	-1.72 (-2.88, -0.57)	0.003
BMI (kg/m^2)	0.02 (-0.02, 0.06)	0.24	-0.001 (-0.07, 0.07)	0.98

OR per 1 SD increase in biomarker.

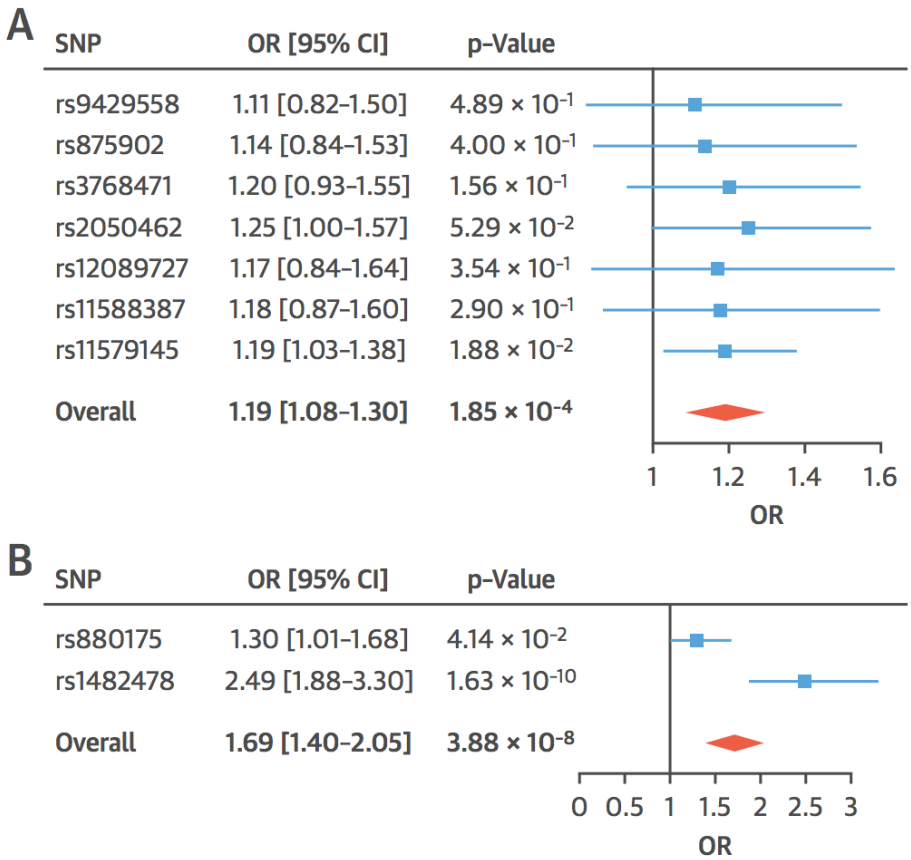


Figure 3-2: Association of CSF1 and CXCL12 with risk of CAD using MR.

Forest plots depict a summary of the MR results for (A) CSF1 (A) and (B) CXCL12. A single SNP MR was conducted for each independent SNP (pairwise $r^2 < 0.1$). ORs were determined by the Wald method by regressing the effect estimates from the CAD association (from CARDIoGRAM) on the biomarker association (from ORIGIN). A two-tailed p-value was calculated using a z-test from 100,000 random simulations. The two SNPs in (B) were significantly different ($p<0.05$).

An important assumption in MR studies is that the genetic variant does not affect the outcome other than through its influence on the exposure. Therefore, to assess for the presence of unmeasured horizontal pleiotropy (i.e. effects of the genetic variants beyond those on the biomarkers of interest), we utilized the MR-Egger(18) method where the y-intercept is allowed to float, rather than be fixed at zero. This method is only suitable when three or more SNPs are used as IVs and was therefore only applied to the CSF1 MR. We found no evidence of pleiotropy as determined by the significance of the y-intercept ($p>0.05$). As a sensitivity analysis, we used a leave-one-out strategy in which MR analyses were repeated, excluding one variant at a time. This technique may also only be applied to MRs with three or more SNPs. When applied to CSF1 data, consistent estimates were obtained for each SNP excluded (see Supplementary Table 8). To further explore the effect of these SNPs on their respective biomarkers, we tested them for expression quantitative trait loci in the GTEx dataset (www.gtexportal.org)(19). Notably, 2 of the 7 CSF1 SNPs included in the MR (rs9429558 and rs11579145) were significantly associated ($p<0.05$) with CSF1 expression in the aorta, with consistent direction of effects. We did not identify any associations in other relevant tissues.

3.6.2 Validation of MR findings using UKBiobank

All CSF1 and CXCL12 variants used in the original MR analyses were also imputed at high quality in the UKB cohort with average INFO score (as defined by IMPUTE2, where values near 1 indicate that a SNP was imputed with high certainty) 0.97 and 0.98 for CSF1 and CXCL12 variants, respectively. No imputed variants had an INFO score below 0.9. MR results were corroborated using genetically predicted biomarker levels in the UKB with estimates obtained from ORIGIN. Genetically

elevated CSF1 and CXCL12 were both associated with an increased risk of CAD in UKB (CXCL12: 1.41 per SD, 1.13 to 1.76, p=0.002; CSF1: 1.12 per SD, 1.03 to 1.22, p=0.01), consistent with MR findings using ORIGIN and CARDIoGRAM estimates. Consistent estimates were obtained after removing 17,203 individuals with diabetes (CXCL12: 1.36 per SD, 1.11 to 1.79, p=0.004; CSF1: 1.09 per SD, 0.99 to 1.19, p=0.08).

3.6.3 Association of CSF1 and CXCL12 concentration with MACE in ORIGIN

The MR-generated hypothesis that these biomarkers promoted CAD was then tested using the ORIGIN data. We therefore assessed the epidemiological relationship of baseline CSF1 and CXCL12 with MACE. We found that increased levels of blood CSF1 and CXCL12 were significantly associated with an increased risk of incident MACE in models adjusting for age, sex, ethnicity (CSF1: hazard ratio (HR)=1.30 per SD; 95% CI, 1.23 to 1.37; p<0.0001 and CXCL12: HR=1.15 per SD; 95% CI, 1.08 to 1.21; p<0.0001). Consistent results were also observed in models fully adjusted for CAD risk factors provided in the supplementary material. To understand potential thresholds for risk stratification, we performed receiver operating characteristic (ROC) curves and determined the optimal threshold for each biomarker. Our analysis revealed optimal thresholds of 0.72 ng/mL and 3.45 ng/mL for CSF1 and CXCL12, respectively (Figure 3-3 and Supplementary Figure 5). Both CSF1 and CXCL12 improved discrimination of MACE in a model adjusting for the same co-variates included in the fully adjusted model (i.e. age, sex, ethnicity, prior diabetes, BMI, smoking status, hypertension and hypercholesterolemia) (CSF1: Net Reclassification Index (NRI)=0.235, p<0.0001; CXCL12: NRI=0.084, p=0.004). As a sensitivity analysis, we also tested for association after removing non-fatal stroke from the MACE composite to create a better proxy for coronary outcomes, with similar results for both CSF1 (HR=1.18 per SD; 95% CI, 1.11 to 1.26; p<0.0001) and CXCL12 (HR=1.10 per SD; 95% CI,

1.03 to 1.16; p=0.003). We tested for an interaction between the ORIGIN treatment arms and the effect of CSF1 and CXCL12 on MACE, and did not identify significant interaction for either biomarker with either treatment arm in a survival model adjusting for age, sex, and ethnicity ($p_{interaction}$ >0.05). Finally, we performed subgroup analyses to assess the consistency of the association of CSF1 and CXCL12 concentration with MACE. No significant difference among subgroups was observed after adjustment for multiple hypothesis testing (Figure 3-4).

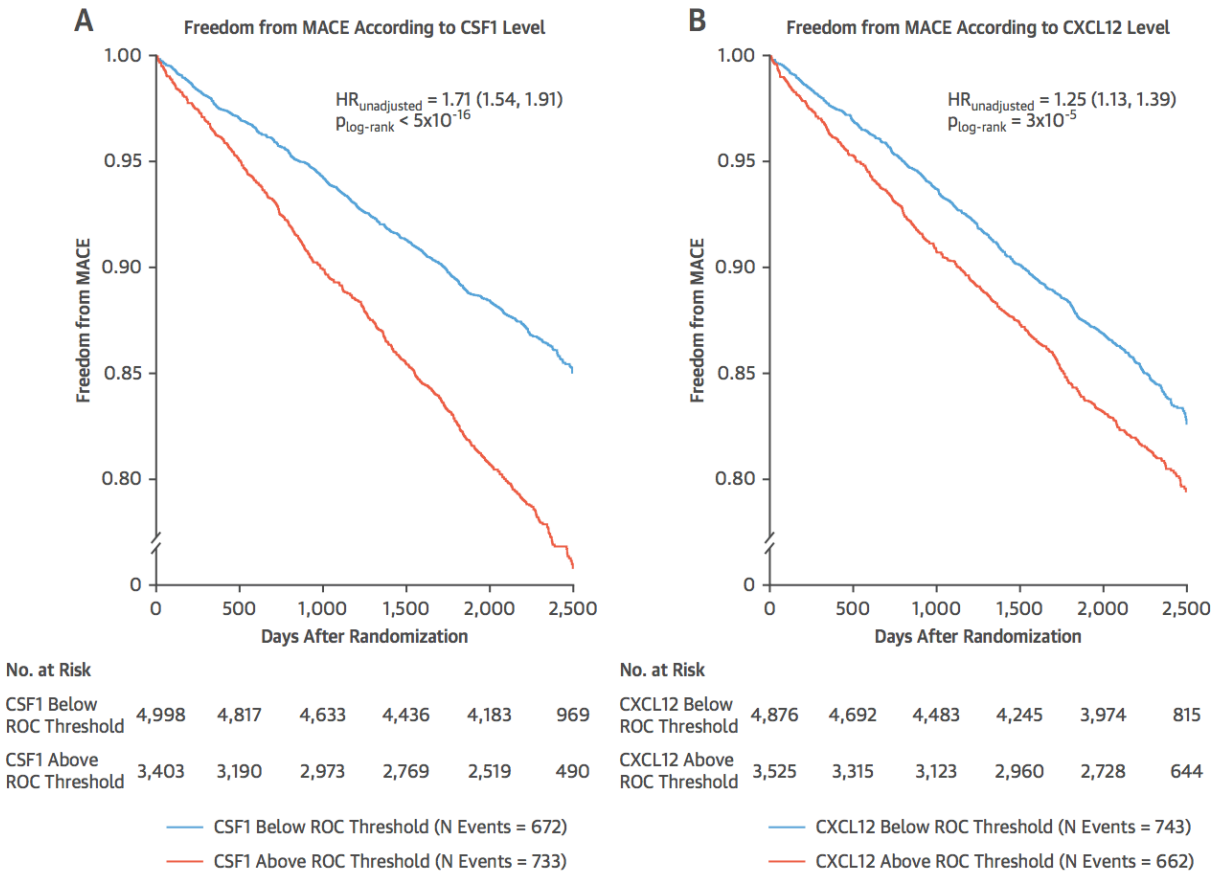


Figure 3-3: Kaplan-Meier curve for MACE-free survival according to CSF1 and CXCL12 levels.

Groups are defined as concentration above or below median biomarker level 0.72 ng/mL and 3.45 ng/mL for CSF1 and CXCL12, respectively. These thresholds were chosen to maximize discrimination based on ROC analysis. Number at risk for 500 day intervals are presented below each plot. Figure represents MACE survival

curves for CSF1 (panel A) and CXCL12 (panel B) using unadjusted models. HR (95% CI) indicates the unadjusted risk of biomarker level above threshold versus below. MACE was defined as nonfatal myocardial infarction, nonfatal stroke or cardiovascular death.

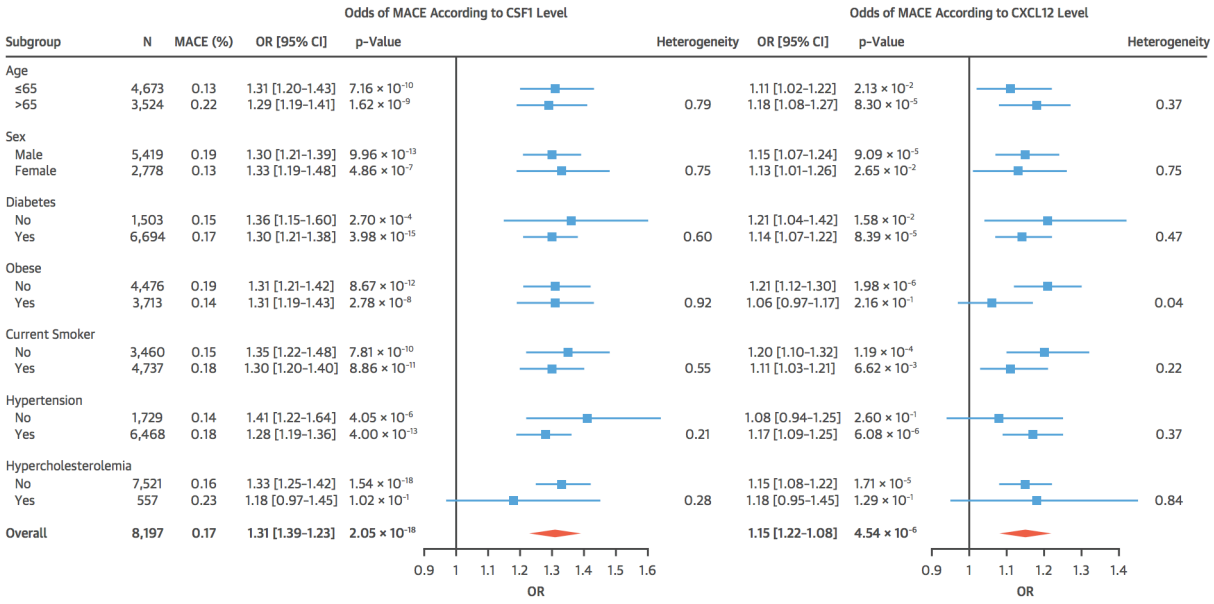


Figure 3-4: Subgroup analysis for association of CSF1 and CXCL12 levels with risk of MACE.

Models are adjusted (where appropriate) for age, sex and ethnicity. Subgroups were as follows: age (=<65, >65), sex (male, female), baseline diabetes status (yes/no), obese (BMI>=30, BMI<30), current smoker, hypertension diagnosis, and hypercholesterolemia (LDL-cholesterol >4.5 mmol/L, LDL-cholesterol=<4.5 mmol/L).

3.6.4 Association of CSF1 and CXCL12 with CAD risk factors and other biomarkers using Mendelian randomization

We assessed whether CSF1 and CXCL12 affect the concentration of other biomarkers or CAD risk factors using MR analysis. We identified an association between CSF1 and C-reactive protein (CRP), whereby increased CSF1 leads to

increased CRP levels (0.47 SD CRP per 1 SD CSF1; 95% CI, 0.21 to 0.73, p=0.0002). No other biomarkers were significant for CSF1 or CXCL12 in the MR analysis. We also tested the effect of CSF1 and CXCL12 on 11 cardiovascular risk factors and endpoints (Table 3-2). We found no significant associations after adjusting for multiple hypothesis testing (p<0.05/11). Because CSF1 is known to be upregulated by interleukin-1 beta (IL-1β)(20–23) and in light of the protective effect of IL-1β antagonist canakinumab on CAD in the CANTOS trial(24), we also tested for a causal effect of blood IL-1β on CSF1 and CXCL12 using MR. Using genetic variants at the *IL1B* locus as IVs in the MR, we identified a nominal association between IL-1β and CSF1 (0.31 SD CSF1 per 1 SD IL-1B; 95% CI, -0.03 to 0.64, p=0.07) suggestive of a causal role for IL-1β in regulating CSF1 levels. However, statistical power was limited as IL-1β levels were below the detection threshold in 93% of ORIGIN participants. We did not identify an association with CXCL12 (p=0.22).

3.7 Discussion

In the current report, we used Mendelian randomization to screen a comprehensive panel of 227 blood biomarkers to identify potential causal mediators of CAD. We identified blood macrophage colony-stimulating factor 1 and stromal cell-derived factor 1 as mediators of CAD, and confirmed a role for lipoprotein(a), interleukin-6 receptor, apolipoprotein C3 and apolipoprotein E. Increased serum CSF1 and CXCL12 were found to be associated with an increased risk of CAD and these findings were corroborated in UKB using genetically predicted biomarker levels against CAD events. We further showed a consistent association between serum concentrations of both novel biomarkers and MACE in ORIGIN, confirming their deleterious effects. Our results suggest increased CSF1 and CXCL12 concentration are causally related to cardiovascular events, paving the way for both risk stratification and therapeutic intervention in susceptible patients.

Macrophage colony-stimulating factor 1 is a cytokine and a hematopoietic growth factor which regulates macrophage survival, differentiation, proliferation and migration from precursor hematopoietic stem cells. Blood CSF1 is detectable under normal conditions and is expressed by several cell types, including endothelial cells, macrophages and smooth muscle cells.(25) Multiple studies have shown CSF1 to play a role in atherosclerosis formation and to be actively expressed in atherosclerosis lesions.(26, 27) Additionally, osteopetrotic (*op/op*) mutant mice lacking functional *CSF1* have been shown to be dramatically protected against atherosclerosis.(28) These models have also demonstrated CSF1 as being involved in monocyte recruitment and survival in atherosclerosis plaque(29). More recently, a large scale proteomic analysis identified CSF1 as a new risk marker for ischemic stroke in two large, independent studies(30), consistent with our findings in ORIGIN. Together with our results, these data suggest a causal, pro-inflammatory role for CSF1 in the development of CAD. Indeed, recent results from the CANTOS trial have identified inflammation as being an important and independent actor in cardiovascular disease. Specifically, canakinumab, an IL-1β inhibitor with anti-inflammatory effects, was associated with significantly lower rates of recurrent cardiovascular events(24). Remarkably, independent studies have shown IL-1β to be an important upregulator of CSF1 levels(20–23) which is consistent with our MR findings. Our MR analysis also points to CSF1 as being a causal regulator of CRP levels, consistent with this notion and the inhibitory effect of canakinumab on CRP. These results point to CSF1 as a tentative link between IL-1β and the protective effect of canakinumab and suggest that individuals with high CSF1 are at increased risk for CAD (Figure 3-1, Central Figure).

Stromal cell-derived factor 1 (CXCL12) is a chemokine which binds to the receptor encoded by CXCR4 and has a prominent role in leukocyte recruitment and hematopoietic stem cell functions.(31) CXCL12 has been thoroughly examined in the context of atherosclerosis and has been shown to be highly expressed by several cell types of importance in CAD such as smooth muscle cells and

endothelial cells of atherosclerotic plaques.(32, 33) However, current experimental data in mice is inconclusive regarding the biological effect of CXCL12 and its ligand (CXCR4) in atherosclerosis, suggesting a complex cell and context-specific role.(34) Genetic studies have identified the *CXCL12* locus as a novel region in CAD, with risk increasing alleles being also associated with increased *CXCL12* gene expression and levels.(35) The top *CXCL12* SNP previously associated with CAD (rs1746048) was not included in the MR analysis present here because it did not pass our significance threshold (p<0.01), although it was nominally associated with CXCL12 levels (p=0.03). However, we report two novel CXCL12 SNPs independent of those previously identified (r^2<0.1 for both SNPs), which show consistent effects with CAD according to our MR. Additionally, consistent associations have been reported between increased serum CXCL12 and incident myocardial infarction and CAD, and also recurrent events in patients with CAD.(36, 37) Taken in combination with our MR estimates, these data support the causal association of higher CXCL12 levels with development of CAD and MI.

There are limitations to our study which need to be taken into consideration. First, there may be issues of statistical power. Although we can be confident in the six associations found, we cannot rule out a causal role of other biomarkers with CAD. For example, we did not detect a significant association between ApoB and CAD after adjusting for multiple hypothesis testing in our MR analysis (p=0.029). This relationship has been seen and replicated in other, larger, MR studies.(38) One explanation for this observation could be due to our two sample MR study design, where genetic estimates were obtained from independent populations. In such a design, weak IVs (in this case weak associations of genetic variants with ApoB) results in estimates which are biased toward the null hypothesis, reducing the likelihood of type 1 error, and as a consequence, decreasing power.(39) Also, while our results point to a causal role of CSF1 and CXCL12 in CAD, the results do not specifically address in which tissue the effect may be mediated. Additionally, genetic pleiotropy is a major consideration and limitation to Mendelian

randomization studies. This phenomenon occurs when a genetic variant has other effects beyond its effect on the biomarker being studied. If present, interpretation of results can be difficult. We mitigated this source of bias by limiting our investigation to variants at or near the gene coding for the biomarker of interest. Furthermore, associated loci were individually inspected for proximity to other potential genes. We found no nearby genes to the *CSF1* and *CXCL12* loci which were plausible sources of pleiotropy and MR-Egger revealed no significant source of pleiotropic bias in the CSF1 MR. However, we were unable to test CXCL12 in the MR-Egger framework as only two IVs were utilized in the MR. Finally, it is worth noting that the MR analysis investigated CAD as the primary endpoint while our corroborated epidemiological associations were on prospective MACE. Although the estimates from the two models were directionally consistent, the magnitudes were different, specifically for CXCL12. This may be a result of confounding in the epidemiological association, difference in the study groups, or due to the differences in the two outcomes analyzed.

Identification of CAD risk factors is instrumental to further our understanding of the disease, evaluate its risk and guide treatment. Using Mendelian randomization, we have investigated a comprehensive panel of biomarkers for involvement in CAD. Our study presents the first MR analysis of CXCL12 and identifies CSF1 as a novel causal mediator of CAD, consistent with previous model systems and the inflammatory role of these biomarkers. Notably, the CANTOS trial has recently shown that an intervention aimed at decreasing inflammation leads to lower rates of recurrent cardiovascular events(24, 40). Indeed, our results suggest CSF1 mediates, at least in part, the beneficial effect of canakinumab on CAD identified in CANTOS. Increased serum CSF1 and CXCL12 concentrations represent independent mechanisms leading to CAD, which can be assessed through a simple blood test. Future research should be aimed at identifying the causal mechanisms and whether interventions targeted at reducing the CSF1 and CXL12 levels can reduce CAD.

3.8 Concluding Remarks

3.8.1 Competency in Medical Knowledge

Identification of causal markers of CAD has led to tremendous advances in both prevention and treatment, however epidemiological studies are limited by reverse causation and confounding such that it is often impossible to distinguish true causal biomarkers. Mendelian randomization, however, is immune to these limitations. A comprehensive investigation of 237 biomarkers identified CXCL12 and CSF1 as causal CAD markers using Mendelian randomization.

3.8.2 Translational Outlook

Our results suggest the clinical utility of CXCL12 and CSF1 for risk stratification and the development of interventions targeted to lower these biomarkers for prevention and reduction of CAD. Our data also point to CSF1 as a potential mediator of the protective effect of canakinumab on CAD.

3.8.3 Conflict of Interest

H.C.G. has received consulting fees from Sanofi, Novo Nordisk, Lilly, AstraZeneca, Boehringer Ingelheim, and GlaxoSmith-Kline and support for research or continuing education through his institution from Sanofi, Lilly, Takeda, Novo Nordisk, Boehringer Ingelheim, and AstraZeneca. G.P. has received consulting fees from Sanofi, Bristol Myers Squibb, Lexicomp, and Amgen and support for research through his institution from Sanofi. S.H. is an employee of Sanofi. S.Y. has received research support for ORIGIN from Sanofi through his institution. D.M., S.A., M.C. and J.S. report no conflicts.

3.8.4 Funding

The ORIGIN trial and biomarker project were supported by Sanofi and CIHR (award 125794). H.C.G., receiving support from Population Health Institute Chair

in Diabetes Research and Care; S.A., receiving support from Canada Research Chair in Ethnicity and Cardiovascular Disease, Michael G. DeGroote Chair in Population Health; D.M.; receiving support from Tier 2 Canada Research Chair in Genetics of Obesity; S.Y., receiving support from Heart and Stroke Foundation of Ontario/MarionW. Burke Chair in Cardiovascular Disease; G.P., receiving support from Canada Research Chair in Genetic and Molecular Epidemiology, CISCO Professorship in Integrated Health Systems.

3.8.5 Acknowledgements

The authors thank all the participants in the ORIGIN-trial, the CARDIoGRAMplusC4D consortia, and UK Biobank for making their data available, CIHR (award 125794) and Sanofi for supporting the project, and the support of Cisco Canada and Canada Research Chair.

3.9 References

1. Yusuf S, Sleight P, Pogue J, Bosch J, Davies R, Dagenais G. Effects of an angiotensin-converting-enzyme inhibitor, ramipril, on cardiovascular events in high-risk patients. The Heart Outcomes Prevention Evaluation Study Investigators. N. Engl. J. Med. 2000;342:145–53.

2. Baigent C, Keech A, Kearney PM, et al. Efficacy and safety of cholesterol-lowering treatment: prospective meta-analysis of data from 90,056 participants in 14 randomised trials of statins. Lancet 2005;366:1267–78.

3. Lawlor DA, Harbord RM, Sterne JAC, Timpson N, Smith GD, Davey Smith G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. Stat. Med. 2008;27:1133–1163.

4. Evans DM, Davey Smith G. Mendelian Randomization: New Applications in the Coming Age of Hypothesis-Free Causality. Annu. Rev. Genomics Hum. Genet. 2015:1–24.

5. Thanassoulis G, O'Donnell CJ. Mendelian randomization: nature's randomized trial in the post–genome era. JAMA 2009;301:2386–2388.

6. Clarke R, Peden JF, Hopewell JC, et al. Genetic variants associated with Lp(a) lipoprotein level and coronary disease. N. Engl. J. Med. 2009;361:2518–2528.

7. Teslovich, T. Musunuru, K. Smith a. E Al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature 2010;466:707–713.

8. Sarwar N, Butterworth AS. Interleukin-6 receptor pathways in coronary heart disease: A collaborative meta-analysis of 82 studies. Lancet 2012;379:1205–1213.

9. Gerstein HC, Yusuf S, Riddle MC, Ryden L, Bosch J. Rationale, design, and baseline characteristics for a large international trial of cardiovascular disease

prevention in people with dysglycemia: The ORIGIN Trial (Outcome Reduction with an Initial Glargine Intervention). Am. Heart J. 2008;155:26. e1-26. e13.

10. Gerstein HC, Paré G, McQueen MJ, et al. Identifying novel biomarkers for cardiovascular events or death in people with dysglycemia. Circulation 2015;132:2297–2304.

11. Nikpay M, Goel A, Won H-H, et al. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nat. Genet. 2015;47:1121–30.

12. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat. Rev. Genet. 2010;11:499–511.

13. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 2007;81:559–575.

14. Burgess S, Small DS, Thompson SG. A review of instrumental variable estimators for Mendelian randomization. Stat. Methods Med. Res. 2017;26:2333–2355.

15. Sudlow C, Gallacher J, Allen N, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Med. 2015;12:1–10.

16. Bennet A, Di Angelantonio E, Ye Z, et al. Association of apolipoprotein E genotypes with lipid levels and coronary risk. JAMA 2007;298:1300–1311.

17. TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung and BI, Crosby J, Peloso GM, et al. Loss-of-function mutations in APOC3, triglycerides, and coronary disease. N. Engl. J. Med. 2014;371:22–31.

18. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. Int. J. Epidemiol. 2015;44:512–25.

19. Carithers LJ, Ardlie K, Barcus M, et al. A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. Biopreserv. Biobank. 2015;13:311–319.

20. Fibbe WE, van Damme J, Billiau A, et al. Interleukin 1 induces human marrow stromal cells in long-term culture to produce granulocyte colony-stimulating factor and macrophage colony-stimulating factor. Blood 1988;71:430–5.

21. Kaushansky K, Lin N, Adamson JW. Interleukin 1 stimulates fibroblasts to synthesize granulocyte-macrophage and granulocyte colony-stimulating factors. Mechanism for the hematopoietic response to inflammation. J. Clin. Invest. 1988;81:92–97.

22. Segal GM, McCall E, Stueve T, Bagby GC. Interleukin 1 stimulates endothelial cells to release multilineage human colony-stimulating activity. J. Immunol. 1987;138:1772–8.

23. Kauma SW. Interleukin-1 beta stimulates colony-stimulating factor-1 production in human term placenta. J. Clin. Endocrinol. Metab. 1993;76:701–3.

24. Ridker PM, Everett BM, Thuren T, et al. Antiinflammatory Therapy with Canakinumab for Atherosclerotic Disease. N. Engl. J. Med. 2017;377:1119–1131.

25. Hamilton JA. GM-CSF in inflammation and autoimmunity. Trends Immunol. 2002;23:403–408.

26. Rajavashisth TB, Andalibi a, Territo MC, et al. Induction of endothelial cell expression of granulocyte and macrophage colony-stimulating factors by modified low-density lipoproteins. Nature 1990;344:254–7.

27. Rosenfeld ME, Ylä-Herttuala S, Lipton BA, Ord VA, Witztum JL, Steinberg D. Macrophage colony-stimulating factor mRNA and protein in atherosclerotic lesions of rabbits and humans. Am. J. Pathol. 1992;140:291–300.

28. Rajavashisth T, Qiao JH, Tripathi S, et al. Heterozygous osteopetrotic (op) mutation reduces atherosclerosis in LDL receptor-deficient mice. J. Clin. Invest. 1998;101:2702–2710.

29. Shaposhnik Z, Wang X, Lusis AJ. Arterial colony stimulating factor-1 influences atherosclerotic lesions by regulating monocyte migration and apoptosis. J. Lipid Res. 2010;51:1962–1970.

30. Lind L, Siegbahn A, Lindahl B, Stenemo M, Sundström J, Ärnlöv J. Discovery of New Risk Markers for Ischemic Stroke Using a Novel Targeted Proteomics Chip. Stroke. 2015;46:3340–7.

31. Zernecke A, Shagdarsuren E, Weber C. Chemokines in atherosclerosis an update. Arterioscler. Thromb. Vasc. Biol. 2008;28:1897–1908.

32. Abi-Younes S, Sauty a, Mach F, Sukhova GK, Libby P, Luster a D. The stromal cell-derived factor-1 chemokine is a potent platelet agonist highly expressed in atherosclerotic plaques. Circ. Res. 2000;86:131–138.

33. Zernecke A, Schober A, Bot I, et al. SDF-1α/CXCR4 axis is instrumental in neointimal hyperplasia and recruitment of smooth muscle progenitor cells. Circ. Res. 2005;96:784–791.

34. Döring Y, Pawig L, Weber C, Noels H. The CXCL12/CXCR4 chemokine ligand/receptor axis in cardiovascular disease. Front. Physiol. 2014;5 JUN:1–23.

35. Schunkert H, König IR, Kathiresan S, et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nat. Genet. 2011;43:333–338.

36. Ghasemzadeh N, Hritani AW, De Staercke C, et al. Plasma stromal cell-derived factor 1alpha/CXCL12 level predicts long-term adverse cardiovascular outcomes in patients with coronary artery disease. Atherosclerosis 2015;238:113–118.

37. Mehta NN, Li M, William D, et al. The novel atherosclerosis locus at 10q11 regulates plasma CXCL12 levels. Eur. Heart J. 2011;32:963–971.

38. Tragante V, Asselbergs FW, Swerdlow DI, et al. Harnessing publicly available genetic data to prioritize lipid modifying therapeutic targets for prevention of coronary heart disease based on dysglycemic risk. Hum. Genet. 2016;135:453–467.

39. Inoue A, Solon G. Two-Sample Instrumental variables estimators. Rev. Econ. Stat. 2010;92:557–561.

40. Verma S, Leiter LA, Bhatt DL. CANTOS Ushers in a New Calculus of Inflammasome Targeting for Vascular Protection—and Maybe More. Cell Metab. 2017;26:703–705.

4 BLOOD HER2 AND UROMODULIN AS CAUSAL MEDIATORS OF CHRONIC KIDNEY DISEASE

Jennifer Sjaarda[1,2,3], Hertzel C. Gerstein MD MSc[1,2], Salim Yusuf DPhil[1], Darin Treleaven MD PhD[6], Michael Walsh MD[1,6,7], Johannes F.E. Mann MD[8], Sibylle Hess PhD[9], Guillaume Paré MD MSc[1,2,4,5]

[1] Population Health Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, 237 Barton Street East, Hamilton, ON L8L 2X2, Canada

[2] Thrombosis and Atherosclerosis Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, 237 Barton Street East, Hamilton, ON L8L 2X2, Canada

[3] Department of Medical Sciences, McMaster University, 1280 Main Street West, Hamilton ON L8S 4K1, Canada

[4] Department of Pathology and Molecular Medicine, McMaster University, Michael G. DeGroote School of Medicine, 1280 Main Street West, Hamilton ON L8S 4K1, Canada

[5] Department of Clinical Epidemiology & Biostatistics, McMaster University, 1280 Main Street West, Hamilton ON L8S 4K1, Canada

[6] Department of Medicine, McMaster University, 1280 Main Street West, Hamilton ON L8S 4K1, Canada

[7] Department of Health Research Methods, Evaluation and Impact, McMaster University, 1280 Main Street West, Hamilton ON L8S 4K1, Canada

[8] KfH Kidney Center, Munich, Department of Medicine IV, University of Erlangen-Nurnberg, Germany

[9] Sanofi Aventis Deutschland GmbH, Research and Development Division, Translational Medicine and Early Development, Biomarkers and Clinical Bioanalyses, Frankfurt 65926, Germany

4.1 Forward

There are limited options to slow the progression of chronic kidney disease (CKD). Inhibition of the renin-angiontensin aldosterone system (RAAS) remains the primary intervention to preserve kidney function, yet the exact mechanism is not fully understood. In this report, we propose an innovative approach to identify novel, causal biomarkers of CKD by screening a comprehensive panel of 237 biomarkers in ORIGIN participants using Mendelian randomization (MR).

Our MR analysis identified human epidermal growth factor receptor 2 (HER2) as a causal mediator of CKD. Further MR exploration of the HER2 pathway also revealed ACE as a regulator of HER2 levels. Both MR findings were then corroborated in epidemiological analyses using blood HER2, incident CKD and BP-lowering medication data in ORIGIN. Taken together, our findings implicate HER2 as a mediator of ACE-inhibitors' protective effect on CKD and as a marker which may help reveal patients likely to benefit from ACE-inhibition. Furthermore, these findings suggest HER2-inhibition as a potential novel treatment for CKD, which may be applied through the use of HER2-inhibitors (e.g. gefitinib). We also identified uromodulin (UMOD) as a causal factor of CKD. Additional exploration of UMOD concentration in an independent sample of healthy nephrectomy donors found a nearly halving of plasma UMOD after transplant, compared with before. As GFR is preserved in these donors, these findings position UMOD as the first blood biomarker of kidney mass, to the best of our knowledge.

In summary, we found strong evidence for a causal relationship of UMOD and HER2 on CKD using a novel MR-based approach. Our results implicate HER2 as the mediator by which ACE-inhibitors are involved in CKD prevention, paving the way for novel therapeutic options as HER2 inhibitors are already clinically available. We further found UMOD to be a first blood marker of kidney mass. This manuscript was published in the *Journal of the American Society of Nephrology* in April 2018, Volume 29, Issue 4 (PMID: 29511113). Hertzel Gerstein and Guillaume

Paré conceptualized and designed the study. Jennifer Sjaarda designed the analysis plan, conducted all statistical analysis, and wrote the manuscript. All authors contributed to the interpretation of findings and to the critical reading and revision of the manuscript.

4.2 Abstract

BACKGROUND: Identification of biomarkers for chronic kidney disease (CKD) may lead to important advances in both prevention and treatment, particularly if they are causally linked. Many biomarkers have been epidemiologically linked with CKD, however, the possibility that such associations may be due to reverse causation or confounding limit their utility. This limitation can be overcome using a Mendelian randomization (MR) approach. We therefore used this technique to identify novel, causal mediators of CKD.

METHODS: MR was performed by first identifying genetic determinants of 227 protein biomarkers assayed in 4,147 ORIGIN trial (Outcome Reduction with Initial Glargine Intervention) participants with early or pre-diabetes, and assessing effects of these biomarkers on CKD in the CKDGen consortium (N=117,165, 12,385 cases) using the Wald method. The relationship between the serum concentration of each biomarker identified using this approach and incident CKD in ORIGIN participants was then estimated.

FINDINGS: Uromodulin (UMOD) and human epidermal growth factor receptor 2 (HER2) were identified as novel, causal mediators of CKD using MR (UMOD: OR=1.30 per SD; 95% CI 1.25 to 1.35; $p<5x10^{-20}$; HER2: OR=1.30 per SD; 95% CI 1.14 to 1.48; $p=8.0x10^{-5}$). Consistent with the MR findings, blood HER2 was also associated with CKD events in ORIGIN (OR=1.07 per SD; 95% CI, 1.01 to 1.13; $p=0.01$). Additional exploratory MR analyses identified Angiotensin converting enzyme (ACE) as a regulator of HER2 levels (β=0.13 per SD, 95%CI 0.08 to 0.16, $p=2.5x10^{-7}$). This latter finding was corroborated by an inverse relationship between ACE-inhibitor use and HER2 levels (0.25 SD decrease with ACE-inhibition, 95%CI -0.30 to -0.20, $p<5x10^{-16}$).

INTERPRETATION: UMOD and HER2 are independent causal mediators of CKD in humans, and HER2 levels are regulated in part by ACE. Both these biomarkers are potential therapeutic targets for CKD prevention.

4.3 Introduction

Chronic kidney disease (CKD) is a growing public health problem that increases the risk of cardiovascular disease, kidney failure and other complications[1]. Whereas glucose lowering, blood pressure lowering, and therapies that target the renin-angiotensin-aldosterone system (RAAS)[2] can slow progression of CKD, the mechanism(s) by which they work are not fully understood. Elucidation of these and other CKD-related mechanisms may identify novel therapies, and the identification of causal biomarkers for CKD represents one promising approach. Unfortunately, candidate biomarkers identified using traditional epidemiological approaches may be confounded with, or caused by other unmeasured biomarkers or mechanisms. Under a strict set of assumptions, Mendelian randomization (MR) can overcome these problems[3].

MR is a powerful genetic methodology that is based on the principle that genetic variants are inherited randomly and independently of other risk factors for disease. If the levels of a particular biomarker are affected by the presence of a genetic variant, and if that variant also affects the incidence of a disease, the variant may be causing the disease by modulating the levels of the risk factor. The random distribution of the genetic variant at birth minimizes the possibility of confounding or reverse causation as explanations for the link between the biomarker and disease, in the same way that the random allocation of a therapy in a randomized controlled trial minimizes this possibility.[4] These principles have been successful in identifying risk factors that are causal for CVD and biomarkers identified using MR have been validated in randomized trials of therapeutic agents[5]. Specifically, MR techniques have confirmed LDL-cholesterol, interleukin-6 receptor, and lipoprotein(a) as causal biomarkers of coronary artery disease.[6–8] Whereas there are fewer examples of MR in the field of nephrology, this approach has recently identified a causal effect of lower iron and ferritin levels on decreased kidney

function and has ruled out a causal relationship between fetuin-A and mortality in patients on dialysis[9,10].

MR methodology has traditionally been used to determine whether or not a candidate biomarker is causally related to a clinical outcome. However, when combined with a large panel of biomarkers measured in a prospective study which accrued many clinical outcomes, it can also be used to identify new, unsuspected biomarkers that are likely to be causally related to the clinical outcome. We therefore sought to identify such CKD biomarkers by applying MR techniques to a comprehensive panel of 237 biomarkers covering cardiovascular, metabolic and inflammatory processes within the recently completed Outcomes Reduction with an Initial Glargine Intervention (ORIGIN) trial that was performed in people with type 2 diabetes or pre-diabetes.[11]

4.4 Results

4.4.1 Identification of CKD biomarkers using MR

Two-hundred twenty-seven serum biomarkers were tested for an association with CKD using a MR approach. After removing biomarkers without any significant *cis* SNP associations with MAF>0.05 (according to CKDGen), 197 biomarkers were retained for downstream analysis. After MR analysis, two biomarkers were found to be significantly associated with CKD after Bonferroni correction for multiple hypothesis testing (p<0.05/197), namely uromodulin (UMOD) and human epidermal growth factor receptor 2 (HER2). As noted in Figure 4-1, the MR analysis suggested a deleterious effect of UMOD (odds ratio (OR)=1.30 per SD; 95% CI 1.25 to 1.35; $p<5\times10^{-20}$, number of SNPs=17) and HER2 (OR=1.30 per SD; 95% CI 1.14 to 1.48; $p=8.0\times10^{-5}$, number of SNPs=5) on CKD. All SNPs used in the MR models had p<0.01 and FDR<0.05 for the SNP-biomarker associations (Supplemental Tables 1 and 2). Regional plots of SNP associations with serum UMOD and HER2 at the *UMOD* and *ERBB2* loci, respectively, are depicted in

Supplementary Figure 1. To assess for the presence of unmeasured horizontal pleiotropy, we utilized the MR-Egger[12] method where the y-intercept is allowed to float, rather than be fixed at zero. We found no evidence of pleiotropy as determined by the significance of the y-intercept (p>0.05). As a sensitivity analysis, we use a leave-one-out strategy in which MR analyses were repeated, excluding one variant at a time and consistent estimates were obtained for each SNP excluded (see Supplemental Tables 3 and 4 material for full results).

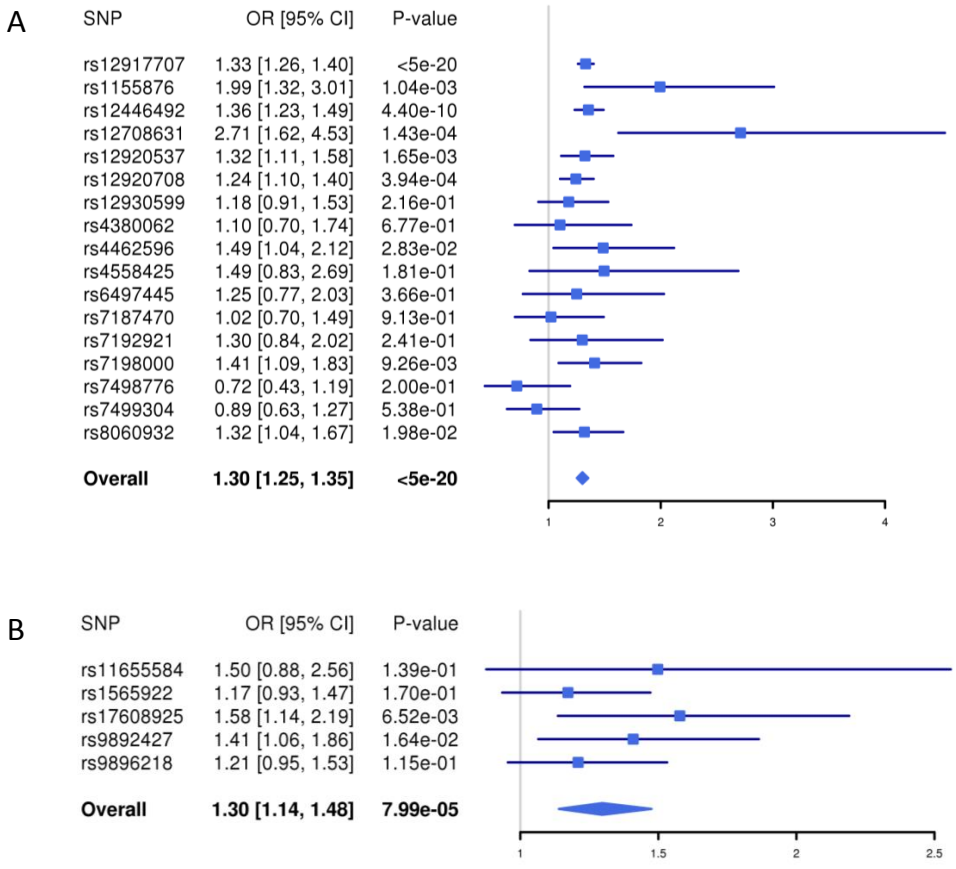


Figure 4-1: Association of UMOD and HER2 with risk of CKD using MR.

UMOD and HER2 identified as novel markers of CKD using MR. Forest plots depict a summary of the MR results for UMOD (A) and HER2 (B) at the *UMOD* and *ERBB2* locus, respectively. A single SNP MR was conducted for each independent SNP (pairwise r2<0.1). ORs were determined by the IVW method by regressing the effect estimates from the CKD association (from CKDGen) on the biomarker

association (from ORIGIN). A two-tailed p-value was calculated using a z-test from 100,000 random simulations.

4.4.2 Association of UMOD and HER2 concentration with CKD in ORIGIN

The MR-generated hypothesis that these biomarkers promoted CKD was then tested using the ORIGIN data. We therefore assessed the epidemiological relationship of baseline UMOD and HER2 with incident CKD. We found that increased levels of blood UMOD was significantly associated with decreased with risk in CKD, while increased levels of blood HER2 were associated with an increased risk of incident CKD in models adjusting for age, sex, and ethnicity (UMOD: OR=0.83 per SD; 95% CI 0.78 to 0.88; p<0.0001 and HER2: OR=1.07 per SD; 95% CI, 1.01 to 1.13; p=0.01). Consistent results were also observed in models fully adjusted for CKD risk factors and are provided in the supplementary material. We performed subgroup analyses to assess the consistency of the association of UMOD and HER2 concentration with CKD. No significant interaction across subgroups was observed after adjustment for multiple hypothesis testing (Figure 4-2).

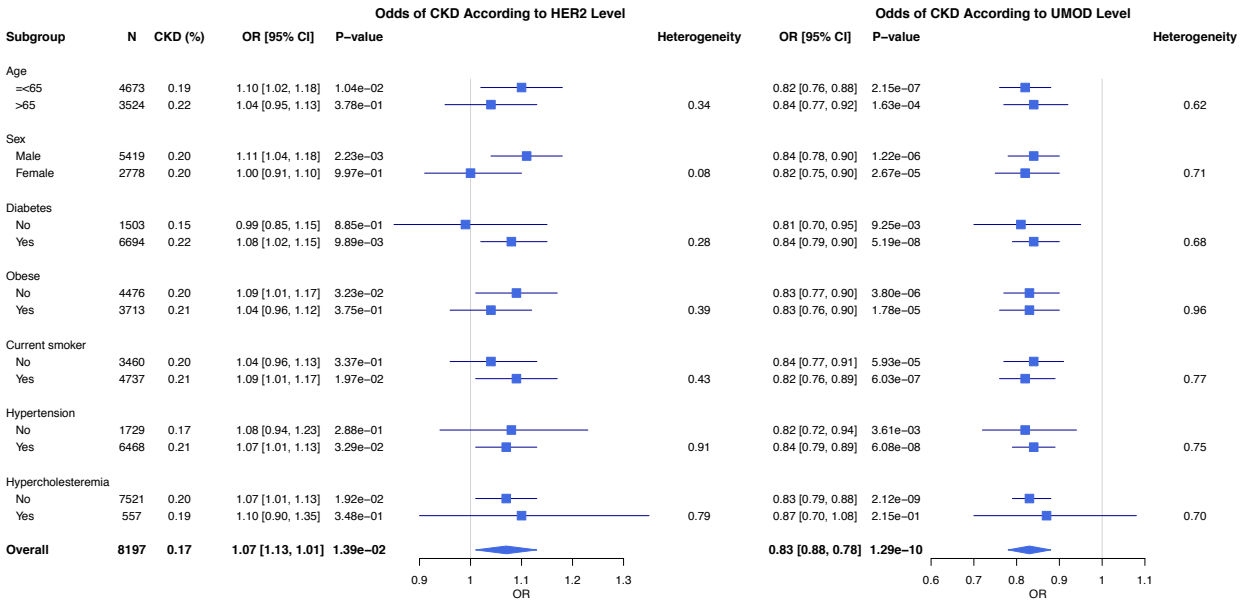


Figure 4-2: Subgroup analysis for association of UMOD and HERs levels with risk of CKD in ORIGIN.

Subgroup analysis for the epidemiological association of serum UMOD and HER2 levels with risk of CKD in ORIGIN. Models are adjusted (where appropriate) for age, sex and ethnicity. Subgroups were as follows: age (=<65, >65), sex (male, female), baseline diabetes status (yes/no), obese (BMI>=30, BMI<30), current smoker, hypertension diagnosis, and hypercholesterolemia (LDL-cholesterol >4.5 mmol/L, LDL-cholesterol=<4.5 mmol/L).

4.4.3 Association of UMOD with kidney mass in healthy nephrectomy patients

We assessed for the possibility that reverse causation could play a role in the apparent discrepant epidemiological association of UMOD with CKD (OR=0.83) and the evidence for UMOD as a CKD risk factor found in the MR (OR=1.30). Because UMOD is exclusively synthesized in the kidney, we hypothesized that UMOD is linked to a reduced kidney mass, and therefore lower UMOD expression, in CKD patients. Therefore, the hypothesis that UMOD concentration is a marker of kidney mass, rather than CKD progression, was explored in 10 healthy kidney donors. Briefly, participants had to meet clinical criteria for kidney donation, namely, normal blood pressure, non-smoker, eGFR>80mL/min and absence of major chronic disease. Mean age was 49.9 years and 30% were male (see supplementary material for further details). Indeed, UMOD blood levels were almost halved after uninephrectomy as compared to the pre-surgery period (μ=217.7 ng/mL, SD=75.6 vs μ=129.5, SD=39.1; p=7.6x10^{-5}, paired samples t-test) (see Figure 4-3). We also tested UMOD levels in urine (indexed to creatinine) and found a significant positive, correlation between blood and urine levels (R^2=0.29, p=0.018).

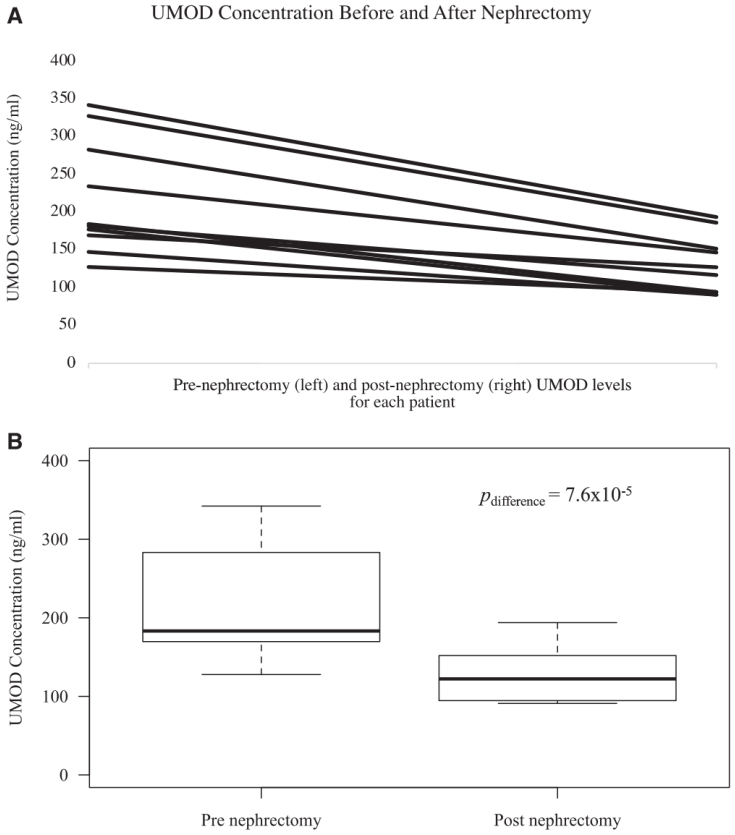


Figure 4-3: Difference between UMOD concentration pre and post nephrectomy in otherwise healthy patients.

Decreased serum UMOD concentration observed after nephrectomy compared with before. Figure A depicts concentration of blood UMOD concentration (ng/mL) in 10 otherwise healthy nephrectomy patients, before (left) and after (right) surgery. Notably, every patient showed a decrease in UMOD after nephrectomy, (paired t-test p=7.6x10-5). Figure B illustrates the box-plot of UMOD levels at the two time-points.

4.4.4 Identification of regulators of UMOD and HER2 using MR

To explore the mechanism by which UMOD and HER2 exert their effect on CKD, we tested the for a causal effect of all other biomarkers on UMOD and HER2 levels using MR. Specifically, we investigated whether any of the other biomarkers play a causal role in the regulation of UMOD and HER2 through a similar MR analysis using *cis* SNPs for the biomarker under study (where possible) as instrumental

variables. Significant cis regulators (p<0.001) were identified for 169 biomarkers (decreased from 197 due to the more stringent threshold) and were thus tested for their effect on UMOD and HER2. No significant biomarkers were found as regulators of UMOD (p<0.05/169). However, after adjusting for multiple hypothesis testing (p<0.05/169), angiotensin-converting enzyme (ACE) was identified as a causal regulator of serum HER2 levels (β=0.13 per SD, 95%CI 0.08 to 0.16, p=2.5x10^{-7}). We identified 16 *cis* ACE SNPs which were associated with ACE levels at p<0.001 to be used as instrumental variables in the MR analysis. We also tested for pleiotropic effects of UMOD and HER2 with other CV traits using MR, but found no significant associations after multiple hypothesis testing (data not shown).

Following the identification of ACE as a causal mediator of HER2, we decide to further explore this relationship by investigating the effect of different classes of blood pressure (BP) lowering medications on HER2 and ACE concentration ORIGIN. The following medications were assessed as dichotomous (yes/no) variables: ACE-inhibitors or angiotensin-II receptor blockers (ARB), diuretics (grouped as one variable), aldosterone inhibitors, beta-blockers, calcium channel blockers (CCBs). Specifically, a linear model was used, with HER2 or ACE concentration as the dependent variable, and use of medication (yes/no) as the independent variable (Table 4-1). The models were adjusted for age, sex, ethnicity, hypertension diagnosis, and prior renal disease, HER2 models were further adjusted for serum ACE levels. We identified a significant association between use of ACE-inhibitors/ARBs and HER2 concentration, indicating lower levels of HER2 in patients using ACE-inhibitors or ARBs (β=0.25 SD decrease with ACE-inhibition, 95%CI -0.30 to -0.20, p<5x10^{-16}), consistent with our MR findings indicating ACE increases HER2 levels. Conversely, no other BP medication was associated with lower levels of ACE after adjusting for multiple hypothesis testing (p<0.05/5), in fact, diuretics showed a marginal increase in HER2 levels consistent with an activation of the renin-angiotensin system with diuretics. Additionally, ACE-

inhibitors/ARBs, diuretics, and beta blockers were associated with increased levels of ACE, consistent with the RAAS inhibitory feedback loop. Aldosterone blockers and CCBs showed no effect on ACE levels after adjusting accounting for test multiplicity (p<0.05/5). A summary of ACE/HER2 findings and their effect on CKD risk can be seen in Figure 4-4.

Table 4-1: Epidemiological association of blood pressure medications on HER2 serum levels.

Medication	No. on Medication	β (95% CI)	P-value
ACE-inhibitors/ARBs	5641	-0.25 (-0.30, -0.20)	$<5x10^{-16}$
Diuretics	1309	0.08 (0.02, 0.13)	0.0096
Aldosterone antagonists	275	0.12 (0.008, 0.24)	0.037
Beta blockers	4426	-0.04 (-0.09, 0.001)	0.054
Calcium channel blockers	1561	-0.02 (-0.07, 0.04)	0.54

Estimates are given for medication use (yes/no) using HER2 concentration as a dependent variable. Models were adjusted for age, sex, ethnicity, hypertension, prior renal disease, and blood ACE levels.

β given as difference between those on medication compared to those not on medication.

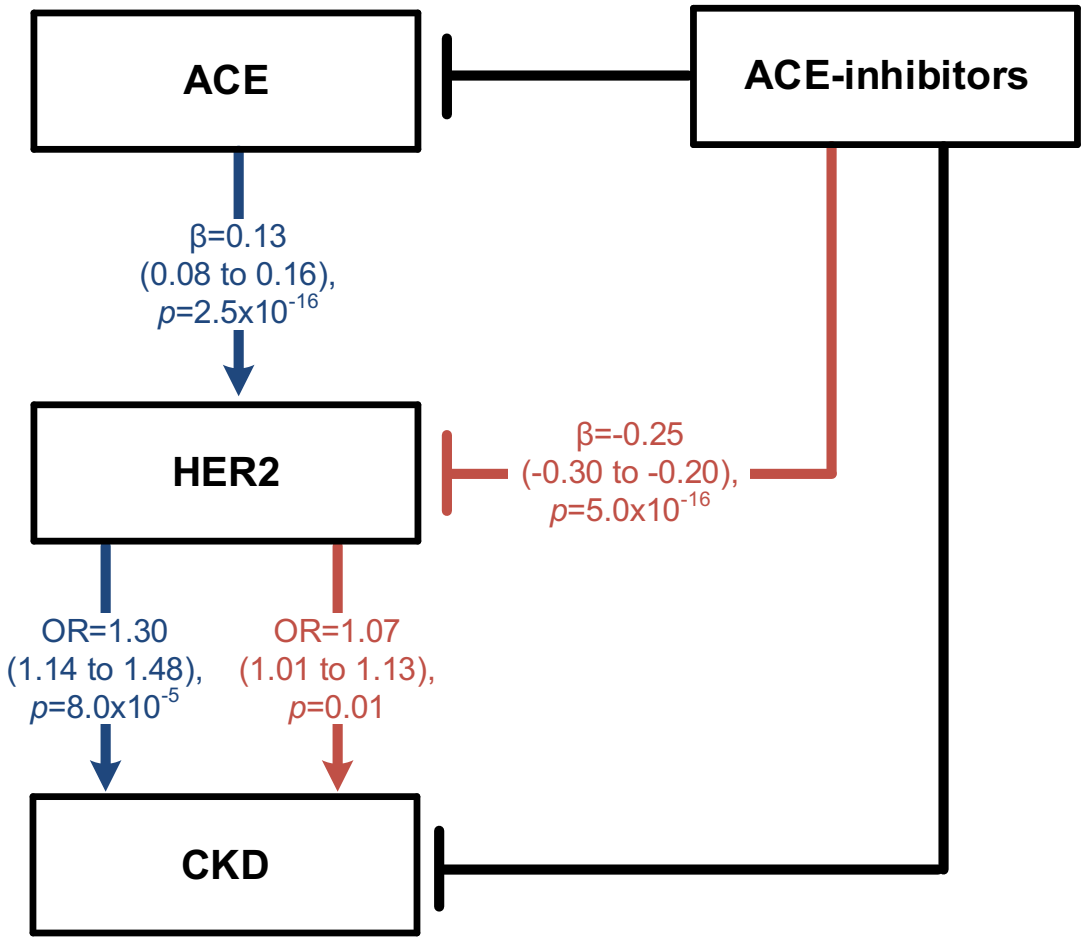


Figure 4-4: Summary of ACE/HER2 findings.

Summary of ACE/HER2 findings and their effect on CKD risk. Flow chart shows summary of ACE/HER2 results, red lines depict epidemiological associations, and blue lines depict MR associations. Black lines indicate previously known relationships.

4.5 Discussion

Using MR to screen a comprehensive panel of 227 blood biomarkers, we identified blood uromodulin and human epidermal growth factor receptor 2 as causal mediators of CKD. Uromodulin, also known as Tamm-Horsfall protein, is a kidney-specific protein ubiquitously expressed by the epithelial cells of the thick ascending loop of Henle. Under normal physiological conditions, UMOD is the most abundant

protein found in urine. Despite much research following its discovery more than 50 years ago, the role of UMOD in renal physiology remains unclear. Clinical and experimental studies implicated UMOD in several forms of bacteria clearance from the urinary tract and inflammatory kidney disease, although these findings were inconsistent[13].

With the advent of large GWAS, UMOD has emerged as an important locus in both CKD and hypertension[14–16]. A meta-analysis by Olden et. al identified a SNP in the *UMOD* promoter to be strongly associated with UMOD urinary levels, confirming the role of variants at the *UMOD* locus in UMOD excretion[17]. Furthermore, this same allele was previously identified to be associated with increased CKD risk[18]. These results indicate a positive relationship between urinary UMOD and CKD, consistent with our MR findings in blood. Notably, Trudu and colleagues have demonstrated that UMOD overexpression in transgenic mice led to salt-sensitive hypertension and activation of the renal sodium co-transporter NKCC2, this is consistent with their data in humans where pharmacological inhibition of NKCC2 was found to be more effective in lowering BP in patients homozygous for *UMOD* risk variants[19]. These findings establish a link between uromodulin, hypertension and CKD. Together with our results, these data point to uromodulin as a therapeutic target for lowering blood pressure and preserving renal function. However, it should be noted that it is impossible to know from these results if UMOD is acting through urine, blood or possibly an intra-cellular mechanism to exert its risk on CKD. In our epidemiological analysis, however, increased UMOD levels were associated with a decreased risk of incident CKD, consistent with other studies[20–22]. Possible explanations for the divergent direction of effect between the MR and epidemiological association include confounding and reverse causation. While adjustment for relevant risk factors did not alter conclusions, our analysis in healthy nephrectomy patients revealed almost halving of UMOD blood levels after uninephrectomy in healthy donors, consistent with previous studies[23]. Although we cannot rule out other biological and statistical explanations, including residual

confounding, these results suggest the protective epidemiological finding may be a result of reverse causation, reflective of loss of nephron mass in progressive kidney disease.

HER2 is a member of the human epidermal growth factor receptor (EGFR) family which are key regulators of cellular proliferation. The EGFR family has been implicated in CKD previously as EGFR signaling is involved in renal physiology through nephrogenesis, tissue repair and electrolyte balance. Numerous experimental studies have shown that pharmacological and genetic blockade of the EGFR system inhibits renal deterioration and fibrosis in animal models of kidney damage[24]. Additionally, in three independent cohorts of CKD patients, low urine excretion of EGF predicted accelerated loss of kidney function[25]. The authors suggested that urine low EGF excretion reflect reserved concentration in the tubules and represent a key factor in CKD progression. EGFR has also been shown to play a role in hypertensive and diabetic nephropathy[26,27]. For instance, EGFR expression is increased in the kidneys of hypertensive rats. Similarly, administration of gefitinib, an EGFR-tyrosine kinase inhibitor, improves renal function in rats with hypertension-induced renal disease[28]. Furthermore, significant reduction of diabetes-associated glomerular hypertrophy and renal enlargement has been seen upon blockade of EGFR signaling[29]. These animal studies are consistent with our MR and with epidemiological findings suggesting a causal effect of HER2 on CKD progression and development. Currently, HER2-inhibitors are used clinically in EGFR-mediated cancers[30,31]. Our data suggest to explore those drugs in models of kidney disease, as others have suggested[32].

We identified ACE as a positive regulator of HER2 levels, consistent with the observation of increased EGFR activity in hypertensive rats. Furthermore, we identified a lower concentration of HER2 in patients on ACE-inhibitors or ARBs versus those not on these medications, consistent with the hypothesis that ACE does increase HER2 levels. Moreover, as medication control we found no evidence

of other BP-lowering medications to have an effect on lowering HER2 levels. Together these results implicate HER2 as a mediator by which ACE-inhibitors and ARBs exert their protective effect on CKD patients, beyond that of other classes of antihypertensive drugs[33,34]. These results signify that blockade of RAAS through ACE-inhibitors and ARBS does indeed reduce HER2, and also suggest that HER2 levels may be able to guide RAAS inhibition. Finally, as noted previously, HER2 inhibitors are commonly used to treat EGFR-mediated cancers. Therefore, given these findings, it is possible that inhibition of ACE would decrease cancer risk by decreasing HER2 levels. Indeed, a very large, cross-sectional, cohort study, of nearly 300,000 individuals, identified a lower cancer risk in individuals on ACE-inhibitors and ARBs as compared to those not on these medications[35]. While these results should be cautiously interpreted and may be confounded, they do suggest a novel application of ACE-inhibitors which should be explored in future studies.

These findings are limited by the two sample MR study design, where genetic estimates were obtained from independent populations. In such a design, weak instrumental variables lead to estimates which are biased toward the null hypothesis, which reduced the likelihood of type 1 error but decreases power.[36] Thus not all of the causal biomarkers may have been identified. Additionally, the genetic variants studied may have other effects beyond its effect on the biomarker being studied (i.e. genetic pleiotropy). We mitigated this source of bias by limiting our investigation to variants at or near the gene coding for the biomarker of interest. Furthermore, associated loci were individually inspected for proximity to other potential genes and we did not identify any genes near the *UMOD* and *ERBB2* loci that were plausible sources of pleiotropy.

Identification of CKD risk factors is instrumental to further our understanding of the disease, evaluate its risk and guide treatment. Using MR, we have investigated a comprehensive panel of biomarkers for involvement in CKD. Our study presents the first MR analysis of blood UMOD and identified HER2 as a novel causal

mediator of CKD, consistent with previous model systems and the known biological role of these biomarkers. We also found compelling evidence to suggest HER2 as a mediator through which ACE inhibitors and ARBs protect again CKD progression and development. Increased serum UMOD and HER2 concentrations represent independent mechanisms leading to CKD, which can be assessed through a simple blood test. These findings pave the way for risk stratification and therapeutic interventions and provide important insights into the pathophysiology of CKD, and its relation to the current clinical practice of ACE-inhibitors and ARBs in CKD treatment. Future research should be aimed at identifying the causal mechanisms and whether interventions targeted at reducing UMOD and HER2 levels can reduce CKD.

4.6 Concise Methods

4.6.1 Study Population - ORIGIN

The design and findings of the ORIGIN trial have been described in detail[11,37]. Briefly 12,537 people with established cardiovascular risk factors who also had diabetes, impaired glucose tolerance, or impaired fasting glucose were studied. After random allocation to 2 therapies using a factorial design (basal insulin glargine versus standard care and omega 3 fatty acid supplements versus placebo) they were followed for a median of 6.2 years for cardiovascular events and other health outcomes. The ethics committee at each participating site approved the trial, and all participants provided written informed consent. As previously described[38], a subset of 8,401 participants from the ORIGIN-trial consented to further biological analysis and were therefore included in the biomarker ORIGIN sub-study. Biomarker levels were analyzed using the serum that was drawn at the beginning of the study (a detailed description of biomarker measurement and quality control is found in the supplement and a complete list of biomarkers analyzed is found in Supplemental Table 5).

4.6.2 CKDGen Consortium data

Genetic data on SNPs associations with CKD (defined as eGFRcrea < 60 ml min $^{-1}$ per 1.73 m^2) were obtained from the CKDGen database and downloaded from https://www.nhlbi.nih.gov/research/intramural/
researchers/ckdgen. Specifically, we used the most recent meta-analysis (released in 2015) of 43 genome-wide association studies (GWAS) with up to 117,165 individuals for CKD (12,385 cases), of European descent[39].

4.6.3 SNP association with biomarkers and CKD

The analysis was restricted to biomarkers directly encoded by a gene(s) on autosomal chromosomes (i.e. chromosome 1-22). Thus, removal of five biomarkers because they are products of genes on the X chromosome, and five biomarkers because they were not a direct gene product (e.g. cortisol), left 227 biomarkers for analysis.

SNP selection was carried out in four steps. First, as noted in Supplementary Figure 2, we restricted our analysis for each of the 227 biomarkers to SNPs within 300 Kb of the gene(s) encoding the corresponding protein or protein component, hereafter referred to as *cis* associations. This process identified 1,067,955 SNP/biomarker *cis* pairs (note that some SNPs are in *cis* with multiple biomarkers). Second, after removing SNPs not found in the CKDGen database and those with a minor allele frequency below 0.05 according to CKDGen, we estimated the relationship between the remaining SNPs and their corresponding *cis* biomarker(s) in ORIGIN, by regressing each SNP against the concentration of its *cis* biomarker (with biomarker concentration as the dependent variable and SNP dosage as the independent variable). In other words, for each biomarker we only tested SNPs near the respective encoding gene(s). The regression models were first computed in each ethnic group separately, adjusting for age, sex, and the first five principal components, using SNPtest.[40] The ethnic specific models were then meta-

analyzed across the two ethnicities using fixed effects models to minimize the risk of confounding caused by population stratification. Third, *cis* SNPs with a biomarker association of $p<0.01$ were selected. Finally, SNPs were pruned for linkage disequilibrium at a stringent threshold of $r^2<0.1$ using the 1000 Genomes data (Europeans) to ensure associations retained for MR analysis were non-redundant. SNPs were selectively prioritized based on the significance of the association with their biomarker. For each biomarker, the *cis* SNP with the most significant association with the biomarker was first retained and all SNPs in linkage ($r^2>0.1$) with that SNP removed. This process was then repeated for any remaining SNPs. 1,307 *cis* SNP/biomarker associations remained after pruning. SNP Filtering was performed in R (version 3.0.1) and PLINK was used to calculate R^2 statistics in 1000G. A summary of the SNP and biomarker selection can be found in Supplemental Figure 3.

4.6.4 Identification of blood mediators of CKD using MR

A two sample MR was performed on the 197 biomarkers which had at least one significant *cis* SNP ($p<0.01$) and that was also found in the CKDGen data. Input variables for the MR analysis for each biomarker, were (1) the beta coefficients of the SNPs on their *cis* biomarker that were estimated using the regression models above, and (2) the beta coefficients of the SNPs on CKD that were estimated from the CKDGen consortium (Supplementary Figure 4). MR associations were performed using the inverse-variance weighted method by regressing genetic effect estimates for CKD (dependent variable) on genetic effect estimates of biomarkers[41]. To determine significance, a bootstrap method was used under the null hypothesis of no effect between CKD and biomarkers. Predicted effects on CKD were sampled from a normal distribution with mean and standard deviations as determined from CKDGen. A two-tailed p-value was calculated using a z-test from 100,000 random simulations. In other words, CKD estimates for each SNP were sampled 100,000 times and regressed onto the corresponding beta estimates

from ORIGIN (this procedure is equivalent to the IVW fixed-effect method). Biomarkers were deemed significant after adjusting for multiple testing hypothesis (p<0.05/197).

4.6.5 Association of biomarker levels with incident CKD in ORIGIN

Once significant biomarkers were identified by the MR analysis, we tested whether biomarker levels showed a consistent association with incident CKD in 8,197 ORIGIN participants with ethnicity information and biomarker levels. CKD was defined as the composite of either doubling of serum creatinine, worsening in albuminuria category, renal replacement therapy, or death due to end stage renal failure. This was assessed using logistic models with incident CKD as the dependent variable and biomarker concentration as the independent variable of interest. Models were adjusted for age, sex and ethnicity to remain consistent with MR model adjustment. We also tested models after further adjusting for prior type 2 diabetes, prior renal disease, BMI, current smoker, diagnosis of hypertension, baseline eGFR, and LDL. Subgroup analyses were performed to test for heterogeneity between groups using models adjusted for age, sex and ethnicity (where appropriate).

4.6.6 Identification of regulators of CKD-biomarkers using MR

To gain further biological insights regarding the novel CKD biomarkers, a second set of MR analyses were then performed to explore whether the levels of the novel CKD-causing biomarkers were determined by any of the other biomarkers. Specifically, we tested all biomarkers for an effect on both UMOD and HER2 levels (i.e. the novel biomarkers identified in the CKD MR), where the input variables for each MR were (1) the effect of SNPs on their cis biomarker as the independent variable (where possible) and (2) the effect of the same set of SNPs on the novel CKD biomarkers as the dependent variable. Due to the fact that both these estimates were obtained from ORIGIN (i.e. one-sample), a more conservative

significance threshold of $p<0.001$ (or F>10) was applied for the inclusion of SNPs into the MR model as weak instruments can bias results towards false-positives[42]. Statistical analyses were performed using R (version 3.0.1), unless stated otherwise. A summary of the analysis plan can be seen in Figure 4-5.

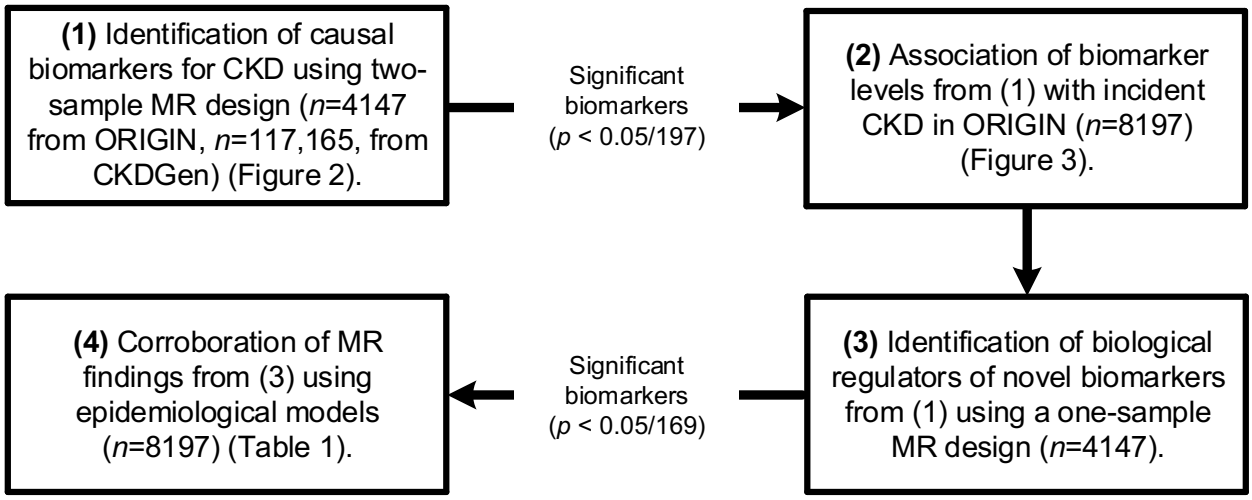


Figure 4-5: Overview of analyses conducted.

Flow chart depicts a summary of the four analyses conducted and their respective sample sizes.

4.7 Concluding Remarks

4.7.1 Significance Statement

Inhibition of the renin-angiotensin-aldosterone system by angiotensin-converting enzyme (ACE) inhibitors is one of the best established strategies to reduce the decline of kidney function in CKD. This study, which uses a novel Mendelian randomization–based approach, identifies human EGF receptor 2 (HER2) and uromodulin (UMOD) as potential causal mediators of CKD, and ACE as a potential regulator of HER2 levels. These findings implicate HER2 as a mediator of ACE inhibitors' protective effect on CKD and as a marker which may help reveal patients

likely to benefit from ACE inhibition. Both UMOD and HER2 inhibition represent potential novel targets for interventions to slow progression of CKD.

4.7.2 Conflict of Interest

H.C.G. has received consulting fees from Sanofi, Novo Nordisk, Lilly, AstraZeneca, Boehringer Ingelheim, and GlaxoSmith-Kline and support for research or continuing education through his institution from Sanofi, Lilly, Takeda, Novo Nordisk, BoehringerIngelheim, and AstraZeneca. G.P. has received consulting fees from Sanofi, Bristol Myers Squibb, Lexicomp, and Amgen and support for research through his institution from Sanofi. S.H. is an employee of Sanofi. J.M. has received consulting fees from Novo Nordisk, AstraZeneca, Amgen, Braun, ACI, Fresenius, Celgene, Gambro, Abbvie, Roche, Sandoz, Lanthio, Sanifit, Relypsam, and ZS Pharma, and grants from European Union, McMaster University Canada, Abbvie, Medice, Novo Nordisk, Roche, and Sandoz. J.S., D.T. and M.W. report no conflicts.

4.7.3 Funding

The ORIGIN trial and biomarker project were supported by Sanofi and CIHR (award 125794). Sanofi was not involved in the design or conduct of the paper, but provided comments on the manuscript. CIHR had no role in the conduct of the study or the manuscript.

4.7.4 Acknowledgements

We are thankful to all the participants having agreed to contribute to this project, and to the CKDGen consortia for making their data available.

4.8 References

1. James MT, Hemmelgarn BR, Tonelli M: Early recognition and prevention of chronic kidney disease. *Lancet* 375: 1296–1309, 2010

2. Shlipak MG, Day EC: Biomarkers for incident CKD: a new framework for interpreting the literature. *Nat. Rev. Nephrol.* 9: 478–83, 2013

3. Sekula P, Del Greco M F, Pattaro C, Köttgen A: Mendelian Randomization as an Approach to Assess Causality Using Observational Data. *J. Am. Soc. Nephrol.* 27: 3253–3265, 2016

4. Thanassoulis G, O'Donnell CJ: Mendelian randomization: nature's randomized trial in the post–genome era. *JAMA* 301: 2386–2388, 2009

5. Ference BA, Majeed F, Penumetcha R, Flack JM, Brook RD: Effect of naturally random allocation to lower low-density lipoprotein cholesterol on the risk of coronary heart disease mediated by polymorphisms in NPC1L1, HMGCR, or both: a 2 × 2 factorial Mendelian randomization study. *J. Am. Coll. Cardiol.* 65: 1552–61, 2015

6. Clarke R, Peden JF, Hopewell JC, Kyriakou T, Goel A, Heath SC, Parish S, Barlera S, Franzosi MG, Rust S, Bennett D, Silveira A, Malarstig A, Green FR, Lathrop M, Gigante B, Leander K, de Faire U, Seedorf U, Hamsten A, Collins R, Watkins H, Farrall M: Genetic variants associated with Lp(a) lipoprotein level and coronary disease. *N. Engl. J. Med.* 361: 2518–2528, 2009

7. Teslovich, T. Musunuru, K. Smith a. E Al: Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 466: 707–713, 2010

8. Sarwar N, Butterworth AS: Interleukin-6 receptor pathways in coronary heart disease: A collaborative meta-analysis of 82 studies. *Lancet* 379: 1205–1213, 2012

9. Del Greco M F, Foco L, Pichler I, Eller P, Eller K, Benyamin B, Whitfield JB, Genetics of Iron Status Consortium, CKDGen Consortium, Pramstaller PP, Thompson JR, Pattaro C, Minelli C: Serum iron level and kidney function: a Mendelian randomization study. *Nephrol. Dial. Transplant* 32: 273–278, 2017

10. Verduijn M, Prein RA, Stenvinkel P, Carrero JJ, Le Cessie S, Witasp A, Nordfors L, Krediet RT, Boeschoten EW, Dekker FW: Is fetuin-A a mortality risk factor in dialysis patients or a mere risk marker? A Mendelian randomization approach. *Nephrol. Dial. Transplant.* 26: 239–245, 2011

11. Gerstein HC, Yusuf S, Riddle MC, Ryden L, Bosch J: Rationale, design, and baseline characteristics for a large international trial of cardiovascular disease prevention in people with dysglycemia: The ORIGIN Trial (Outcome Reduction with an Initial Glargine Intervention). *Am. Heart J.* 155: 26. e1-26. e13, 2008

12. Bowden J, Davey Smith G, Burgess S: Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* 44: 512–25, 2015

13. Scolari F, Izzi C, Ghiggeri GM: Uromodulin: From monogenic to multifactorial diseases. *Nephrol. Dial. Transplant.* 30: 1250–1256, 2015

14. Padmanabhan S, Melander O, Johnson T, Di Blasio AM, Lee WK, Gentilini D, Hastie CE, Menni C, Monti MC, Delles C, Laing S, Corso B, Navis G, Kwakernaak AJ, van der Harst P, Bochud M, Maillard M, Burnier M, Hedner T, Kjeldsen S, Wahlstrand B, Sj??gren M, Fava C, Montagnana M, Danese E, Torffvit O, Hedblad B, Snieder H, Connell JMC, Brown M, Samani NJ, Farrall M, Cesana G, Mancia G, Signorini S, Grassi G, Eyheramendy S, Erich Wichmann H, Laan M, Strachan DP, Sever P, Shields DC, Stanton A, Vollenweider P, Teumer A, V??lzke H, Rettig R, Newton-Cheh C, Arora P, Zhang F, Soranzo N, Spector TD, Lucas G, Kathiresan S, Siscovick DS, Luan J, Loos RJF, Wareham NJ, Penninx BW, Nolte IM, McBride M, Miller WH, Nicklin SA, Baker AH, Graham D, McDonald RA, Pell

JP, Sattar N, Welsh P, Munroe P, Caulfield MJ, Zanchetti A, Dominiczak AF: Genome-wide association study of blood pressure extremes identifies variant near UMOD associated with hypertension. *PLoS Genet.* 6: 1–11, 2010

15. Köttgen A, Hwang S-J, Larson MG, Van Eyk JE, Fu Q, Benjamin EJ, Dehghan A, Glazer NL, Kao WHL, Harris TB, Gudnason V, Shlipak MG, Yang Q, Coresh J, Levy D, Fox CS: Uromodulin levels associate with a common UMOD variant and risk for incident CKD. *J. Am. Soc. Nephrol.* 21: 337–344, 2010

16. Pattaro C, Köttgen A, Teumer A, Garnaas M, Böger CA, Fuchsberger C, Olden M, Chen M-H, Tin A, Taliun D, Li M, Gao X, Gorski M, Yang Q, Hundertmark C, Foster MC, O'Seaghdha CM, Glazer N, Isaacs A, Liu C-T, Smith A V., O'Connell JR, Struchalin M, Tanaka T, Li G, Johnson AD, Gierman HJ, Feitosa M, Hwang S-J, Atkinson EJ, Lohman K, Cornelis MC, Johansson Å, Tönjes A, Dehghan A, Chouraki V, Holliday EG, Sorice R, Kutalik Z, Lehtimäki T, Esko T, Deshmukh H, Ulivi S, Chu AY, Murgia F, Trompet S, Imboden M, Kollerits B, Pistis G, CARDIoGRAM Consortium, ICBP Consortium, CARe Consortium, Wellcome Trust Case Control Consortium 2 (WTCCC2), Harris TB, Launer LJ, Aspelund T, Eiriksdottir G, Mitchell BD, Boerwinkle E, Schmidt H, Cavalieri M, Rao M, Hu FB, Demirkan A, Oostra BA, de Andrade M, Turner ST, Ding J, Andrews JS, Freedman BI, Koenig W, Illig T, Döring A, Wichmann H-E, Kolcic I, Zemunik T, Boban M, Minelli C, Wheeler HE, Igl W, Zaboli G, Wild SH, Wright AF, Campbell H, Ellinghaus D, Nöthlings U, Jacobs G, Biffar R, Endlich K, Ernst F, Homuth G, Kroemer HK, Nauck M, Stracke S, Völker U, Völzke H, Kovacs P, Stumvoll M, Mägi R, Hofman A, Uitterlinden AG, Rivadeneira F, Aulchenko YS, Polasek O, Hastie N, Vitart V, Helmer C, Wang JJ, Ruggiero D, Bergmann S, Kähönen M, Viikari J, Nikopensius T, Province M, Ketkar S, Colhoun H, Doney A, Robino A, Giulianini F, Krämer BK, Portas L, Ford I, Buckley BM, Adam M, Thun G-A, Paulweber B, Haun M, Sala C, Metzger M, Mitchell P, Ciullo M, Kim SK, Vollenweider P, Raitakari O, Metspalu A, Palmer C, Gasparini P, Pirastu M, Jukema JW, Probst-Hensch NM, Kronenberg F,

Toniolo D, Gudnason V, Shuldiner AR, Coresh J, Schmidt R, Ferrucci L, Siscovick DS, van Duijn CM, Borecki I, Kardia SLR, Liu Y, Curhan GC, Rudan I, Gyllensten U, Wilson JF, Franke A, Pramstaller PP, Rettig R, Prokopenko I, Witteman JCM, Hayward C, Ridker P, Parsa A, Bochud M, Heid IM, Goessling W, Chasman DI, Kao WHL, Fox CS: Genome-wide association and functional follow-up reveals new loci for kidney function. *PLoS Genet.* 8: e1002584, 2012

17. Olden M, Corre T, Hayward C, Toniolo D, Ulivi S, Gasparini P, Pistis G, Hwang S-J, Bergmann S, Campbell H, Cocca M, Gandin I, Girotto G, Glaudemans B, Hastie ND, Loffing J, Polasek O, Rampoldi L, Rudan I, Sala C, Traglia M, Vollenweider P, Vuckovic D, Youhanna S, Weber J, Wright AF, Kutalik Z, Bochud M, Fox CS, Devuyst O: Common Variants in UMOD Associate with Urinary Uromodulin Levels: A Meta-Analysis. *J. Am. Soc. Nephrol.* 1–14, 2014

18. Köttgen A, Glazer NL, Dehghan A, Hwang S-J, Katz R, Li M, Yang Q, Gudnason V, Launer LJ, Harris TB, Smith A V, Arking DE, Astor BC, Boerwinkle E, Ehret GB, Ruczinski I, Scharpf RB, Chen Y-DI, de Boer IH, Haritunians T, Lumley T, Sarnak M, Siscovick D, Benjamin EJ, Levy D, Upadhyay A, Aulchenko YS, Hofman A, Rivadeneira F, Uitterlinden AG, van Duijn CM, Chasman DI, Paré G, Ridker PM, Kao WHL, Witteman JC, Coresh J, Shlipak MG, Fox CS: Multiple loci associated with indices of renal function and chronic kidney disease. *Nat. Genet.* 41: 712–7, 2009

19. Trudu M, Janas S, Lanzani C, Debaix H, Schaeffer C, Ikehata M, Citterio L, Demaretz S, Trevisani F, Ristagno G, Glaudemans B, Laghmani K, Dell'Antonio G, Loffing J, Rastaldi MP, Manunta P, Devuyst O, Rampoldi L: Common noncoding UMOD gene variants induce salt-sensitive hypertension and kidney damage by increasing uromodulin expression. *Nat. Med.* 19: 1655–60, 2013

20. Steubl D, Block M, Herbst V, Nockher WA, Schlumberger W, Satanovskij R, Angermann S, Hasenau A-L, Stecher L, Heemann U, Renders L, Scherberich J:

Plasma Uromodulin Correlates With Kidney Function and Identifies Early Stages in Chronic Kidney Disease Patients. *Medicine (Baltimore).* 95: e3011, 2016

21. Delgado GE, Kleber ME, Scharnagl H, Krämer BK, März W, Scherberich JE: Serum Uromodulin and Mortality Risk in Patients Undergoing Coronary Angiography. *J. Am. Soc. Nephrol.* 28: 2201–2210, 2017

22. Devuyst O, Olinger E, Rampoldi L: Uromodulin: From physiology to rare and complex kidney disorders. *Nat. Rev. Nephrol.* 13: 525–544, 2017

23. Pruijm M, Ponte B, Ackermann D, Paccaud F, Guessous I, Ehret G, Pechére-Bertschi A, Vogt B, Mohaupt MG, Martin PY, Youhanna SC, Nägele N, Vollenweider P, Waeber G, Burnier M, Devuyst O, Bochud M: Associations of urinary uromodulin with clinical characteristics and markers of tubular function in the general population. *Clin. J. Am. Soc. Nephrol.* 11: 70–80, 2016

24. Tang J, Liu N, Zhuang S: Role of epidermal growth factor receptor in acute and chronic kidney injury. *Kidney Int.* 83: 804–10, 2013

25. Ju W, Nair V, Smith S, Zhu L, Shedden K, Song PXK, Mariani LH, Eichinger FH, Berthier CC, Randolph A, Lai JY-C, Zhou Y, Hawkins JJ, Bitzer M, Sampson MG, Thier M, Solier C, Duran-Pacheco GC, Duchateau-Nguyen G, Essioux L, Schott B, Formentini I, Magnone MC, Bobadilla M, Cohen CD, Bagnasco SM, Barisoni L, Lv J, Zhang H, Wang H-Y, Brosius FC, Gadegbeku CA, Kretzler M, ERCB, C-PROBE, NEPTUNE and P-IC: Tissue transcriptome-driven identification of epidermal growth factor as a chronic kidney disease biomarker. *Sci. Transl. Med.* 7: 316ra193, 2015

26. Swaminathan N, Vincent M, Sassard J, Sambhi MP: Elevated epidermal growth factor receptor levels in hypertensive Lyon rat kidney and aorta. *Clin. Exp. Pharmacol. Physiol.* 23: 793–6, 1996

27. Benter IF, Canatan H, Benboubetra M, Yousif MHM, Akhtar S: Global upregulation of gene expression associated with renal dysfunction in DOCA-salt-induced hypertensive rats occurs via signaling cascades involving epidermal growth factor receptor: A microarray analysis. *Vascul. Pharmacol.* 51: 101–109, 2009

28. François H, Placier S, Flamant M, Tharaux P-L, Chansel D, Dussaule J-C, Chatziantoniou C: Prevention of renal vascular and glomerular fibrosis by epidermal growth factor receptor inhibition. *FASEB J.* 18: 926–8, 2004

29. Wassef L, Kelly DJ, Gilbert RE: Epidermal growth factor receptor inhibition attenuates early kidney enlargement in experimental diabetes. *Kidney Int.* 66: 1805–14, 2004

30. Govindan R: A review of epidermal growth factor receptor/HER2 inhibitors in the treatment of patients with non-small-cell lung cancer. *Clin. Lung Cancer* 11: 8–12, 2010

31. Schroeder RL, Stevens CL, Sridhar J: Small molecule tyrosine kinase inhibitors of ErbB2/HER2/Neu in the treatment of aggressive breast cancer. *Molecules* 19: 15196–15212, 2014

32. Melenhorst WBWH, Mulder GM, Xi Q, Hoenderop JGJ, Kimura K, Eguchi S, van Goor H: Epidermal growth factor receptor signaling in the kidney: key roles in physiology and disease. *Hypertens. (Dallas, Tex. 1979)* 52: 987–93, 2008

33. Sarafidis PA, Khosla N, Bakris GL: Antihypertensive Therapy in the Presence of Proteinuria. *Am. J. Kidney Dis.* 49: 12–26, 2007

34. Molnar MZ, Kalantar-Zadeh K, Lott EH, Lu JL, Malakauskas SM, Ma JZ, Quarles DL, Kovesdy CP: Angiotensin-converting enzyme inhibitor, angiotensin receptor blocker use, and mortality in patients with chronic kidney disease. *J. Am. Coll. Cardiol.* 63: 650–658, 2014

35. Chiang YY, Chen KB, Tsai TH, Tsai WC: Lowered Cancer Risk With ACE Inhibitors/ARBs: A Population-Based Cohort Study. *J. Clin. Hypertens.* 16: 27–33, 2014

36. Inoue A, Solon G: Two-Sample Instrumental variables estimators. *Rev. Econ. Stat.* 92: 557–561, 2010

37. Bosch J, Gerstein HC, Dagenais GR, Diaz R, Dyal L, Jung H, Maggiono AP, Probstfield J, Ramachandran A, Riddle MC, Ryden LE, Yusuf S, Díaz R, Dyal L, Jung H, Maggiono AP, Probstfield J, Ramachandran A, Riddle MC, Rydén LE, Yusuf S, Hospital HG: n-3 fatty acids and cardiovascular outcomes in patients with dysglycemia. *N. Engl. J. Med.* 367: 309–18, 2012

38. Gerstein HC, Paré G, McQueen MJ, Haenel H, Lee SF, Pogue J, Maggioni AP, Yusuf S, Hess S: Identifying novel biomarkers for cardiovascular events or death in people with dysglycemia. *Circulation* 132: 2297–2304, 2015

39. Pattaro C, Teumer A, Gorski M, Chu AY, Li M, Mijatovic V, Garnaas M, Tin A, Sorice R, Li Y, Taliun D, Olden M, Foster M, Yang Q, Chen M-H, Pers TH, Johnson AD, Ko Y-A, Fuchsberger C, Tayo B, Nalls M, Feitosa MF, Isaacs A, Dehghan A, D'Adamo P, Adeyemo A, Dieffenbach AK, Zonderman AB, Nolte IM, van der Most PJ, Wright AF, Shuldiner AR, Morrison AC, Hofman A, Smith A V, Dreisbach AW, Franke A, Uitterlinden AG, Metspalu A, Tonjes A, Lupo A, Robino A, Johansson Å, Demirkan A, Kollerits B, Freedman BI, Ponte B, Oostra BA, Paulweber B, Krämer BK, Mitchell BD, Buckley BM, Peralta CA, Hayward C, Helmer C, Rotimi CN, Shaffer CM, Müller C, Sala C, van Duijn CM, Saint-Pierre A, Ackermann D, Shriner D, Ruggiero D, Toniolo D, Lu Y, Cusi D, Czamara D, Ellinghaus D, Siscovick DS, Ruderfer D, Gieger C, Grallert H, Rochtchina E, Atkinson EJ, Holliday EG, Boerwinkle E, Salvi E, Bottinger EP, Murgia F, Rivadeneira F, Ernst F, Kronenberg F, Hu FB, Navis GJ, Curhan GC, Ehret GB, Homuth G, Coassin S, Thun G-A, Pistis G, Gambaro G, Malerba G, Montgomery

GW, Eiriksdottir G, Jacobs G, Li G, Wichmann H-E, Campbell H, Schmidt H, Wallaschofski H, Völzke H, Brenner H, Kroemer HK, Kramer H, Lin H, Mateo Leach I, Ford I, Guessous I, Rudan I, Prokopenko I, Borecki I, Heid IM, Kolcic I, Persico I, Jukema JW, Wilson JF, Felix JF, Divers J, Lambert J-C, Stafford JM, Gaspoz J-M, Smith JA, Faul JD, Wang JJ, Ding J, Hirschhorn JN, Attia J, Whitfield JB, Chalmers J, Viikari J, Coresh J, Denny JC, Karjalainen J, Fernandes JK, Endlich K, Butterbach K, Keene KL, Lohman K, Portas L, Launer LJ, Lyytikäinen L-P, Yengo L, Franke L, Ferrucci L, Rose LM, Kedenko L, Rao M, Struchalin M, Kleber ME, Cavalieri M, Haun M, Cornelis MC, Ciullo M, Pirastu M, de Andrade M, McEvoy MA, Woodward M, Adam M, Cocca M, Nauck M, Imboden M, Waldenberger M, Pruijm M, Metzger M, Stumvoll M, Evans MK, Sale MM, Kähönen M, Boban M, Bochud M, Rheinberger M, Verweij N, Bouatia-Naji N, Martin NG, Hastie N, Probst-Hensch N, Soranzo N, Devuyst O, Raitakari O, Gottesman O, Franco OH, Polasek O, Gasparini P, Munroe PB, Ridker PM, Mitchell P, Muntner P, Meisinger C, Smit JH, ICBP Consortium, AGEN Consortium, CARDIOGRAM, CHARGe-Heart Failure Group, ECHOGen Consortium, Kovacs P, Wild PS, Froguel P, Rettig R, Mägi R, Biffar R, Schmidt R, Middelberg RPS, Carroll RJ, Penninx BW, Scott RJ, Katz R, Sedaghat S, Wild SH, Kardia SLR, Ulivi S, Hwang S-J, Enroth S, Kloiber S, Trompet S, Stengel B, Hancock SJ, Turner ST, Rosas SE, Stracke S, Harris TB, Zeller T, Zemunik T, Lehtimäki T, Illig T, Aspelund T, Nikopensius T, Esko T, Tanaka T, Gyllensten U, Völker U, Emilsson V, Vitart V, Aalto V, Gudnason V, Chouraki V, Chen W-M, Igl W, März W, Koenig W, Lieb W, Loos RJF, Liu Y, Snieder H, Pramstaller PP, Parsa A, O'Connell JR, Susztak K, Hamet P, Tremblay J, de Boer IH, Böger CA, Goessling W, Chasman DI, Köttgen A, Kao WHL, Fox CS: Genetic associations at 53 loci highlight cell types and biological pathways relevant for kidney function. *Nat. Commun.* 7: 10023, 2016

40. Marchini J, Howie B: Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* 11: 499–511, 2010

41. Burgess S, Butterworth A, Thompson SG: Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet. Epidemiol.* 37: 658–665, 2013

42. Burgess S, Thompson SG: Avoiding bias from weak instruments in mendelian randomization studies. *Int. J. Epidemiol.* 40: 755–764, 2011

5 INFLUENCE OF GENETIC ANCESTRY ON HUMAN SERUM PROTEOME

Jennifer Sjaarda[1,2,3], Hertzel C. Gerstein MD MSc[1,2], Pedrum Mohammadi-Shemirani [1,2,3], Marie Pigeyre MD PhD [1,2,3], Salim Yusuf DPhil[1], Sibylle Hess PhD[6], Guillaume Paré MD MSc[1,2,4,5]

1 Population Health Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, 237 Barton Street East, Hamilton, ON L8L 2X2, Canada

2 Thrombosis and Atherosclerosis Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, 237 Barton Street East, Hamilton, ON L8L 2X2, Canada

3 Department of Pathology and Molecular Medicine, McMaster University, Michael G. DeGroote School of Medicine, 1280 Main Street West, Hamilton ON L8S 4K1, Canada

4 Department of Clinical Epidemiology & Biostatistics, McMaster University, 1280 Main Street West, Hamilton ON L8S 4K1, Canada

5 Department of Biochemistry, McMaster University, 1280 Main Street West, Hamilton ON L8S 4K1, Canada

6 Sanofi Aventis Deutschland GmbH, Research and Development Division, Translational Medicine and Early Development, Biomarkers and Clinical Bioanalyses, Frankfurt 65926, Germany

5.1 Forward

Disease risk is known to vary significantly between ethnic groups, but the clinical significance and implications of these observations is not well understood. Investigating ethnic differences within the human proteome may shed light on the impact of ancestry on health and disease, as the proteome offers a direct window into biological processes. However, through epidemiological studies alone, it is very difficult to resolve the involvement of genetics versus environmental factors in ethnic disparities. Using a genetic approach known as admixture mapping, it is possible to overcome these barriers and ascertain the biological, unconfounded, and genetic differences between populations. In this study, we used admixture mapping to explore the impact of ancestry on a large, comprehensive panel of biomarkers in the ORIGIN-trial.

Our results revealed that 0.19 of biomarkers are affected by genetically-predicted ancestry. Notably, our strongest association was C-peptide, whereby 0.31 of the variation in C-peptide levels was attributed to genetic ancestry. Additional analyses revealed that a genetic risk score of ancestry determined C-peptide levels was associated with an increased risk of type-2 diabetes and measures of insulin resistance. Together, these results point to an effect of C-peptide on diabetes risk effect mediated through insulin resistance rather than β-cell dysfunction. Our results also revealed novel loci involved in regulating biomarker levels and demonstrate that specific genetic polymorphisms may partially explain the observed differences in biomarker concentrations between populations. These results may have implications regarding the interpretation of clinical markers in different ethnic groups.

This manuscript is currently under review by co-authors with plans to submit to the *American Journal of Human Genetics*. Guillaume Paré conceptualized and designed the study. Jennifer Sjaarda designed the analysis plan, conducted all

statistical analysis, and wrote the manuscript. All authors contributed to the interpretation of findings and to the critical reading and revision of the manuscript.

5.2 Abstract

BACKGROUND: Disease risk varies significantly between ethnic groups, but the clinical significance and implications of these observations is poorly understood. Investigating ethnic differences within the human proteome may shed light on the impact of ancestry on health and disease. However, it is difficult to determine if ethnic disparities are a result of genetic or environmental factors. Admixture mapping is a powerful method of gene mapping which overcomes these barriers and may help elucidate the impact of genetics on ethnic differences and identify genes conferring differential risk.

METHODS: Admixture mapping was used to explore the impact of ancestry on a comprehensive panel of 237 biomarkers in 2,216 Latin American participants within the ORIGIN-trial (Outcome Reduction with Initial Glargine Intervention). We developed a variance component model to determine the proportion of variance explained by local ancestral differences, and applied this model to the ORIGIN biomarker panel. Multivariable linear regression was used to identify and localize genetic loci affecting biomarker variability between ethnicities.

RESULTS: Variance component analysis revealed 0.05 of biomarkers to have a significant effect of ancestry, after adjusting for multiple hypothesis testing ($p < 0.05/237$), including C-peptide, apolipoprotein-E and intercellular adhesion molecule 1. We also identified 46 regional associations across 40 different biomarkers. An independent analysis revealed 34 of these 46 regions were associated at genome-wide significance ($p < 5x10^{-8}$) with their respective biomarker in either ORIGIN Europeans or Latins. Additional analyses revealed that a genetic risk score based on ancestral differences of C-peptide levels was

associated with an increased risk of diabetes (OR=4.47 per SD, 95% CI 1.70 to 11.76, p=0.002) and measures of insulin resistance.

DISCUSSION: Our results demonstrate the importance of ancestry on biomarker levels, suggesting some of the observed differences in disease prevalence likely has a biological basis, and that use of reference intervals for those biomarkers should be tailored to ancestry. Specifically, our results point to a role of ancestry in insulin metabolism and diabetes risk.

5.3 Introduction

The human proteome plays a principal role in biological processes such as signaling, transport, growth, repair, and defense against infection. These proteins represent intermediate phenotypes and are often directly and causally involved in disease pathophysiology. Indeed, many biomarkers are measured clinically and used as a non-invasive marker of a patient's overall health, guiding diagnosis, prognosis and treatment management[1]. However, biomarker profiles have been shown to vary widely between ethnic groups and the clinical significance and implications of these observed differences is poorly understood[2]. Furthermore, it is unknown whether these differences correspond to ethnic-specific susceptibility to disease. Disease risk varies significantly between ethnic groups as well. For instance, Mexican, Latin American and African populations have a higher risk of type 2 diabetes (T2D) compared to European ancestries[3–5]. This disparity in risk has been hypothesized to be due, at least in part, to genetic and biological factors rather than confounding[6,7]. These findings suggest that other phenotypes may similarly harbor genetic variants that account for differences between ancestries.

Biomarker differences that exist between populations may also lead to clinical challenges. Consistent differences have been reported for many biomarkers used in patient management, including C-reactive protein, vitamin D binding protein and many circulating adipokines[8–12]. Clinical interpretation of these markers are based on reference intervals which are defined using population values. However, for biomarkers that are markers of disease, this might lead to erroneous diagnosis if ancestry leads to differences in concentrations. For biomarkers that are causal mediators, this might lead to wrongful evaluation of risk if ancestry leads to increased risk through that mediator. Ideally, these intervals should be determined from a random sample of healthy individuals from a population similar to the patient. Traditionally, reference intervals have been determined using

predominantly Caucasian reference intervals, which do not necessarily extend to other ethnic groups[13].

While differences in biomarkers levels have been observed between ethnic groups, the reasons for these differences are difficult to determine through classic epidemiological studies. Admixture mapping is a powerful tool used in genetic epidemiological studies that may shed light on these observations. Genetic admixture occurs when two or more previously independent populations interbreed, resulting in the introduction of new genetic lineages. This has occurred in Latin Americans, for instance, which are an admixed population from Native American, European and African ancestors. Admixture mapping is a method applied to recently admixed populations used to localize disease causing genetic variants that differ in frequency across ancestral groups[14]. The approach is based on the assumption that increased ancestry from the population with a greater risk of the disease will be observed in patients near a disease causing gene. In this way, differential risk across ancestral groups can be observed at specific genetic loci[15]. This approach has been particularly effective in African Americans in identifying novel loci for various diseases[16–18]. Most recently, this technique has been used to reveal novel susceptibility loci in atherosclerosis and albuminuria[19,20].

In this report, we used admixture mapping to investigate the impact of ancestry on health and disease through a comprehensive investigation of a multiplex biomarker panel. Specifically, we evaluated the effect of genetic ancestry on 237 serum biomarker concentrations measured in the Latin American population from the recently completed ORIGIN (Outcomes Reduction with an Initial Glargine Intervention) trial[21]. Although ethnicity has been determined to be a strong predictor of biomarker concentrations, few studies have leveraged the known genetic admixture of Latin Americans to assess the impact of ancestry on biomarker variability and discover novel regions associated with their levels that may, in turn, impact their risk of disease. Furthermore, admixture mapping studies offer the

unique advantage over single-ancestry genetic studies at identifying genes conferring differential risk between populations. By examining a large, comprehensive panel of serum biomarkers, we sought to explore the broad impact of ancestry in human health and disease. The human proteome provides an ideal paradigm for such an investigation as it provides a direct window into biological processes.

5.4 Methods

5.4.1 Study Population - ORIGIN

The design and findings of the ORIGIN trial have been described in detail. Briefly 12,537 people with established cardiovascular risk factors who also had T2D, impaired glucose tolerance, or impaired fasting glucose were studied. After random allocation to 2 therapies using a factorial design (basal insulin glargine versus standard care and omega 3 fatty acid supplements versus placebo) they were followed for a median of 6.2 years for cardiovascular events and other health outcomes. As previously described[22] biomarker levels were analyzed in the serum of 8,401 people that was drawn at the beginning of the study. The analysis was done using a customized human discovery multi-analyte profile (MAP) on the Luminex 100/200 platform and the biomarkers were selected based on their implication in physiologic processes related to cardiovascular diseases.

Between September 2003 and December 2005, 578 clinical sites in 40 countries screened 15,374 individuals and randomized 12,537 participants for the original ORIGIN trial[21]. A subset of 8,401 participants provided consent for collection and storage of a blood sample for future measurement and analysis were included in the ORIGIN biomarker study (66% men; mean age 63.7 years).[22] A further subset of 5,078 participants consented to genetic analyses and 4,147 (1,931 Europeans and 2,216 Latins) passed quality control. Study characteristics were similar across the two groups.

5.4.2 Genotyping

A subset of 5,078 ORIGIN individuals who consented to genetic analyses were genotyped on Illumina's HumanCore Exome chip. Standard quality control measures were assessed. SNPs were excluded on the basis of low call rate (<99%), deviation from Hardy-Weinberg ($p<10^{-6}$), and low minor allele frequency (MAF<0.01 in all ethnic groups). Samples with low call rates (<99%), sex or ethnicity mismatches, or cryptic relatedness were also removed. We also removed ethnicities with small sample sizes (n<100). All quality control steps were performed using PLINK[23] and GCTA.[24] After quality control, the sample consisted of 4,390 participants and 284,024 SNPs from three ethnic groups (European, Latin American and African). Imputation was then performed on the post QC data through to predict unobserved genotypes in the study population. Over 30 million single nucleotide polymorphisms (SNPs) were imputed, allowing for comprehensive coverage of known genetic variants. The 1000 Genomes Project[25] was used as the reference panel for ORIGIN imputation and was performed using the software IMPUTE2[26,27]. We removed SNPs imputed with low certainty (info <0.6, as defined by IMPUTE2)[27]. patients. For the current report, participants of self-reported Latin American ancestry comprised the primary analysis (n=2,216) and Europeans were used for validation and replication analyses (n=1,931).

5.4.3 Genetic Ancestry Estimation

We used phased, consensus data from the 1000 Genomes Project to create reference panels for Europeans (CEU, FIN, GBR, IBS and TSI), Africans (ASW, LWK and YRI), and Asians alleles (CHB, CHS and JPT; which were used as a proxy for Native American ancestry, as previously described[15]). After removing ambiguous SNPs and phasing ORIGIN genotypes using Beagle[28], we inferred the local ancestry at 259,778 SNPs in 2,216 Latin Americans using RFMix[15]. Probabilities of Native American, European and African ancestry were derived for each SNP, thus accounting for uncertainty in ancestry ascertainment. Probabilities

at each SNP ranged from 0 to 2, where a value of 2 for the European local component at given SNP would represent both alleles having European ancestry, for example. The procedure has been described in detail elsewhere[15]. To calculate individual-level ancestries, a set of minimally pruned sites was generated according to an LD correlation matrix based on local SNP European ancestry components, in R (pairwise r^2 < 0.95). Specifically, a square matrix was constructed for each chromosome containing the Pearson's r^2 correlation coefficient between all site (i.e. pairwise correlation). For example, for any two sites (x and y) the r^2 was calculated between the local Asian components at $site_x$ and $site_y$. The resulting matrix was pruned agnostically at a threshold of r^2 <0.95. Currently, there is no standard method for pruning local admixture signals, and this threshold was chosen to reduce redundant (identical) associations while retaining as much ancestry information as possible. Pruning using genotype LD (rather than local ancestry LD is not sensible here, as admixture regions are much larger than haplotype blocks across the genome. Therefore, this threshold was selected in an effort to balance over-pruning, and ultimately losing local admixture signals, and under-pruning, resulting in redundant signals. Following pruning, 7,246 local components remained. This set was used for all subsequent analyses. Individual-level (global) ancestry was then obtained for each individual by averaging the ancestry at each of the retained sites. Thus, following this procedure, each individual had three local ancestry components (1 for each of the three ancestral ethnicities) for each site ranging from 0 to 2 and three global ancestry components ranging from 0 to 2 representing the average of all locally derived estimates.

5.4.4 Genetic Association Models to Determine Contribution of Local Ancestry on Phenotypic Variation

We evaluated the performance of genetic association models to capture the phenotypic variance explained by local ancestry using simulations. Because associations with global ancestry may represent confounding by environmental or

societal factors rather than a true biological difference, we sought to distinguish between local and global effects to determine the variance explained according to biological differences (i.e. local ancestry) between ethnic groups. Continuous phenotypes were simulated for each of 2,216 Latins in ORIGIN using the derived local and global ancestry components as predictors. We explored various parameters for their impact on estimated local ancestry variance, including: effect of non-directional versus directional local effects (i.e. restricting local effects to be positive for a given ancestry in the directional case), number of causal loci associated with the simulated trait, and presence and absence of a global ancestry effect. Local directional effects were evaluated to test the model's ability to distinguish between many local signals exerting an effect in the same direction versus a single global (confounding) effect. Total trait variance and mean were set at 1 and 0, respectively, in all simulations. For each simulation, a pre-specified set of causal loci ranging from 1 to 10 (1, 2, 3, 5, and 10) were randomly selected from a stringently pruned set of 46 local components (r^2 <0.05) to ensure independent regions were selected. The genetic effect of each causal locus was proportionally set according to the number of causal loci specified and a pre-defined, unobserved, true local variance ranging from 0 to 0.05. Similarly, the effect of each global component was standardized and fixed according to a pre-defined overall variance value of either 0 or 0.05. The remaining phenotypic variance was randomly determined. The effect of each locus on the simulated trait was evaluated using adjusted linear models.

Because ancestry tends to be highly correlated over longer regions of the chromosome as opposed to genotype data, the number of independent tests estimated is small despite the inclusion of genome-wide level ancestry data in the model. Therefore, to determine an appropriate significance level, we performed 10,000 simulations under the null hypothesis, assuming no effect of local ancestry on the simulated phenotype. For each simulation, a continuous phenotype was derived with no effect of local ancestry and both with and without an effect of global

ancestry. Next, using European ancestry as a reference, each local Asian and African component were tested independently in a linear model adjusted for global Asian and global African components. In other words, for each simulation, 14,492 (7,246 loci times 2 ethnicities) linear models were tested for an association with the simulated trait. The lowest p-value from the 14,492 independent tests was recorded ($p_{minimum}$). We did not identify any difference in distribution of $p_{minimum}$ with and without a global effect. We selected a p-value threshold which corresponded to < 1% of the $p_{minimum}$ resulting in a significance threshold level of $p < 1.13 \times 10^{-6}$.

For each set of conditions, 100 simulations were completed and both the effect of local ancestry and global ancestry on trait variance estimated. We used variance component (VC) models to assess the overall effect of ancestry on the simulated trait using the mmer2 function in the *sommer* R package[29]. Genetic related matrices (GRMs) were calculated for each ancestry using local ancestry estimates at the remaining 7,246 sites after pruning (described above). The local ancestry matrices (2,216 x 7,246) were scaled to have mean of 0 and standard deviation of 1. Next, the GRM was calculated as the cross-product of the scaled local ancestries. Global Asian and African ancestry were each included in the model as fixed effects. Proportion of variance explained by global and local ancestry (both together and separately) was then estimated for each model and compared to the value specified for each simulation. Global ancestry variance was estimated using the regression coefficients from the fixed effect estimates in the VC model. The mmer2 function provided variance-covariance components for each random effect (i.e. the two local ancestry GRMs and residual variance) and were used to estimate local and residual variance accordingly. Total trait variance was estimated as the sum of global, local and residual variance estimates. Next, we calculated the proportion of variance explained due to local, global and the sum of local and global ancestry as their respective estimated variance divided by total trait variance. Estimates were recorded for each simulation and the average (±SD) of each set of

conditions was calculated and compared to their unobserved, true, respective values.

We sought to identify the individual loci selected for a causal association with the simulated trait. Specifically, each local ancestry component for both Asian and African ethnicities was independently tested in a linear model with the simulated trait as the dependent variable, adjusted for global Asian and global African components. A forward-selection approach was then used to identify the local components that independently and cumulatively predicted the dependent variable, with a p-value for inclusion set at the pre-specified threshold according to simulations under the null ($p < 1.13 \times 10^{-6}$). The minimum p-value of each of the 14,492 models representing all local ancestry components for both African and Asian ancestries was evaluated, and if it fell below the threshold for inclusion, the respective local component was added in the predictive model in addition to global African and global Asian components. This process was repeated until no local component association p-value fell below 1.13×10^{-6}. Because 45 of the biomarkers were analyzed as ordinal variables, all simulations were repeated using a simulated ordinal trait to test the models ability to perform with a non-continuous dependent variable.

We then evaluated the proportion of identified loci that matched the randomly selected causal loci for a given simulation (i.e. true positives). When a locus was identified by our algorithm that was not randomly selected to have an effect on the simulated phenotype, we evaluated if it was a false positive or near a true, unobserved, causal locus, representing a regional association. We used a threshold of $r^2 > 0.8$ with a causal locus, to define a regional association. Identified loci with $r^2 < 0.8$ with all randomly selected causal loci were classified as false positives.

5.4.5 Estimation of Effect of Local Ancestry on Serum Biomarkers in ORIGIN

The variance component model and forward selection process described above were then performed on the 237 measured biomarkers in ORIGIN to determine the proportion of variance explain by local ancestry for each serum biomarker. A predictive model was constructed for each biomarker according to the following procedure. First, biomarkers were linearly residualized for age and sex. Second, VC models were used to assess the proportion of trait variance explained by local ancestry with global components as fixed effects (as in simulations). Third, linear models were used to test each local component independently for an effect on the residualized biomarker, in a model adjusted for global Asian and global African components. As described above, the minimum p-value of all local component was assessed and if less than our inclusion threshold ($p < <1.13 \times 10^{-6}$), was added to the predictive model. This process was then repeated until no p-values were less than 1.13×10^{-6}. Therefore, for each biomarker, one VC model was used to assess overall local and global variance and a linear predictive model was constructed, including global African and global Asian components in addition to local components selected from the forward selection algorithm. This forward selection process revealed specific genetic regions which were independently associated with serum biomarkers, residualized for age and sex. Associations were classified as in *cis* if the identified locus had $R^2 < 0.8$ with any SNP ± 3000 KB for the respective gene.

We sought to further elucidate identified local ancestry associations by testing genotype associations in European and Native Latin ORIGIN samples (n=1,931). For each identified local ancestry association we implemented the following process. First, an investigation window surrounding the local ancestry signal was created according pairwise r^2 of European local ancestry data. Pairwise r^2 was examined both up and down stream of the identified locus until a local ancestry estimate had $r^2 <0.8$ with the identified locus to create a window of association.

Second, the association of each SNP in this window was tested in ORIGIN Europeans with the respective biomarker using a linear model, adjusted for age, sex and the first five principal components. Third, the association of each SNP in this window was also tested in ORIGIN Native Latins with the respective biomarker using a linear model, adjusted for age, sex, global ancestry and the corresponding local ancestry components for the SNP under investigation. SNPs with MAF <0.01 or INFO <0.6 were removed.

5.4.6 Prediction of Global Ancestry Using Biomarker Score

We sought to determine the ability of the 237 serum biomarker to predict global ancestry components in ORIGIN. Participants were randomly divided into model building and model assessment groups composed of 1,477 people (i.e. 0.67) and 739 people (0.33), respectively. Using the model building subset, those biomarkers that independently predicted global European and global Asian components were identified using a multivariate analysis approach after accounting for age and sex. To assess the significance of each biomarker, a multivariate ANOVA was used to account for the two outcomes (African and Asian global ancestry components) comparing a multivariate model with and without each biomarker. Similar to the method to determine local ancestry variance, a forward-selection approach was used to identify those biomarkers that independently and cumulatively predicted global ancestry according to the multivariate ANOVA, with a p-value for inclusion set below 0.05 divided by 237 ($p < 0.00021$), to account for the 237 comparisons. The biomarkers identified to independently predict global ancestry components in the model building group was then validated in the model assessment group and the resulting correlation for each model were compared. All statistical analysis was done in R.

5.5 Results

5.5.1 Evaluation of Genetic Association Models Using Simulations

We evaluated the performance of our variance component models to estimate the phenotypic variance explained by local admixture associations. In these simulations, we assumed that varying number of loci (1, 2, 3, 5, 10) had an ancestry effect on the quantitative trait and that the proportion of variance explained was 0.00, 0.01, 0.02, 0.03, 0.04, and 0.05. We then tested conditions with and without a directional condition on the causal ancestry effects (i.e. all effects greater than 0 for a given ancestry) and with and without an effect of global admixture. When a global effect was specified, it was split evenly over the two components (African and Asian), each with a proportion of variance explained of 0.025. Our simulations show that total variance attributed to local and global ancestry can be determined using VC models (Figure 5-1). Similarly, our simulations show that it is possible to derive unbiased estimates of local variance using VC models (Figure 5-2). These estimates are stable both with a directional local effect and in the presence of a global effect. However, local estimates were lower in the directional scenarios compared to non-directional. For instance, considering a scenario with 10 causal loci and local variance specified at 0.05, 2-way ANOVA revealed significant differences between directional and non-directional simulations (Figure 5-2, panels A and B versus panels C and D, $p < 5 \times 10^{-16}$) and no difference between simulations with and without a global effect (Figure 5-2 panels A and C versus panels B and D, $p = 0.73$). This is likely due to the fact that it is difficult for the model to distinguish a global effect from a directional local signal, particularly when many causal loci are present.

We also sought to determine the ability of the model to select the true, unobserved causal loci. The proportion of causal SNPs selected increased as specified local variance increased (Figure 5-3) number of and did not vary significantly across conditions. When only one SNP was specified to have an effect on the phenotype,

the algorithm performed well, and identified this locus greater than 95% of the time when local variance was greater than 0.05. Conversely, as number of causal loci increased, the resulting effects were diluted across the randomly selected SNPs, and power to detect individual loci decreased. Consequently, the algorithm was unable to detect all of the true, causal SNPs. This pattern was apparent for all conditions, however this was strongest in the presence directional local effect (panels C and D). For example, in the simulations with 10 causal loci and local variance specified at 0.05, ANOVA showed significant differences in the proportion of causal loci selected between those with and without a global effect (Figure 5-3 panels A and B versus C and D, $p < 5 \times 10^{-16}$). Similar results were found by simulating an ordinal rather than a continuous phenotype (see supplementary material).

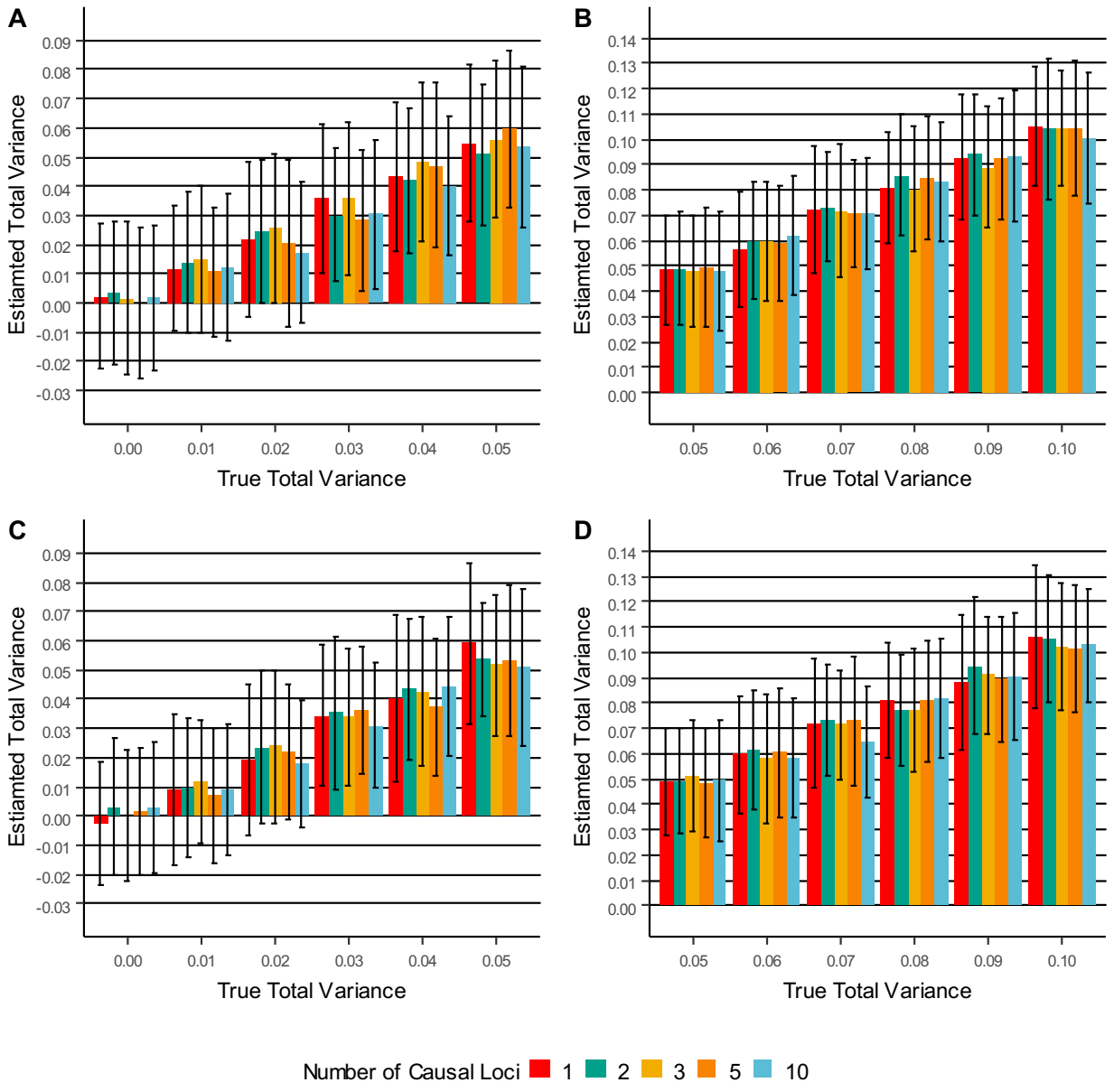


Figure 5-1: Estimated proportion of variance explained by local and global ancestry under various conditions.

Average (±SD) estimated local and global variance explained under various simulated conditions. The sum of the true, unobserved local and global variances were pre-specified. Global variance was set at either 0 (panel A and C) or 0.05 (panel B and D) and local varied as 0.01, 0.02, 0.03, 0.04 and 0.05 as determined by 1, 2, 3, 5, or 10 causal SNPs. The sum of the two variances is shown on the x-axis. Each bar represents an average of 100 simulations, error bars show ±SD. Panels A and B illustrates simulated conditions with no directional condition, while

panels C and D restrict local effects to be greater than 0. Panel A and C illustrate simulated conditions with no global effect, and Panel B and D have a pre-specified global effect.

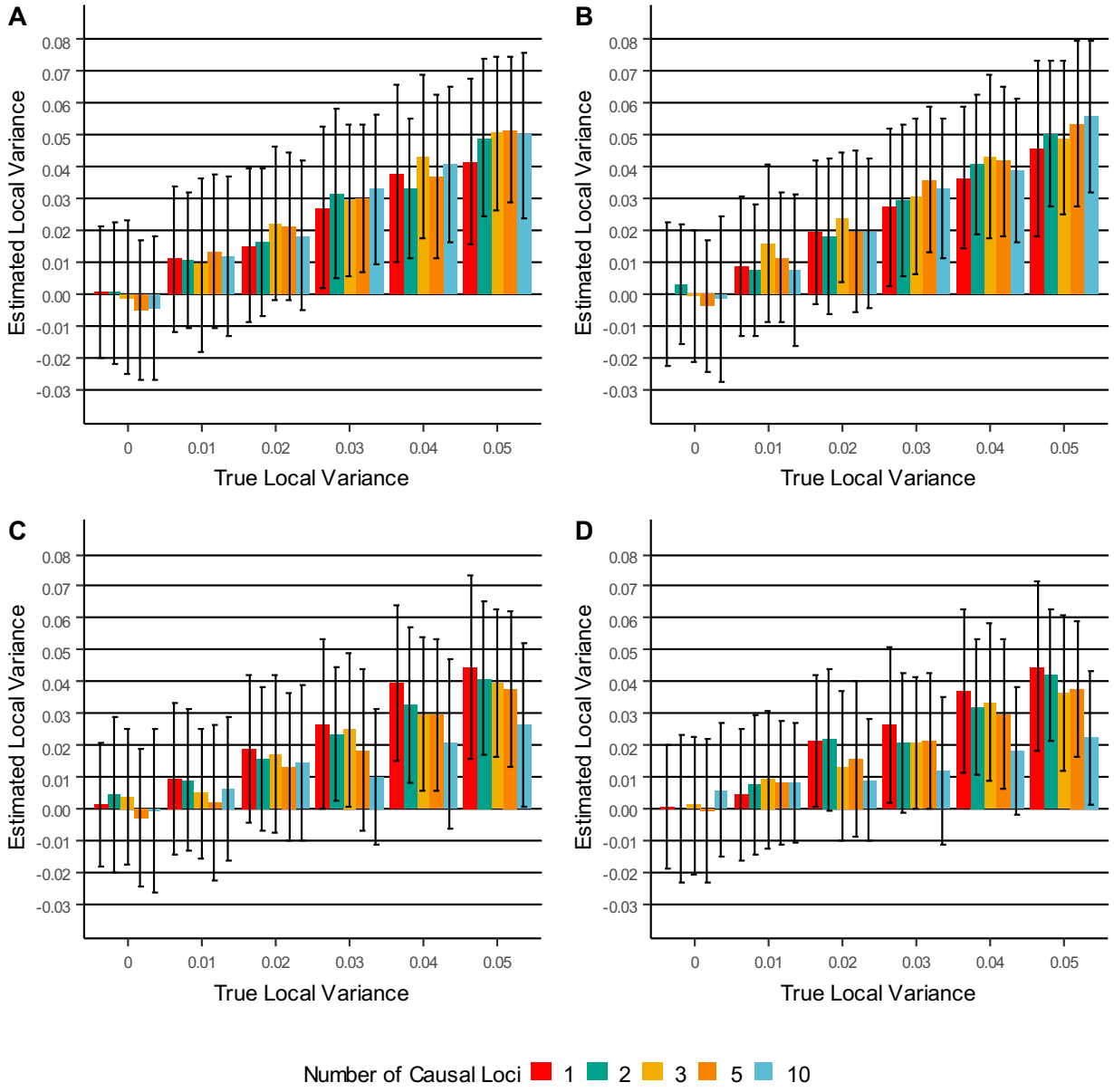


Figure 5-2: Estimated proportion of variance explained by local ancestry under various conditions.

Average (±SD) estimated local variance explained under various simulated conditions. True, unobserved local variances were pre-specified at 0, 0.01, 0.02,

0.03, 0.04 and 0.05 (x-axis) as determined by 1, 2, 3, 5, or 10 causal SNPs. Each bar represents an average of 100 simulations, error bars show ±SD. Panels A and B illustrates simulated conditions with no directional condition, while panels C and D restrict local effects to be greater than 0. Panel A and C illustrate simulated conditions with no global effect, and Panel B and D have a pre-specified global effect.

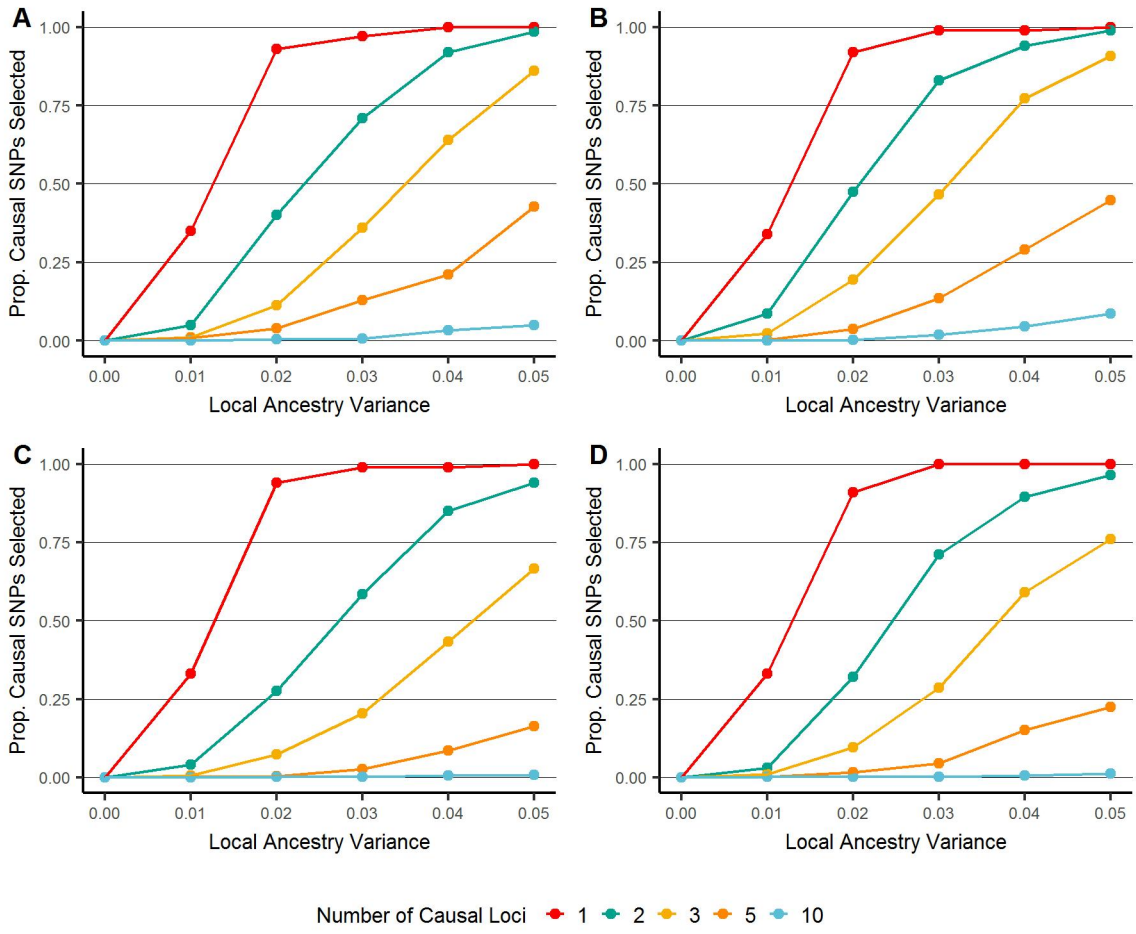


Figure 5-3: Proportion of causal SNPs selected under various conditions.

Proportion of selected SNPs which were causal or regional selected by the forward selection algorithm under various simulated conditions. Proportion of causal SNPs (y-axis) was calculated as: (number of selected causal SNPs + number of selected regional SNPs) / (number of true, unobserved causal SNPs). True, unobserved local variances were pre-specified at 0, 0.01, 0.02, 0.03, 0.04 and 0.05 (x-axis) as determined by 1, 2, 3, 4, or 5 causal SNPs. Panels A and B illustrates simulated conditions with no directional condition, while panels C and D restrict local effects

to be greater than 0. Panel A and C illustrate simulated conditions with no global effect, and Panel B and D have a pre-specified global effect.

5.5.2 Estimation of Effect of Local Ancestry in ORIGIN

The VC and forward selection models tested through simulations were then performed on the 237 ORIGIN biomarkers. A model was built for each biomarker comprising of global Asian and African components in addition to local components selected according to the forward selection algorithm. The proportion of variance attributed to local variance was estimated from the VC model and evaluated and the individual associated loci ($p < 1.13e-06$) identified in the linear model were inspected. VC models revealed 11 biomarkers to have a significant proportion ($p < 0.05/237$) of variance explained by both local ancestries ranging from 0.11 to 0.24 (Table 5-1). The global associations and estimated variance were also evaluated as fixed effects from the VC model (Figure 5-4). We identified 23 and 6 global African and Asian associations, respectively, after adjusting for multiple hypothesis testing ($p < 0.05/237$), representing 0.12 of biomarkers (29/237 biomarkers) (Table 5-2).

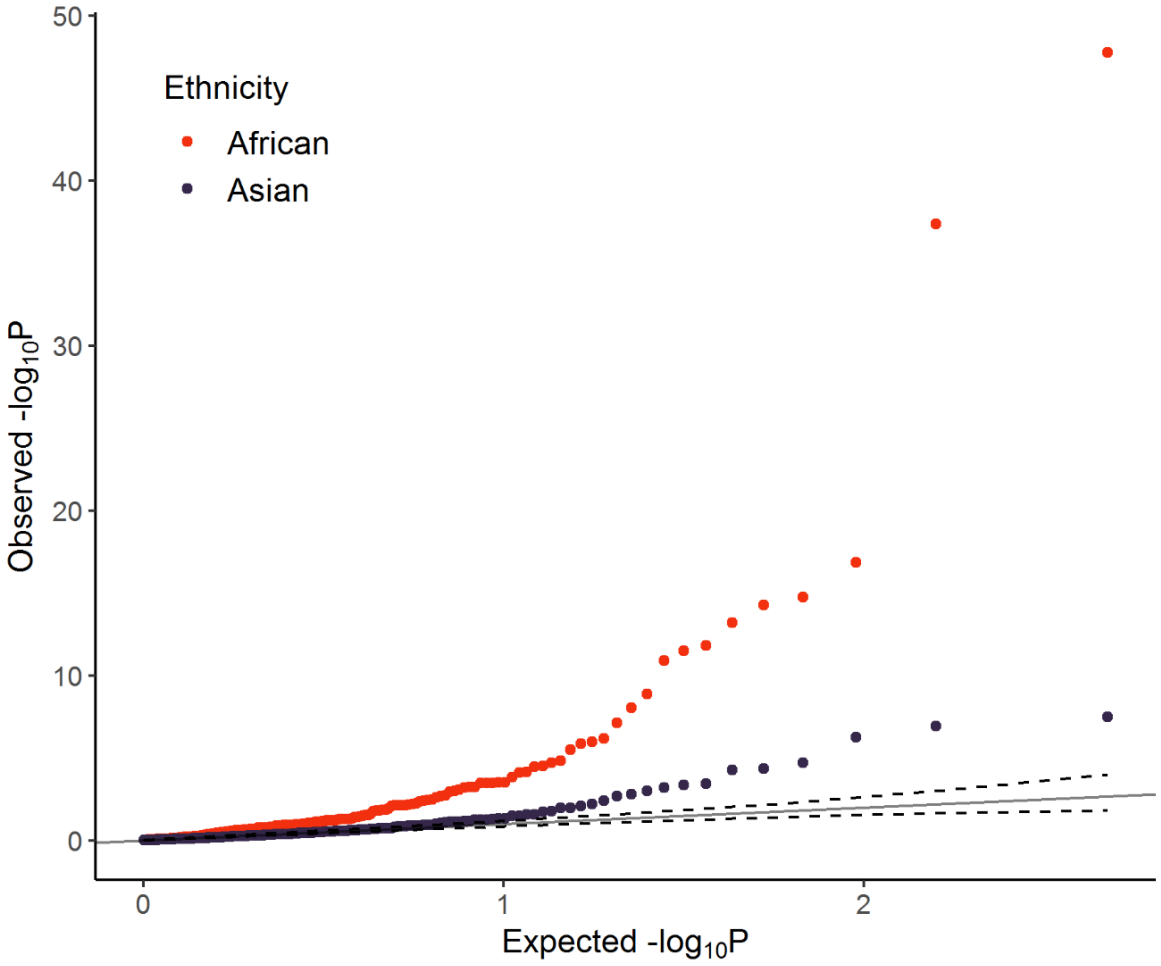


Figure 5-4: Global ancestry QQ plot.

QQ plot of association statistics of the effect of African (red) and Asian (blue) global ancestry on biomarker levels. Variance component models were used to assess the effect of global ancestry, with local ancestry included as random effects. Biomarker levels were first residualized for age and sex.

Table 5-1: Summary of biomarkers with significant proportion of variation explained by local ancestry in using VC analysis ($p < 0.05/237$).

Biomarker	Both		African		Asian	
	Proportion explained (95% CI)	P-value	Proportion explained (95% CI)	P-value	Proportion explained (95% CI)	P-value
C-Peptide	0.24 (0.16, 0.32)	1.6E-11	0.02 (-0.02, 0.06)	9.9E-12	0.22 (0.15, 0.29)	0.13
Eotaxin-3	0.22 (0.14, 0.29)	5.0E-10	0.19 (0.12, 0.25)	0.02	0.03 (0.00, 0.07)	8.1E-09
Clusterin	0.18 (0.11, 0.25)	2.2E-08	0.02 (-0.01, 0.06)	1.6E-08	0.16 (0.10, 0.22)	0.10
Fatty acid-binding protein liver	0.13 (0.07, 0.19)	2.4E-06	0.02 (-0.02, 0.06)	1.3E-06	0.11 (0.06, 0.16)	0.13
Intercellular adhesion molecule-1	0.14 (0.08, 0.21)	3.1E-06	0.06 (0.01, 0.11)	0.00018	0.08 (0.03, 0.13)	0.0042
Apolipoprotein E	0.14 (0.07, 0.20)	4.0E-06	0.03 (-0.01, 0.08)	9.4E-06	0.10 (0.05, 0.15)	0.05
Fas ligand	0.14 (0.08, 0.20)	4.6E-06	0.06 (0.02, 0.11)	0.00030	0.08 (0.03, 0.12)	0.0037
Alpha-2 macroglobulin	0.12 (0.06, 0.18)	1.2E-05	0.09 (0.04, 0.14)	0.038	0.03 (0.00, 0.06)	0.00010
Apolipoprotein A-IV	0.12 (0.06, 0.18)	3.2E-05	0.05 (0.00, 0.09)	0.00039	0.07 (0.03, 0.12)	0.017
Interleukin-2	0.12 (0.06, 0.18)	4.0E-05	0.02 (-0.02, 0.06)	1.1E-05	0.11 (0.06, 0.16)	0.22
Paraoxanase-1	0.11 (0.05, 0.17)	8.7E-05	0.04 (-0.01, 0.08)	0.00028	0.07 (0.03, 0.12)	0.048

Biomarkers were residualized for age and sex.

Table 5-2: Summary of biomarkers with significant global ancestry association for either African or Asian global ancestry ($p < 0.05/237$).

Biomarker	Global African Ancestry		Global Asian Ancestry	
	β (95% CI)	P-value	β (95% CI)	P-value
Kallikrein 5	-0.33 (-0.40, -0.25)	< 5E-10	0.21 (0.02, 0.40)	0.027
Vitronectin	-0.29 (-0.33, -0.25)	< 5E-10	0.07 (-0.45, 0.59)	0.79
Factor VII	-0.24 (-0.27, -0.21)	< 5E-10	0.03 (-0.19, 0.25)	0.81
Insulin-like growth factor binding protein 5	0.17 (0.13, 0.22)	1.8E-15	-0.15 (-0.50, 0.20)	0.40

Immunoglobulin M	0.24 (0.18, 0.30)	5.3E-15	0.00 (-0.31, 0.31)	0.98
Apolipoprotein B	-0.74 (-0.93, -0.55)	6.2E-14	0.04 (-0.33, 0.40)	0.84
Monocyte chemotactic protein 4	0.54 (0.39, 0.69)	1.6E-12	0.08 (0.02, 0.14)	0.011
Interleukin-12 subunit p40	-0.08 (-0.10, -0.06)	3.2E-12	0.01 (-0.22, 0.24)	0.95
Resistin	0.34 (0.24, 0.44)	1.3E-11	0.01 (-0.20, 0.22)	0.95
Hepatocyte growth factor receptor	-0.27 (-0.35, -0.18)	1.3E-09	0.10 (-0.34, 0.55)	0.65
Ficolin-3	-0.13 (-0.17, -0.09)	9.0E-09	-0.19 (-0.48, 0.09)	0.19
Protein S100-A4	0.55 (0.35, 0.75)	7.3E-08	0.07 (-0.04, 0.17)	0.22
Cortisol	-0.41 (-0.58, -0.25)	6.5E-07	-0.13 (-0.38, 0.13)	0.33
6Ckine	-0.27 (-0.38, -0.16)	1.1E-06	-0.18 (-0.39, 0.02)	0.080
Immunoglobulin E	0.37 (0.22, 0.52)	1.4E-06	0.45 (0.16, 0.74)	0.0020
Hepatocyte growth factor	-0.46 (-0.65, -0.26)	3.2E-06	-0.18 (-0.37, 0.02)	0.073
Ferritin	-0.36 (-0.52, -0.20)	1.5E-05	-0.03 (-0.36, 0.30)	0.86
Adrenomedullin	-0.62 (-0.90, -0.33)	2.0E-05	-0.23 (-0.54, 0.07)	0.13
Creatine kinase-MB	-0.61 (-0.89, -0.32)	3.0E-05	-0.10 (-0.48, 0.28)	0.61
Prostatic acid phosphatase	-0.26 (-0.38, -0.14)	3.4E-05	-0.07 (-0.17, 0.02)	0.13
Methylglyoxal	-0.50 (-0.74, -0.25)	6.8E-05	-0.05 (-0.54, 0.44)	0.83
Glucose-6-phosphate isomerase	0.40 (0.20, 0.60)	8.1E-05	0.17 (-0.37, 0.71)	0.54
Sex hormone-binding globulin	0.26 (0.13, 0.40)	0.00015	0.17 (-0.11, 0.45)	0.24
Mesothelin	0.25 (-0.01, 0.52)	0.064	0.22 (0.14, 0.29)	3.2E-08
Thrombospondin-1	0.00 (-0.24, 0.24)	0.99	0.12 (0.08, 0.17)	1.21E-07
Pulmonary and activation-regulated chemokine	0.27 (-0.08, 0.62)	0.13	0.25 (0.16, 0.35)	5.4E-07
T lymphocyte-secreted protein I-309	0.23 (0.05, 0.41)	0.014	-0.07 (-0.11, -0.04)	1.9E-05
Pigment epithelium derived factor	-0.01 (-0.15, 0.14)	0.92	-0.26 (-0.39, -0.14)	4.6E-05
Chemokine CC-4	0.00 (-0.09, 0.09)	0.98	-0.53 (-0.79, -0.27)	5.5E-05

β per SD increase in global ancestry. Global ancestry included as fixed effects in variance component model. Biomarkers were residualized for age and sex.

Using the fixed effect forward selection framework, 0.17 (40/237) of biomarkers were found to have at least one significant local association, 5 of these 40 biomarkers overlapped with the 11 associations identified using VC analysis. A total of 46 local components were associated with these biomarkers (i.e. some biomarkers were associated with more than one local component). Of the 46 local ancestry associations identified, 0.55 (25/46) were in *trans* and 0.45 (21/46) were in *cis* with the gene encoding the corresponding protein or protein component of the biomarker for which an association was found. Five biomarkers investigated were not a direct gene product (e.g. cortisol), and therefore could not have any *cis* associations by this definition. One local association identified with Asian ancestry, on chromosome 14, was with one such biomarker, methylglyoxal (included in the 25 *trans* associations). The number of local associations was similar across ethnicities, with 21 and 25 African and Asian associations, respectively. Replication in ORIGIN Europeans revealed that 33 of these 46 regions had genotype associations at genome-wide significance with their corresponding biomarkers ($p < 5x10^{-8}$), 9 in *trans* and 24 in *cis.* Notably, five biomarkers were significantly associated with rs12075, located in the *ARCK1* gene which encodes Duffy antigen receptor responsible for the Duffy blood group system. Replication in ORIGIN Native Latin participants, revealed an additional *cis* association at genome-wide significant. Therefore, only 34 local associations had no corresponding GW-significant association in either European or Native Latin participants. A summary of local associations and their corresponding genotypic associations in European and Native Latins can be found in supplementary Tables 2-4.

5.5.3 Evaluation of the Role of C-Peptide in Disparities of Diabetes Risk Among Ethnic Groups

Our analysis revealed C-peptide as the most significant biomarker with involvement of ancestry in determining its levels. Specifically, we found that 0.24 (95% CI 0.16 to 0.32, $p=2.6x10^{-11}$) of the variance of C-peptide is due to local ancestry in Latin Americans, largely due to an effect of African ancestry. We also identified two local ancestry components associated with its levels (Figure 5-5). Furthermore, C-peptide has direct medical relevance because of its physiological importance as a marker of insulin production, and also as a biomarker used clinically in the diagnosis and categorization of diabetes into type 1 and 2. Because diabetes risk is well known to vary between ethnicities, we sought to explore the role of C-peptide and genetic ancestry in the context of T2D to further elucidate this relationship. First, we sought to determine the impact of including a glycemic-related weighted genetic risk score (GRS) as a fixed effect in the VC model. Using estimates from public consortia, we tested a GRS for T2D, HBA1C, fasting glucose, fasting insulin, and 2-hour glucose[30,31]. The effect of local ancestry remained significant in all models (data not shown). Next, we evaluated if any SNPs in the in local ancestry window derived using admixture LD were associated with T2D and glycemic traits in DIAGRAM or MAGIC databases[30,31]. After adjusting for multiple hypothesis testing, no significant associations were found. Third, we tested the effect of the two C-peptide local components for an association with baseline T2D, HOMA-IR and HOMA-β in ORIGIN, both separately and using a weighted local ancestry GRS. The GRS was derived using estimates from a linear model including both local components (scaled from 0 to 1), age and sex, and C-peptide levels as the dependent variable. The C-peptide local ancestry GRS was found to be significantly associated with an increased risk of T2D (OR=4.47 per SD, 95% CI 1.70 to 11.76, p=0.002) and HOMA-IR (β=2.73 per SD, 95% CI 0.95 to 4.51, p=0.003), but not with HOMA-β ($p>0.05$) after adjusting for age, sex and global ancestry. Models assessing local associations independently demonstrated that

the African local estimate at rs4149261 was similarly associated with T2D (OR=6.07 per percent increase in African ancestry, 95% CI 1.44 to 25.56, p=0.01) and HOMA-IR (β=2.86, 95% CI 0.93 to 4.80, p=0.004), but not HOMA-β. However, African local estimates at rs3769050 were not associated with T2D, HOMA-IR or HOMA-β (p>0.05). Finally, we investigated if variants previously linked to diabetes in a Mexican population were associated with C-peptide and metabolic traits in ORIGIN. We found that African local ancestry at the *SLC16A11* locus reported by Williams et al. and Hara et al. was inversely associated with C-peptide (β=-0.25 SD per percent increase in ancestry, 95% CI -0.49 to -0.02, p=0.03), fasting plasma glucose (β=-0.80 mmol/L, 95% CI -1.39 to -0.20, p=0.009), and risk of diabetes (OR=4.47, 95% CI 1.70 to 11.76, p=0.002), in models adjusted for age sex and global ancestry[6,7]. These results are inconsistent with the findings in these reports, as Williams et al. identified a risk haplotype common to African populations which increases diabetes risk. Other novel SNPs in these reports showed no association in ORIGIN.

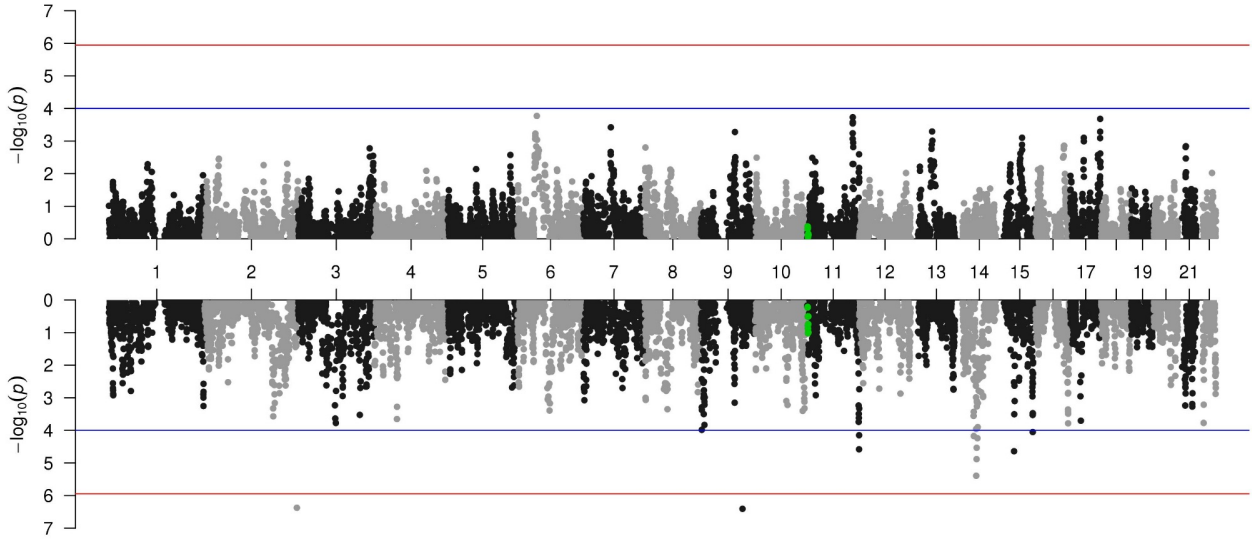


Figure 5-5: Manhattan plot of admixture mapping of C-peptide protein.

Mirror Manhattan plot of the association between local Asian ancestry (top) and local African ancestry (bottom) with C-Peptide using the additive estimated model as determined by the forward selection algorithm. In other words, each point

represents the association between a single local component and C-Peptide, after residualizing for age and sex, adjusted for global African and Asian ancestry and the three significant SNPs falling above (for African) or below (for Asian) the selected threshold (shown in red), based on experiment-wide adjusted p-vale ($1.13x10^{-6}$). Negative log 10 P-values are plotted against each local component's respective position on each chromosome. The blue line corresponds to nominal significance of $p=1x10^{-4}$. Local components in LD ($r^2 > 0.8$) with any SNP within 300KB of the gene encoding c-peptide protein are shown in green.

5.5.4 Prediction of Global Ancestry Using Serum Biomarkers in ORIGIN

The forward-selection multivariate algorithm revealed 31 biomarkers significantly and independently predictive of global ancestry European and Asian components, after adjusting for multiple hypothesis testing ($p < 0.05/237$) according to the multivariate ANOVA test. A multivariable linear model was then computed in the model building set for each global ancestry component using two independent linear models, with global ancestry as the dependent variable and age, sex and the 31 predictive biomarkers as the independent variables. The resulting correlation was 0.69 (0.66, 0.71) and 0.64 (0.62, 0.67) for European and Asian global ancestry, respectively. These models were then tested in model assessment group revealing consistent estimates, with the proportion of variance explained 0.67 (0.63, 0.71) and 0.62 (0.57, 0.66) for European and Asian global ancestry, respectively. Fitted versus actual values can be seen in Figure 5-6.

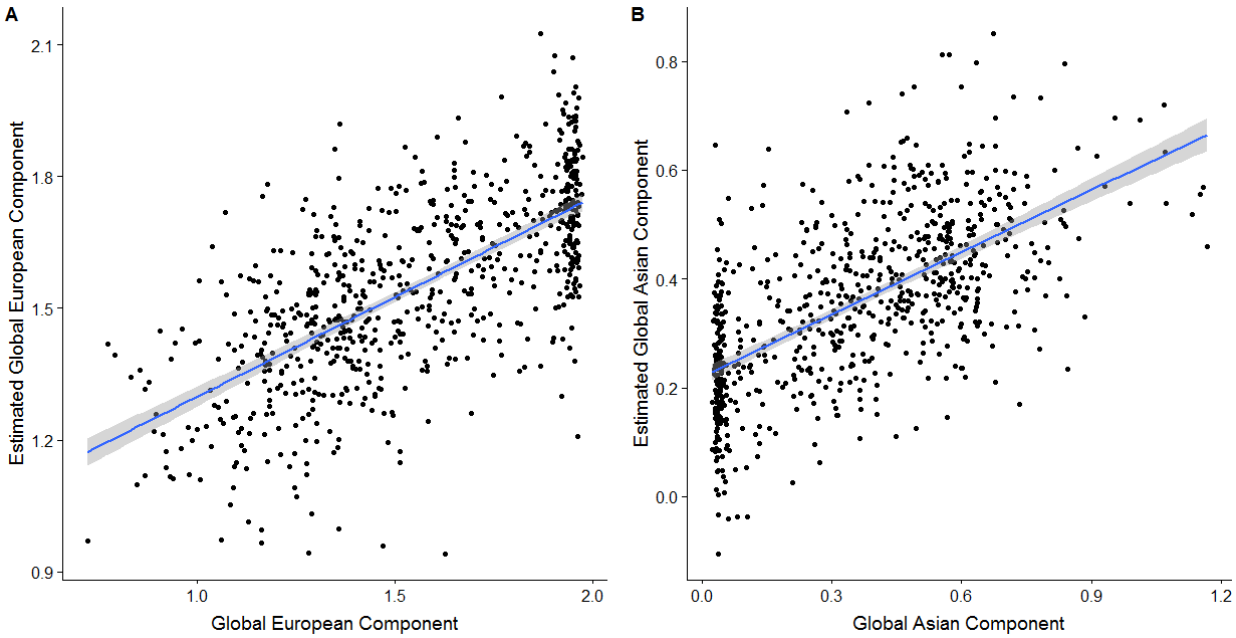


Figure 5-6: Global fitted versus true estimates.

Scatter plot illustrates fitted global components based on biomarker levels versus true global components for European (A) and Asian (B) ancestries. Line of best fit is shown in blue and the corresponding 95% CI is shaded.

5.6 Discussion

Marked regional differences in biomarker profiles have been investigated and previously reported[2]. However, these observations have not been fully elucidated and causes may include genetic, lifestyle or socio-economic factors. In this report, we sought to explore the impact of ancestry on the human proteome and its implications on health and disease. We first developed a model to investigate phenotypic variation among admixed individuals. Through simulations, we show that an unbiased proportion of variance due to local admixture can be elucidated using a VC model and a highly specific forward selection model can be used to reveal causal loci associated with biomarker levels. Using these models, we found that local ancestry affects at least 0.19 (46 of 237) of biomarkers with 0.05 of

biomarkers having more than 0.10 of phenotypic variance explained by local ancestry in Latin Americans. Additionally, 0.12 biomarkers had a significant global association, however, these associations may be confounded due to legacy effects across the entire genome. Local associations, conversely, represent true biological effects and implicate genomic regions involved in phenotypic variance.

C-peptide was identified to have a significant effect of ancestry, almost entirely due to an effect of African ancestry. C-peptide is a well-established clinical biomarker most commonly used to distinguish type 1 and 2 diabetes, among other uses[32]. Indeed, diabetes is well known to exhibit differential risk patterns among ethnic groups, consistent with these findings[3–5]. Similarly, we also identified two African local components to have an effect on C-peptide. A GRS using these local ancestry components was associated with an increased risk of diabetes and increased HOMA-IR. Our findings shed light on the complex influence of ancestry on T2D, as C-peptide is a strong marker of insulin secretion. Together, these findings point to insulin resistance rather than β-cell dysfunction as a mediator of this relationship. While both local components showed that increased levels of C-peptide leads to increased risk of T2D and higher HOMA-IR, the ancestry specific effect on C-peptide levels were inconsistent. In other words, African ancestry increased C-peptide levels at one locus and decreased C-peptide levels at another locus. Therefore, more studies are needed to further resolve the disparity in diabetes risk among ethnic groups.

The genetics of biomarker concentrations have been extensively investigated in the context of genome-wide association studies (GWAS)[34]. Numerous loci have been identified for many biomarkers and these loci have also been linked to disease, suggestive of causal relationships and potential drug targets. However, few studies have leveraged genetic admixture as a complimentary approach to discover novel chromosomal regions impacting biomarker levels. Our analysis revealed 46 regions linked to biomarker concentration, many of which are novel. Specifically, we identified an association of local Asian ancestry ACE levels, and

further genotypic mapping suggests this is an effect at the *ACE* locus. ACE has a well-established role in regulating blood pressure and is also used to diagnose sarcoidosis. Response to ACE-inhibitors has been shown to differ between ethnic groups raising the possibility that current guidelines may not be applicable to non-Caucasian ethnicities[35–37]. Additionally, we identified 5 biomarkers associated with rs12075 within the Duffy antigen receptor gene, which encodes for the glycosylated membrane protein and is non-specific receptor for several cytokines. This gene exhibits known genetic admixture, and variation in this gene are responsible for the Duffy blood group system[38]. The association of the *ARCK1* variant with multiple protein levels replicates previous findings results, substantiating a potential role for DARC in the regulation of serum cytokines[39].

Understanding the impact of genetics on biomarker profiles also has clinical implications. Predictive thresholds for each specific ethnic group are necessary for accurate risk stratification. Otherwise, there is potential for misclassification of risk and inappropriate use of pharmacotherapies. Notably, we found that 0.05 of biomarkers are affected by local ancestry after multiple hypothesis testing, and 0.30 showed nominal significance ($p < 0.05$), ranging from 0.05 to 0.31 proportion of variation due to an effect of ancestry. These findings suggest that these biomarkers harbor true biological inter-ancestry differences in concentration that are genetically determined. These differences may lead to differences in disease risk and clinical diagnosis. We also identified local associations with clinically-relevant biomarkers, including vitamin-D binding protein, apolipoprotein-E, and vascular endothelial growth factor. These results are consistent with previous reports and demonstrate that specific genetic polymorphisms may partially explain the observed differences in concentrations between populations[40,41]. These findings may have implications for the interpretation of clinical markers across different ethnic groups.

A few limitations are worth mentioning. First, for the 11 biomarkers for which we identified a significant effect of local ancestry, we did not identify a specific genetic

locus contributing to the variation of 6 of these biomarkers. These results are consistent with the polygenic model of inheritance, hypothesized to underlie many complex traits. According to this model, a large number of loci of small effect sizes together explain the variation of a single trait, such as a biomarker. If hundreds of genetic variants contribute to the observed differences between Asian and African ancestry for a single biomarker, relative to European ancestry, then the proportion of Asian and African ancestry in Latins will act as a proxy for the overall contribution of variants. However, identification of any specific variant will require an appropriately large sample size. Indeed, our simulation have shown that even with 10 loci the power to detect local associations was very weak, particularly in the presence of directional associations. Second, we identified a local association for 0.17 (40/237) biomarkers, however the VC models identified only 0.05 biomarkers to have a significant effect of ancestry after multiple hypothesis testing. These results suggest that power was limited in our VC analysis compared to the linear model, and larger studies are needed to identify additional markers with an effect of ancestry using VC analysis. Finally, we were not able to identify a significant, corresponding genotype association in either Europeans or Native Latins for all local associations identified. This could be because multiple causal variants account for the local association for which we were underpowered or the causal variant was not well tagged in our study. Likewise, in the Native Latin GWAS, the causal variant(s) could be perfectly correlated with ancestry, and therefore impossible to distinguish from local ancestry itself. It is also worth noting that we did not have access to an African or Asian cohort to assess these genotypic relationships in these ancestries.

Genetically admixed populations provide a powerful model to dissect the contribution of genetics to difference in biomarker concentrations between populations. Studying Latin Americans within the framework of a large, international study, we provide evidence for an effect of genetic ancestry on biomarker variability. Our results show that ancestry has plays a role in the

concentration of at least 0.05 of biomarkers, although this is likely a lower bound. This has many implications, namely that differences in disease prevalence likely has a biological basis in many cases, and that use of reference intervals for those biomarkers should be tailored to ancestry. These results highlight the need for specific cutoff values and prognostic measures for each ethnicity and implemented accordingly. Finally, we also show that some loci appear to have pleiotropic ancestry effects and therefore appear to be of particular importance. As serum proteins are frequently dysregulated in disease, identification of factors that determine protein variability is a clinical priority. These findings shed light on the contribution of ancestry in disease and pave the way for better informed, ethnic specific defined cut-offs. Further research will be needed to identify specific factors responsible for these differences and gain a better understanding of underlying biological mechanisms.

5.7 Concluding Remarks

5.7.1 Conflict of Interest

H.C.G. has received consulting fees from Sanofi, Novo Nordisk, Lilly, AstraZeneca, Boehringer Ingelheim, and GlaxoSmith-Kline and support for research or continuing education through his institution from Sanofi, Lilly, Takeda, Novo Nordisk, Boehringer Ingelheim, and AstraZeneca. G.P. has received consulting fees from Sanofi, Bristol Myers Squibb, Lexicomp, and Amgen and support for research through his institution from Sanofi. Dr. Hess is an employee of Sanofi. S.Y. has received research support for ORIGIN from Sanofi through his institution. D.M. and S.A. report no conflicts.

5.7.2 Funding

The ORIGIN trial and biomarker project were supported by Sanofi and CIHR (award 125794). H.C.G., receiving support from Population Health Institute Chair

in Diabetes Research and Care; S.A., receiving support from Canada Research Chair in Ethnicity and Cardiovascular Disease, Michael G. DeGroote Chair in Population Health; D.M.; receiving support from Tier 2 Canada Research Chair in Genetics of Obesity; S.Y., receiving support from Heart and Stroke Foundation of Ontario/MarionW. Burke Chair in Cardiovascular Disease; G.P., receiving support from Canada Research Chair in Genetic and Molecular Epidemiology, CISCO Professorship in Integrated Health Systems.

5.8 References

1. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin. Pharmacol. Ther.* **69,** 89–95 (2001).

2. Tahmasebi, H., Trajcevski, K., Higgins, V. & Adeli, K. Influence of ethnicity on population reference values for biochemical markers. *Crit. Rev. Clin. Lab. Sci.* **55,** 359–375 (2018).

3. Florez, J. C. *et al.* Strong association of socioeconomic status with genetic ancestry in Latinos: implications for admixture studies of type 2 diabetes. *Diabetologia* **52,** 1528–36 (2009).

4. Spanakis, E. K. & Golden, S. H. Race/ethnic difference in diabetes and diabetic complications. *Curr. Diab. Rep.* **13,** 814–23 (2013).

5. Knowler, W. C., Pettitt, D. J., Saad, M. F. & Bennett, P. H. Diabetes mellitus in the Pima Indians: incidence, risk factors and pathogenesis. *Diabetes. Metab. Rev.* **6,** 1–27 (1990).

6. Hara, K. *et al.* Genome-wide association study identifies three novel loci for type 2 diabetes. *Hum. Mol. Genet.* **23,** 239–246 (2014).

7. Williams, A. L. *et al.* Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. *Nature* **506,** 97–101 (2013).

8. Gijsberts, C. M. *et al.* Biomarkers of Coronary Artery Disease Differ Between Asians and Caucasians in the General Population. *Glob. Heart* **10,** 301–311.e11 (2015).

9. Morimoto, Y. *et al.* Ethnic differences in serum adipokine and C-reactive protein levels: the multiethnic cohort. *Int. J. Obes. (Lond).* **38,** 1416–22 (2014).

10. Khan, U. I. *et al.* Race-ethnic differences in adipokine levels: the Study of Women's Health Across the Nation (SWAN). *Metabolism.* **61,** 1261–9 (2012).

11. Talib, H. J., Ponnapakkam, T., Gensure, R., Cohen, H. W. & Coupey, S. M. Treatment of Vitamin D Deficiency in Predominantly Hispanic and Black Adolescents: A Randomized Clinical Trial. *J. Pediatr.* **170,** 266–72.e1 (2016).

12. Nielson, C. M. *et al.* Role of Assay Type in Determining Free 25-Hydroxyvitamin D Levels in Diverse Populations. *N. Engl. J. Med.* **374,** 1695–6 (2016).

13. Lim, E., Miyamura, J. & Chen, J. J. Racial/Ethnic-Specific Reference Intervals for Common Laboratory Tests: A Comparison among Asians, Blacks, Hispanics, and White. *Hawaii. J. Med. Public Health* **74,** 302–10 (2015).

14. Sankararaman, S., Sridhar, S., Kimmel, G. & Halperin, E. Estimating local ancestry in admixed populations. *Am. J. Hum. Genet.* **82,** 290–303 (2008).

15. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *Am. J. Hum. Genet.* **93,** 278–288 (2013).

16. Zhu, X. *et al.* Admixture mapping for hypertension loci with genome-scan markers. *Nat. Genet.* **37,** 177–181 (2005).

17. Freedman, M. L. *et al.* Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 14068–14073 (2006).

18. Kopp, J. B. *et al.* MYH9 is a major-effect risk gene for focal segmental glomerulosclerosis. *Nat. Genet.* **40,** 1175–1184 (2008).

19. Brown, L. A. *et al.* Admixture Mapping Identifies an Amerindian Ancestry Locus Associated with Albuminuria in Hispanics in the United States. *J. Am. Soc. Nephrol.* ASN.2016091010 (2017). doi:10.1681/ASN.2016091010

20. Shendre, A. *et al.* Admixture Mapping of Subclinical Atherosclerosis and Subsequent Clinical Events Among African Americans in 2 Large Cohort Studies. *Circ. Cardiovasc. Genet.* **10,** e001569 (2017).

21. Gerstein, H. C., Yusuf, S., Riddle, M. C., Ryden, L. & Bosch, J. Rationale, design, and baseline characteristics for a large international trial of cardiovascular disease prevention in people with dysglycemia: The ORIGIN Trial (Outcome Reduction with an Initial Glargine Intervention). *Am. Heart J.* **155,** 26. e1-26. e13 (2008).

22. Gerstein, H. C. *et al.* Identifying novel biomarkers for cardiovascular events or death in people with dysglycemia. *Circulation* **132,** 2297–2304 (2015).

23. 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).

24. 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).

25. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

26. 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).

27. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5,** e1000529 (2009).

28. Browning, S. R. & Browning, B. L. Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.* **12,** 703–14 (2011).

29. Covarrubias-Pazaran, G. Genome-Assisted prediction of quantitative traits using the r package sommer. *PLoS One* **11,** 1–15 (2016).

30. Scott, R. A. *et al.* An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. *Diabetes* **66,** db161253 (2017).

31. Scott, R. A. *et al.* Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nat. Genet.* **45,** 1274–83 (2012).

32. Jones, A. G. & Hattersley, A. T. The clinical utility of C-peptide measurement in the care of patients with diabetes. *Diabet. Med.* **30,** 803–817 (2013).

33. Larsen, P. B., Linneberg, A., Hansen, T. & Friis-Hansen, L. Reference intervals for C-peptide and insulin derived from a general adult Danish population. *Clin. Biochem.* **50,** 408–413 (2017).

34. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558,** 73–79 (2018).

35. Ferdinand, K. C. & Armani, A. M. The Management of Hypertension in African Americans. *Crit. Pathways Cardiol. A J. Evidence-Based Med.* **6,** 67–71 (2007).

36. Cohn, J. N. *et al.* Clinical experience with perindopril in African-American hypertensive patients: a large United States community trial. *Am. J. Hypertens.* **17,** 134–8 (2004).

37. ALLHAT Officers and Coordinators for the ALLHAT Collaborative Research Group. The Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial. Major outcomes in high-risk hypertensive patients randomized to

angiotensin-converting enzyme inhibitor or calcium channel blocker vs diuretic: The Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT). *JAMA* **288,** 2981–97 (2002).

38. Howes, R. E. *et al.* The global distribution of the Duffy blood group. *Nat. Commun.* **2,** (2011).

39. Voruganti, V. S. *et al.* Genome-wide association replicates the association of Duffy antigen receptor for chemokines (DARC) polymorphisms with serum monocyte chemoattractant protein-1 (MCP-1) levels in Hispanic children. *Cytokine* **60,** 634–8 (2012).

40. Powe, C. E. *et al.* Vitamin D-binding protein and vitamin D status of black Americans and white Americans. *N. Engl. J. Med.* **369,** 1991–2000 (2013).

41. Braithwaite, V. S. *et al.* Vitamin D binding protein genotype is associated with plasma 25OHD concentration in West African children. *Bone* **74,** 166–170 (2015).

6 CONCLUSION

6.1 General Overview

Identification of biological mediators of disease is a valuable research endeavor. Biomarkers are widely used clinical measures as they can be measured both cost-effectively and non-invasively. These measurements provide the opportunity to identify subclinical disease states before the development of disease and apply preventative measures, facilitate research and understanding of disease mechanisms, and allow for the assessment of therapeutic measures. However, due to innate complexities of many chronic diseases, a single biomarker is often unsuitable to both estimate risk and serve as a valid pharmacological target.

Indeed, the pathogenesis behind chronic diseases results from a number of modifiable and unmodifiable risk factors and is caused by an interaction of many biomarkers simultaneously. Therefore, a major challenge in the field of biomarker research is to discern cause and effect, and as a result, association has often been mistaken for causation. Furthermore, biomarker levels show marked differences between ethnic groups, and it is difficult to distinguish whether this is a result of environmental or genetic factors. Longitudinal, genetic, multi-biomarker studies, offer a valid approach to ameliorate these challenges. Cumulatively, the thesis addresses how advancements in genetic and biomarker research may help to gain novel insights into both known and novel biomarkers of cardiovascular disease, inform and guide clinical decision-making and validate potential disease target pathways. This section briefly summarizes the main research findings, the clinical implications, biologic significance, current challenges and future directions of biomarker research for cardiovascular disease.

6.2 Chapter 3 Summary

Mendelian randomization (MR) was employed to identify novel, causal mediators of coronary artery disease in the ORIGIN-trial. The MR analysis revealed six biomarkers to be causally associated with CAD. Four of these biomarkers (lipoprotein(a), interleukin-6 receptor and apolipoprotein E, apolipoprotein C3) had been previously linked to CAD, however, blood colony stimulating factor 1 (CSF1) and stromal cell derived factor 1 (CXCL12) were established as novel CAD markers. These MR results were then corroborated through epidemiological association of CSF1 and CXCL12 levels with prospective MACE in ORIGIN. Furthermore, analysis in the large UKBiobank cohort revealed that genetically elevated CSF1 and CXCL12 were also associated with increased risk of CAD. Both biomarkers have been previously linked to inflammatory processes characteristic of atherosclerosis and consistent with previous reports, including results from the CANTOS study, showing that an intervention aimed at decreasing inflammation

through interleukin-1 beta inhibition can lead to lower rates of recurrent cardiovascular events. Together, these results support a role for CXCL12 and CSF1 for both risk stratification and as therapeutic targets.

6.3 Chapter 4 Summary

MR was used to reveal new mediators of chronic kidney disease (CKD) in the ORIGIN-trial. Human epidermal growth factor receptor 2 (HER2) and uromodulin (UMOD) were both identified as causal mediators of CKD. Further MR exploration of the HER2 pathway also revealed ACE as a regulator of HER2 levels. These findings were then corroborated in epidemiological analyses using blood HER2, incident CKD and BP-lowering medication data in ORIGIN. These results implicate HER2 as a mediator of ACE-inhibitors' protective effect on CKD and as a marker which may help reveal patients likely to benefit from ACE-inhibition. Furthermore, these findings suggest HER2-inhibition as a potential novel treatment for CKD, which may be applied through the use of HER2-inhibitors (e.g. gefitinib). Additional exploration of UMOD concentration in an independent sample of healthy nephrectomy donors found a nearly halving of plasma UMOD after transplant, compared with before. Therefore, in addition to its causal effect on CKD revealed in ORIGIN, UMOD also represents a blood biomarker of kidney mass. In summary, UMOD and HER2 were found to be causally involved in CKD and potential therapeutic targets for CKD prevention.

6.4 Chapter 5 Summary

Admixture mapping was used to explore the impact of ancestry on a comprehensive panel of 237 biomarkers in 2,216 Latin American participants within the ORIGIN-trial. We first determined the proportion of European, Asian and African ancestry of each participant using genotypes and global ancestry was obtained for each individual by averaging the ancestry across the genome. Next, we determined the proportion of variance explained by these local ancestral

differences using variance component models, revealed 11 biomarkers to have a significant effect of ancestry, after adjusting for multiple hypothesis testing ($p < 0.05/237$). Among these findings, included clinically relevant markers such as C-peptide, apolipoprotein-E and intercellular adhesion molecule 1. Additionally, multivariable linear regression was used to identify specific genetic regions associated with biomarker levels, whereby 46 regional associations including 40 different biomarkers were identified. An independent analysis revealed 34 of these 46 regions were associated at genome-wide significance ($p < 5x10^{-8}$) with their respective biomarker in Europeans within ORIGIN. These results demonstrate an importance of ancestry on determining biomarker levels. This has many implications, namely that differences in disease prevalence likely has a biological basis in many cases, and that use of reference intervals for those biomarkers should be tailored to ancestry.

6.5 Clinical and Research Implications

6.5.1 Novel Drug Targets

In our analysis investigating novel mediators of CAD, we identified both CSF1 and CXCL12 as two new causal markers of CAD. Further analysis using epidemiological models and genetically predicted biomarker levels, confirmed their deleterious effects[1]. These results position both CSF1 and CXCL12 as potential therapeutic targets, suggesting that pharmacological inhibition and subsequent lowering of blood levels would function to decrease risk of CAD. However, these findings should be explored further. Future research should be aimed at identifying the causal mechanisms behind these observations and if interventions which reduce CSF1 and CXL12 levels can reduce CAD.

Similarly, we established both HER2 and UMOD as mediators of CKD and potential therapeutic targets for CKD treatment[2]. HER2-inhibitors have been designed and tested previously, and are currently used in cancer treatment[3,4]. These

interventions may be extended to CKD treatment following proper evaluation through a well-designed RCT. Therefore, future studies should investigate the impact of HER2 inhibitors to reduce the risk of CKD and also explore the effects of lowering UMOD and its impact on CKD.

6.5.2 Inform Treatment Decisions

Our MR analysis investigating novel markers for CKD, revealed important insights into the mechanism by which ACE-inhibitors exert their protective effect on CKD. Upon identifying HER2 as a novel CKD mediator, subsequent MR-analyses revealed ACE as a regulator of HER2 levels. Further epidemiological analyses using BP-lowering medication data show that blockade of RAAS through ACE-inhibitors and ARBS reduces HER2. These findings not only implicate HER2 as a potential mediator of ACE-inhibitors' known protective effect on CKD, but also as a marker which may be able to guide RAAS inhibition. Specifically, these results suggest that individuals with high HER2 levels may be more likely to benefit from RAAS inhibition. However, these findings should be tested in future studies to confirm these observations. A well designed RCT which randomizes individuals to HER2-guided RAAS inhibition versus standard care would shed light on the clinical utility of HER2 to inform RAAS inhibitor treatment.

6.5.3 Validate Targeting of Known Pathways

Our MR analysis of the effects of biomarkers on CAD demonstrated that lipoprotein(a), interleukin-6 receptor and apolipoprotein E, and apolipoprotein C3 are important actors in CAD, consistent with previous reports[5–8]. These findings support investigating these markers for use therapeutically to reduce risk of CAD. Similarly, we identified CSF1 and CXCL12 as novel biomarkers involved in CAD development. Both biomarkers are involved in inflammation and macrophage proliferation and survival. Notably, the CANTOS trial has recently shown that an intervention aimed at decreasing inflammation leads to lower rates of recurrent

cardiovascular events, consistent with our findings in ORIGIN[9]. Furthermore, our findings point to CSF1 as a potential mediator of the beneficial effect of canakinumab on CAD shown in CANTOS. These results validate the importance of inflammation in CAD, among other previously implicated pathways. Given this compelling evidence presented here and in previous studies, these biomarkers should be explored further to see if pharmacological modification of their levels is a safe and effective way to reduce CAD.

6.5.4 Clinical Interpretation of Laboratory Results

Our admixture mapping analysis in Native Latins has revealed the importance of ancestry in biomarker variability and more generally, in health and disease. These results have both research and clinical implications. Our findings suggest that predictive thresholds for each specific ethnic group are necessary in many cases, for accurate risk stratification. Otherwise, there is potential for misclassification of risk and inappropriate use of pharmacotherapies. Notably, we found that 0.05 of biomarkers are affected by local ancestry after multiple hypothesis testing, and 0.30 showed nominal significance ($p < 0.05$). These differences may lead to differences in disease risk, clinical diagnosis and response to medical interventions. These findings suggest that these biomarkers harbor true biological inter-ancestry differences in concentration that are genetically determined. These results pave the way for better informed, ethnic specific defined cut-offs. Further research will be needed to identify specific factors responsible for these differences and gain a better understanding of underlying biological mechanisms.

6.5.5 Extension to Other Diseases and Biomarkers

The statistical approaches and tools employed in this thesis have revealed important insights into the role of biomarkers in health and disease. However, these methods can easily be extended to other data sets, biomarkers and diseases for further biomarker discovery, validation and scientific insights. Indeed, additional

biomarker measurement is currently ongoing within the ORIGIN dataset and other databases under the supervision of Dr. Pare. This thesis has laid the groundwork for many future projects, paving the way for new, exciting scientific findings. Our work in admixed individuals may be applied to other admixed populations in an effort to determine the impact of ancestry on marked ethnic differences among both clinical phenotypes and biomarker levels. Additionally, our biomarker discovery framework using MR has proven to be a valuable technique to agnostically screen a comprehensive panel of markers for causal evidence with a disease, which can be performed in other studies where genotypes have been measured in addition to a biomarker panel.

6.6 Limitations and Considerations

MR provides a unique tool for assessing causality, however the analysis requires that the following assumptions be met: (1) the instrument must be associated with the exposure of interest, (2) the instrument must not be associated with confounding factors in the exposure-outcome association, and (3) the genetic variant must only affect the outcome through the exposure variable[10]. If these assumption are not met, there are potential limitations that threaten the validity of the results. First, is the issue of pleiotropy, whereby a genetic instrument has effects beyond its effect on the exposure of interest. In the presence of pleiotropy, both false positive and false negative results can occur. Confounding due to pleiotropy is least likely when genetic instruments are used that lie near the gene for the exposure under study[11]. Second, linkage disequilibrium (LD) can result in misinterpretations of MR results, similar to pleiotropic bias[12]. For example, a SNP affecting the expression of gene A may be in LD with a SNP affecting the expression of gene B. If biomarker B, encoded by gene B, exerts a causal effect on the disease, then a MR analysis investigating the effect of the biomarker produced by gene A on a disease could result in false positive findings. In this scenario, biomarker B is a causal factor, while biomarker A is merely a bystander

with no causal effect[13]. This phenomenon is especially problematic in gene clusters, as it is often impossible to distinguish the true, causal biomarker unless all biomarkers for each gene in the cluster are measured. Third, population stratification can bias results if not properly taken into consideration. For instance, if differing genotype frequencies and risk of disease exists in ethnic subpopulations, false positive findings may occur if the prevalence of the variant allele parallels the incidence of the study outcomes[14]. Fourth, MR may be confounded by canalization, which is a developmental compensation where a phenotype is selected despite the genetic variability. Finally, limited statistical power can influence the ability of a MR study to inform on a causal relationship between an exposure and an outcome. Specifically, the effect of SNP on a phenotype (both exposure and outcome) can be difficult to ascertain. There are usually multiple genetic and environmental factors influencing the variability of a trait and consequently, the effect of a single SNP can be very small. Additionally, risk factors often act together to exert their effect on a disease such that a causal biomarker may only be responsible for a portion of the resulting outcome.

Based on these aforementioned limitations, there are some issues to address in our Mendelian randomization analyses. The genetic variants studied may have other effects beyond its effect on the biomarker being studied (i.e. genetic pleiotropy). However, we mitigated this source of bias by limiting our investigation to variants at or near the gene coding for the biomarker of interest. Furthermore, each novel, causal marker was individually inspected for proximity to other relevant genes which may have biased our results and none were identified for the causal biomarkers presented in this thesis. We also employed sensitivity analyses where possible to assess the presence of pleiotropy (MR-Egger), which were non-significant in all cases[15]. There may also be issues of statistical power. Although we can be confident in the associations we did find, we cannot rule out a causal role of other biomarkers investigated. For example, we did not detect a significant association between ApoB and CAD after adjusting for multiple hypothesis testing

in our MR analysis (p=0.029), although this relationship has been seen and replicated in other, larger, MR studies[16]. One possibility for this observation is our two sample MR study design, where genetic estimates were obtained from independent populations. In such a design, weak genetic instruments results in estimates which are biased toward the null hypothesis, reducing the likelihood of type 1 error, and as a consequence, decreasing power[17].

There are also limitations to the admixure analysis that are worth mentioning. While we identified 46 genetic loci associated with 40 serum biomarkers, we did not identify a specific genetic locus contributing to the variation of 197 biomarkers despite having identified a significant local association for 11 of these using the VC model. These results are consistent with the polygenic model of inheritance, hypothesized to underlie many complex traits. According to this model, a large number of loci of small effect sizes together explain the heritability of a single trait, such as a biomarker or disease. If hundreds of genetic variants contribute to the observed differences between Asian and African ancestry for a single biomarker, relative to European ancestry, then the proportion of Asian and African ancestry in Latins will act as a proxy for the overall contribution of variants. However, identification of any specific variant will require an appropriately large sample size. Additionally, admixture studies are limited by low resolution in comparison to genotype studies. This is due to the fact that genetic studies indirectly measure recombination back to the most recent common ancestor. Because admixed populations are a relatively recent (<20 generations), the resolution of admixture mapping is inferior to genome-wide association studies (GWAS)[18]. We were therefore unable to pinpoint the causal gene involved in the 46 regional associations identified.

Finally, all analyses presented in this thesis were limited to the biomarkers included on the customized Human Discovery Multi-Analyte Profile (MAP) 250+ panel on the LUMINEX 100/200 platforms measured by Myriad RBM Inc. Indeed,

other biomarkers not included here may be causally relevant in CAD and CKD and also harbor true ancestry-dependent biological differences, not due to confounding. As technology continues to advance and measurement of larger panels becomes more cost effective, larger, more comprehensive panels will shed additional light on these research questions. Additionally, these findings were based on analyses of people with moderate degrees of dysglycemia at baseline and may not apply to normoglycemic individuals, younger individuals, or those at low risk for CV outcomes. However, MR results were generated using public consortia in addition to the ORIGIN data, which was not limited to dysglycemic individuals and therefore the results extend beyond dysglycemia to a more generalizable population. Additionally, because genetic variants are determined at birth, they are immune to reverse causation biases that easily plague observational studies, limiting their generalizability. Furthermore, our successful replication in UKB, in the case of the CAD MR, provides strong evidence for application of these findings beyond a dysglycemic population as the UKB cohort represents a healthy population from the UK.

6.7 Conclusion

Using two major statistical methods employed in genetic epidemiology we have revealed important insights into the role of biomarkers in health and disease. Taken together, this thesis implicates new biomarkers for CAD and CKD and also indicates the importance of ancestry in disease and paves the way for better research in complex admixed populations using our innovative approach to determine the impact of ancestry on human phenotypes. MR analysis identified both known and novel pathways for diseases and revealed individuals which are likely to benefit therapeutic treatment and can be easily applied in clinical settings. The admixture mapping analyses presented in this thesis have shown ancestry to have an important role in biomarker levels beyond confounding factors and also identified specific markers which should be re-examined regarding their clinical use

in non-European individuals. Despite these findings, more work is needed to further elucidate the pathophysiology and mechanisms behind these observations. However, with additional research, the results presented here have the potential to guide, inform and transform clinical practice, particularly in the context of cardiovascular disease.

6.8 References

1. Sjaarda, J. *et al.* Blood CSF1 and CXCL12 as Causal Mediators of Coronary Artery Disease. *J. Am. Coll. Cardiol.* **72,** 300–310 (2018).

2. Sjaarda, J. *et al.* Blood HER2 and Uromodulin as Causal Mediators of CKD. *J. Am. Soc. Nephrol.* **29,** 1326–1335 (2018).

3. Govindan, R. A review of epidermal growth factor receptor/HER2 inhibitors in the treatment of patients with non-small-cell lung cancer. *Clin. Lung Cancer* **11,** 8–12 (2010).

4. Schroeder, R. L., Stevens, C. L. & Sridhar, J. Small molecule tyrosine kinase inhibitors of ErbB2/HER2/Neu in the treatment of aggressive breast cancer. *Molecules* **19,** 15196–15212 (2014).

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

6. Sarwar, N. & Butterworth, A. S. Interleukin-6 receptor pathways in coronary heart disease: A collaborative meta-analysis of 82 studies. *Lancet* **379,** 1205–1213 (2012).

7. Bennet, A. *et al.* Association of apolipoprotein E genotypes with lipid levels and coronary risk. *JAMA* **298,** 1300–1311 (2007).

8. TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung, and B. I. *et al.* Loss-of-function mutations in APOC3, triglycerides, and coronary disease. *N. Engl. J. Med.* **371,** 22–31 (2014).

9. Ridker, P. M. *et al.* Antiinflammatory Therapy with Canakinumab for Atherosclerotic Disease. *N. Engl. J. Med.* **377,** 1119–1131 (2017).

10. Bennett, D. A. & Holmes, M. V. Mendelian randomisation in cardiovascular research: an introduction for clinicians. *Heart* heartjnl-2016-310605 (2017). doi:10.1136/heartjnl-2016-310605

11. Hemani, G., Bowden, J. & Davey Smith, G. Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Hum. Mol. Genet.* **0,** ddy163 (2018).

12. Ardlie, K. G., Kruglyak, L. & Seielstad, M. Patterns of linkage disequilibrium in the human genome. *Nat. Rev. Genet.* **3,** 299–309 (2002).

13. Vanderweele, T. J., Tchetgen Tchetgen, E. J., Cornelis, M. & Kraft, P. Methodological challenges in Mendelian randomization. *Epidemiology* **25,** 427–435 (2014).

14. Hellwege, J. N. *et al.* Genome-wide family-based linkage analysis of exome chip variants and cardiometabolic risk. *Genet. Epidemiol.* **38,** 345–52 (2014).

15. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44,** 512–25 (2015).

16. Tragante, V. *et al.* Harnessing publicly available genetic data to prioritize lipid modifying therapeutic targets for prevention of coronary heart disease based on dysglycemic risk. *Hum. Genet.* **135,** 453–467 (2016).

17. Inoue, A. & Solon, G. Two-Sample Instrumental variables estimators. *Rev. Econ. Stat.* **92,** 557–561 (2010).

18. Shriner, D. Overview of Admixture Mapping. *Curr. Protoc. Hum. Genet.* **94,** 1.23.1-1.23.8 (2017).

SUPPLEMENTARY MATERIALS

Supplementary Appendix for Chapter 3

Identification of blood CSF1 and CXCL12 as causal mediators of coronary artery disease using Mendelian randomization in the ORIGIN trial

Biomarker Assay Methodology in ORIGIN

At Myriad RBM Inc., the samples were thawed at room temperature (RT), vortexed, spun at 13,000g for 5 minutes for clarification. An aliquot was removed into a master microtiter plate for analysis. Using automated pipetting, an aliquot of each sample was introduced into one of the capture microsphere multiplexes of the Human DiscoveryMAP. The mixtures of sample and capture microspheres were thoroughly mixed and incubated at RT for 1 hour. Next, multiplexed cocktails of biotinylated reporter antibodies for each multiplex were added robotically. After thorough mixing, they were incubated for an additional hour at RT. Multiplexes were developed using an excess of streptavidin-phycoerythrin solution which was thoroughly mixed into each multiplex and incubated for 1 hour at RT. The volume of each multiplexed reaction was reduced by vacuum filtration and then increased by dilution into matrix buffer for analysis. Analysis was performed in Luminex 100 and 200 instruments and the resulting data stream was interpreted using proprietary data analysis software developed at RBM. For each multiplex, both calibrators and controls were included on each microtiter plate. Eight-point calibrators were run in the first and last column of each plate and 3-level quality controls were included in duplicate. Testing results were determined first for the high, medium and low controls for each multiplex to ensure proper assay performance. Unknown values for each of the analytes localized in a specific multiplex were determined using 4 and 5 parameter, weighted and non-weighted curve fitting algorithms included in the data analysis package.

Determining the Distribution of Each Biomarker in ORIGIN

Biomarkers were scrutinized in 5 steps. First, 26 biomarkers with undetectable levels in > 8409 (i.e. 99%) participants were excluded from further analyses. Second, scrutiny of the mean, median, and distribution of results and biologic literature pertaining to another 64 biomarkers with undetectable levels in > 1000 (i.e. 12%) participants led to exclusion of a further 21, leaving 237 biomarkers for analysis. Third, those biomarkers with levels below the level of quantification in < 10% of participants (n=850) were assigned a level corresponding to the lower limit of quantification. Fourth, biomarkers with levels below the level of quantification in > 10% of participants were identified and analyzed as ordinal variables as follows. A level of 1 was assigned to the participants with unquantifiable levels, dividing the remaining participants into 4 groups using quartiles. Values of 2, 3, 4 and 5 were assigned to participants within each progressively higher group. This approach was used to manage skewed biomarker distributions. Fifth, biomarkers with levels above the level of quantification were identified and those affected were assigned a level 1% above the upper limit of quantification. This approach led to 192 biomarkers for analysis as continuous variables and 45 biomarkers for analysis as 5-level ordinal variables.

The distributions of each of the 192 continuous biomarkers' levels were then scrutinized to identify extreme outliers with levels more than 4 standard deviations above or below the mean; levels that met those criteria were assigned the value corresponding to the mean plus or minus the 4th standard deviation respectively. Subsequently, the levels of 125 biomarkers with distributions that were not normally distributed were log-transformed using the natural logarithm. Finally, data from 93 participants in whom all 237 biomarkers were not analyzed due to insufficient volume of serum were excluded.

Genotyping Quality Control and Imputation in ORIGIN

SNPs were excluded on the basis of low call rate (<99%), deviation from Hardy-Weinberg ($p<10^{-6}$), and low minor allele frequency (MAF<0.01). Samples with low call rates (<99%), sex or ethnicity mismatches, or cryptic relatedness were also removed. We also removed ethnicities with small sample sizes (n<500). All quality control steps were performed using PLINK(1) and GCTA.(2) The 1000 Genomes Project(3) was used as the reference panel for ORIGIN imputation and was performed using the software IMPUTE2.(4, 5) We removed SNPs imputed with low certainty (info<0.6, as defined by IMPUTE2).(5)

Biomarker Gene Identification

A distance of 300Kb was chosen based on observation of regional associations extending several hundred kilobases (Kb) away from known loci.(6) Genes falling within 300 Kb of each SNP were identified using the Reference Sequence gene list compiled by the UCSC (University of California Santa Cruz) Genome Table Browser.(7) Gene names were identified through the GeneCards Encyclopedia.(8)

Patient characteristics

Between September 2003 and December 2005, 578 clinical sites in 40 countries screened 15,374 individuals and randomized 12,537 participants for the original ORIGIN trial(9). A subset of 8,401 participants were included in the ORIGIN biomarker study (66% men; mean age 63.7 years).(10) A further subset of 5,078 participants consented to genetic analyses and 4,147 passed quality control, and were thus included in Mendelian randomization analyses. Study characteristics were similar across the three groups. Key clinical characteristics of these populations are shown in Supplementary Tables 1-3.

Sensitivity MR analyses

In addition to the bootstrap method employed, we have also tested CSF1 and CXCL12 in an IVW fixed-effect model using the 'MR-base' package in R, where consistent estimates were obtained (CSF1: OR=1.19 per SD; 95% CI 1.08 to 1.30; p=2.0x10^{-4}; CXCL12: OR=1.75 per SD; 95% CI 1.75 to 2.11; p=4.2x10^{-9}). To the CSF1 MR, we also performed a leave-one-out strategy to ensure that no single variant was driving the MR findings and found consistent results for each SNP excluded (Supplementary Table 8). This method was not applied to the CXCL12 data as three or more SNPs are required.

MR analyses were also performed in each ethnic group separately to rule out sources of confounding due to population stratification. Results were consistent across the independent ethnicities for CXCL12 (European: OR=1.63 per SD; 95% CI 1.37 to 1.95; p=5.2x10^{-8}, Latin: OR=1.75 per SD; 95% CI 1.43 to 2.13; p=5.6x10^{-8}) and CSF1 (European: OR=1.18 per SD; 95% CI 1.08 to 1.29; p=4.5x10^{-4}, Latin: OR=1.18 per SD; 95% CI 1.08 to 1.28; p=2.2x10^{-4}).

MR analyses were also performed at a MAF threshold of 0.01 (in addition to 0.05) and consistent estimates were obtained for the two novel biomarkers (CSF1: OR=1.18, 95% CI 1.10 to 1.27, p=6.0x10^{-6}, CXCL12: OR=1.22, 95% CI 1.09 to 1.37, p=3.9x10^{-4}).

Replication of MR findings in the UKBiobank

All analyses involving UKBiobank (UKB) were conducted under data application number 1525. UKB is a prospective cohort study including more than 500,000 individuals (40-69 years) recruited from the United Kingdom during 2006-2010. Samples were genotyped on either the UK Biobank Array or the UK BiLEVE array. Phasing and imputation were performed using SHAPEIT3 and IMPUTE3, respectively, against a combined haplotype reference panel including UK10K and 1000 Genomes Phase 3. Of the 487,406 individuals with imputed genotypes available, 343,735 unrelated British individuals remained for analyses after excluding based on non-British ancestry, excess heterozygosity, low call rates, gender inconsistencies, and relatedness with other study participants. Filtering criteria were determined by the "het.missing.outliers", "in.kinship.table", and "in.white.British.ancestry.subset" from the "ukb_sqc_v2.txt" bulk data download file containing sample quality control metrics.

Genetic Risk Score (GRS) was calculated using the allelic dosage at each variant site was weighted by the predicted change in biomarker level conferred per additional effect allele (in ORIGIN). Subsequently, the weighted contribution at each variant site was summed to create an overall score for each individual in UKB. Using these biomarker GRS, the association between coronary artery disease and genetically elevated biomarkers was tested using a logistic regression model with age and sex as covariates. Coronary artery disease was defined based on a composite of UK biobank data fields including: 1) heart attack diagnosed by a doctor (data fields 3894 & 6150), and 2) relevant in-hospital ICD-10 diagnoses including unstable angina (ICD code I200), acute myocardial infarction (I210-I214,I219), silent myocardial ischemia (I256), and subsequent myocardial infarction (I220,I221,I228,I229) (data fields 41202 & 41204).

Association of CSF1 and CXCL12 concentration with MACE in ORIGIN

In addition to a minimally adjusted epidemiological model, we also tested models further adjusting for

traditional CAD risk factors as a sensitivity analysis, namely prior type 2 diabetes, BMI, serum cystatin-c, current smoker, diagnosis of hypertension, and LDL (mmol/L). Consistent with our minimally adjusted models, we found that increased levels of blood CSF1 and CXCL12 were significantly associated with an increased risk of incident MACE in models adjusting for age, sex, ethnicity (CSF1: hazard ratio (HR)=1.15 per SD; 95% CI, 1.09 to 1.23; $p<0.0001$ and CXCL12: HR=1.08 per SD; 95% CI, 1.02 to 1.14; $p=0.005$).

Supplementary Table 1: Participant characteristics for the genetic and biomarker sub-study subsets of the ORIGIN study.

Variable	Genetic Study participants (n=4,147)	All biomarker participants (n=8,197)
Age (years), mean (SD)	63.45 (7.98)	63.72 (7.94)
Gender (% male)	64.14	66.11
Ethnicity (%)		
European	46.56	55.41
Black	0	4.36
South Asian	0	5.49
South East Asian	0	0.46
Latin	53.44	34.28
Current smoker (% yes)	55.79	57.79
LDL (mmol/L)	3.07 (1.05)	2.89 (1.03)
HDL (mmol/L)	1.17 (0.32)	1.18 (0.32)
Fasting plasma glucose (mmol/L)	7.58 (2.17)	7.33 (2.02)
Hypertension (% yes)	82.90	78.91
Hypercholesterolemia (% yes)[a]	8.85	6.8
HbA_{1c} (%)	6.56 (0.98)	6.50 (0.95)
Body mass index (kg/m^2)	30.45 (5.33)	30.04 (5.27)
Prior diabetes (% yes)	87.56	81.66
Prospective MACE (% with event)	17.70	16.81

Data are presented as mean (SD) unless stated otherwise. [a]Hypercholesterolemia defined as LDL-cholesterol level >4.5 mmol/L.

Supplementary Table 2: Participant characteristics for all individuals included in the ORIGIN-trial.

		Overall		Glargine		Standard Care	
	N	N/Mean	%/SD	N/Mean	%/SD	N/Mean	%/SD
Categorical Variables							
--N.America + Australia	12537	1516	12.1	762	12.2	754	12.0
--S.America	12537	3853	30.7	1925	30.7	1928	30.7
--Europe	12537	6060	48.3	3027	48.3	3033	48.4
--India	12537	390	3.1	194	3.1	196	3.1
Prior CV Event	12533	7378	58.9	3712	59.3	3666	58.4
Reported or measured Microalb/Alb	12537	3968	31.7	1984	31.7	1984	31.6
Male	12536	8150	65.0	4181	66.8	3969	63.3
Male>=55y or female >=65y	12537	8765	69.9	4432	70.8	4333	69.1
Current Smoking	12533	1552	12.4	781	12.5	771	12.3
Prior diabetes	12536	10321	82.3	5162	82.4	5159	82.2
Hypertension	12533	9963	79.5	4974	79.5	4989	79.5
Age	12537	63.54	7.82	63.55	7.79	63.54	7.85
Continuous Variables							
--Cholesterol (mmol/L)	12521	4.90	1.20	4.91	1.20	4.90	1.20
--LDL Cholesterol (mmol/L)	12328	2.90	1.03	2.91	1.04	2.90	1.03
--HDL Cholesterol (mmol/L)	12471	1.19	0.32	1.19	0.32	1.20	0.32
Outcome Variables							
Coprimary outcome 1	12537	2054	16.4	1041	16.6	1013	16.1
Coprimary outcome 2	12537	3519	28.1	1792	28.6	1727	27.5
Microvascular	12537	2686	21.4	1323	21.1	1363	21.7
New Diabetes	12536	760	6.1	365	5.8	395	6.3
Death	12537	1916	15.3	951	15.2	965	15.4
A1C <6% at 2 year visit	11417	5729	50.2	3362	59.4	2367	41.1
Diabetes duration	11081	5.41	5.98	5.49	6.05	5.33	5.92
IFG or IGT	12537	1452	11.6	735	11.7	717	11.4
Statins	12533	6740	53.8	3373	53.9	3367	53.7

ACE-I or ARB	12533	8681	69.3	4330	69.2	4351	69.4
Beta blockers	12533	6598	52.6	3273	52.3	3325	53.0
Antiplatelets	12533	1706	13.6	855	13.7	851	13.6

Supplementary Table 3: Participant characteristics for all individuals included in biomarker sub-study.

		Overall		Glargine		Standard Care	
	N	N/Mean	%/SD	N/Mean	%/SD	N/Mean	%/SD
Categorical Variables							
--N.America + Australia	8401	1425	17.0	710	16.9	715	17.0
--S.America	8401	2772	33.0	1388	33.1	1384	32.9
--Europe	8401	3822	45.5	1903	45.4	1919	45.6
--India	8401	382	4.5	191	4.6	191	4.5
Prior CV Event	8400	4991	59.4	2513	60.0	2478	58.9
Reported or measured Microalb/Alb	8401	2656	31.6	1330	31.7	1326	31.5
Male	8401	5553	66.1	2834	67.6	2719	64.6
Male>=55y or female >=65y	8401	5928	70.6	2997	71.5	2931	69.6
Current Smoking*	8400	1050	12.5	525	12.5	525	12.5
Prior diabetes	8401	6840	81.4	3422	81.6	3418	81.2
Hypertension	8400	6638	79.0	3320	79.2	3318	78.8
Age	8401	63.71	7.94	63.71	7.93	63.70	7.96
Continuous Variables							
--Cholesterol (mmol/L)	8393	4.89	1.18	4.89	1.17	4.89	1.18
--LDL Cholesterol (mmol/L)	8278	2.90	1.03	2.90	1.03	2.89	1.02
--HDL Cholesterol (mmol/L)	8370	1.18	0.32	1.17	0.31	1.18	0.32
Outcome Variables							
Coprimary outcome 1	8401	1405	16.7	727	17.3	678	16.1
Coprimary outcome 2	8401	2435	29.0	1245	29.7	1190	28.3
Microvascular	8401	1794	21.4	887	21.2	907	21.5
New Diabetes	8401	550	6.5	259	6.2	291	6.9
Death	8401	1340	16.0	672	16.0	668	15.9
A1C <6% at 2 year visit	7668	4042	52.7	2389	62.7	1653	42.8
Diabetes duration	7390	5.26	5.82	5.40	5.95	5.12	5.69
IFG or IGT	8401	1008	12.0	510	12.2	498	11.8
Statins	8400	4616	55.0	2302	54.9	2314	55.0
ACE-I or ARB	8400	5793	69.0	2873	68.6	2920	69.4
Beta blockers	8400	4526	53.9	2249	53.7	2277	54.1
Antiplatelets	8400	1120	13.3	553	13.2	567	13.5

Supplementary Table 4: List of all biomarkers tested and their corresponding genes.

	Biomarker	Gene
1	6Ckine	*CCL21*
2	Adiponectin	*ADIPOQ*
3	Adrenomedullin	*ADM*
4	Agouti-Related Protein	*AGRP*
5	Aldose Reductase	*AKR1B1*
6	Alpha-1-acid glycoprotein 1	*ORM1*
7	Alpha-1-Antichymotrypsin	*SERPINA3*
8	Alpha-1-Antitrypsin	*SERPINA1*
9	Alpha-1-Microglobulin	*AMBP*
10	Alpha-2-Macroglobulin	*A2M*
11	Angiogenin	*ANG*
12	Angiopoietin-2	*ANGPT2*
13	Angiopoietin-related protein 3	*ANGPTL3*
14	Angiotensin-Converting Enzyme	*ACE*
15	Angiotensinogen	*AGT*
16	Antithrombin-III	*SERPINC1*
17	Apolipoprotein A-I	*APOA1*
18	Apolipoprotein A-II	*APOA2*
19	Apolipoprotein A-IV	*APOA4*
20	Apolipoprotein B	*APOB*

21	Apolipoprotein C-I	*APOC1*
22	Apolipoprotein C-III	*APOC3*
23	Apolipoprotein D	*APOD*
24	Apolipoprotein E	*APOE*
25	Apolipoprotein H	*APOH*
26	Apolipoprotein(a)	*LPA*
27	AXL Receptor Tyrosine Kinase	*AXL*
28	B cell-activating factor	*TNFSF13B*
29	B Lymphocyte Chemoattractant	*CXCL13*
30	Beta Amyloid 1-40	*APP*
31	Beta-2-Microglobulin	*B2M*
32	Brain-Derived Neurotrophic Factor	*BDNF*
33	C-Peptide	*INS*
34	C-Reactive Protein	*CRP*
35	Cathepsin D	*CTSD*
36	CD 40 antigen	*CD40*
37	CD163	*CD163*
38	CD40 Ligand	*CD40LG*
39	CD5 Antigen-like	*CD5L*
40	Cellular Fibronectin	*FN1*
41	Chemerin	*RARRES2*
42	Chemokine CC-4	*CCR4*
43	Chromogranin-A	*CHGA*
44	Clusterin	*CLU*
45	Collagen IV	*COL4A1,COL4A2,COL4A3,COL4A4,COL 4A5,COL4A6*
46	Complement C3	*C3*
47	Complement Factor H Related Protein 1	*CFHR1*
48	Cortisol	*NA*
49	Creatine Kinase-MB	*CKM,CKB*
50	Cystatin-C	*CST3*
51	E-Selectin	*SELE*
52	EN-RAGE	*S100A12*
53	Endoglin	*ENG*
54	Endostatin	*COL18A1*
55	Eotaxin-1	*CCL11*
56	Eotaxin-2	*CCL24*
57	Eotaxin-3	*CCL26*
58	Epithelial-Derived Neutrophil-Activating Protein 78	*CXCL5*
59	Erythropoietin	*EPO*
60	Ezrin	*EZR*
61	Factor VII	*F7*
62	Fas Ligand	*FASLG*
63	FASLG Receptor	*TNFRSF6B*
64	Fatty Acid-Binding Protein adipocyte	*FABP4*
65	Fatty Acid-Binding Protein liver	*FABP1*
66	Ferritin	*FTL,FTH1*
67	Fetuin-A	*AHSG*
68	Fibroblast Growth Factor 21	*FGF21*
69	Fibroblast growth factor 23	*FGF23*
70	Fibulin-1C	*FBLN1*
71	Ficolin-3	*FCN3*
72	Follicle-Stimulating Hormone	*FSHB,CGA*
73	Galectin-3	*LGALS3*
74	Gastric inhibitory polypeptide	*GIP*
75	Gelsolin	*GSN*
76	Glucagon-like Peptide 1 total	*GCG*
77	Glucose-6-phosphate Isomerase	*GPI*
78	Glutathione S-Transferase alpha	*GSTA1,GSTA2,GSTA3,GSTA4,GSTA5*
79	Glycogen phosphorylase isoenzyme BB	*PYGB*
80	Granulocyte Colony-Stimulating Factor	*CSF3*
81	Growth differentiation factor 15	*GDF15*
82	Growth Hormone	*GH1,GH2*
83	Growth-Regulated alpha protein	*CXCL1*
84	Haptoglobin	*HP*

85	Heat-Shock protein 70	*HSPA1A,HSPA1B,HSPA1L,HSPA2,HSPA 4,HSPA4L,HSPA5,HSPA6,HSPA8,HSPA9, HSPA12A,HSPA12B,HSPA13,HSPA14*
86	Hemopexin	*HPX*
87	Hepatocyte Growth Factor	*HGF*
88	Hepatocyte Growth Factor receptor	*MET*
89	Hepsin	*HPN*
90	Human Epidermal Growth Factor Receptor 2	*ERBB2*
91	Immunoglobulin A	*IGH*
92	Immunoglobulin E	*IGH*
93	Immunoglobulin M	*IGH*
94	Insulin	*INS*
95	Insulin-like Growth Factor Binding Protein 4	*IGFBP4*
96	Insulin-like Growth Factor Binding Protein 5	*IGFBP5*
97	Insulin-like Growth Factor Binding Protein 6	*IGFBP6*
98	Insulin-like Growth Factor I	*IGF1*
99	Insulin-like Growth Factor-Binding Protein 1	*IGFBP1*
100	Insulin-like Growth Factor-Binding Protein 2	*IGFBP2*
101	Insulin-like Growth Factor-Binding Protein 3	*IGFBP3*
102	Intercellular Adhesion Molecule 1	*ICAM1*
103	Interferon gamma	*IFNG*
104	Interferon gamma Induced Protein 10	*CXCL10*
105	Interferon-inducible T-cell alpha chemoattractant	*CXCL11*
106	Interleukin-1 beta	*IL1B*
107	Interleukin-1 receptor antagonist	*IL1RN*
108	Interleukin-10	*IL10*
109	Interleukin-12 Subunit p40	*IL12B*
110	Interleukin-16	*IL16*
111	Interleukin-17	*IL17A*
112	Interleukin-18	*IL18*
113	Interleukin-2	*IL2*
114	Interleukin-2 receptor alpha	*IL2RA*
115	Interleukin-23	*IL23A,IL12B*
116	Interleukin-6	*IL6*
117	Interleukin-6 receptor	*IL6R*
118	Interleukin-6 receptor subunit beta	*IL6ST*
119	Interleukin-7	*IL7*
120	Interleukin-8	*CXCL8*
121	Kallikrein 5	*KLK5*
122	Kidney Injury Molecule-1	*HAVCR1*
123	Lactoferrin	*LTF*
124	Lactoylglutathione lyase	*GLO1*
125	Latency-Associated Peptide of Transforming Growth Factor beta 1	*LTBP1*
126	Lectin-Like Oxidized LDL Receptor 1	*OLR1*
127	Leptin	*LEP*
128	Leptin Receptor	*LEPR*
129	Leucine-rich alpha-2-glycoprotein	*LRG1*
130	Luteinizing Hormone	*LHB,CGA*
131	Macrophage Colony-Stimulating Factor 1	*CSF1*
132	Macrophage inflammatory protein 3 beta	*CCL19*
133	Macrophage Inflammatory Protein-1 alpha	*CCL3*
134	Macrophage Inflammatory Protein-1 beta	*CCL4*
135	Macrophage Inflammatory Protein-3 alpha	*CCL20*
136	Macrophage Migration Inhibitory Factor	*MIF*
137	Macrophage-Derived Chemokine	*CCL22*
138	Macrophage-Stimulating Protein	*MST1*
139	Matrix Metalloproteinase-1	*MMP1*
140	Matrix Metalloproteinase-10	*MMP10*
141	Matrix Metalloproteinase-3	*MMP3*
142	Matrix Metalloproteinase-7	*MMP7*
143	Matrix Metalloproteinase-9	*MMP9*
144	Matrix Metalloproteinase-9 total	*MMP9*
145	Mesothelin	*MSLN*
146	Methylglyoxal	*NA*
147	MHC class I chain-related protein A	*MICA*

148	Monocyte Chemotactic Protein 1	*CCL2*
149	Monocyte Chemotactic Protein 2	*CCL8*
150	Monocyte Chemotactic Protein 3	*CCL7*
151	Monocyte Chemotactic Protein 4	*CCL13*
152	Monokine Induced by Gamma Interferon	*CXCL9*
153	Myeloid Progenitor Inhibitory Factor 1	*CCL23*
154	Myeloperoxidase	*MPO*
155	Myoglobin	*MB*
156	N-terminal prohormone of brain natriuretic peptide	*NPPB*
157	Neuronal Cell Adhesion Molecule	*NRCAM*
158	Neuropilin-1	*NRP1*
159	Neutrophil Activating Peptide 2	*PPBP*
160	Neutrophil Gelatinase-Associated Lipocalin	*LCN2*
161	Omentin	*ITLN1*
162	Osteocalcin	*BGLAP*
163	Osteopontin	*SPP1*
164	Osteoprotegerin	*TNFRSF11B*
165	P-Selectin	*SELP*
166	Pancreatic Polypeptide	*PPY*
167	Paraoxanase-1	*PON1*
168	Pentraxin-3	*PTX3*
169	Pepsinogen I	*NA*
170	Peptide YY	*PYY*
171	Periostin	*POSTN*
172	Peroxiredoxin-4	*PRDX4*
173	Phosphoserine Aminotransferase	*PSAT1*
174	Pigment Epithelium Derived Factor	*SERPINF1*
175	Plasminogen Activator Inhibitor 1	*SERPINE1*
176	Platelet-Derived Growth Factor BB	*PDGFB*
177	Progesterone	*NA*
178	Progranulin	*GRN*
179	Proinsulin Intact	*INS*
180	Proinsulin Total	*INS*
181	Prolactin	*PRL*
182	Prostasin	*PRSS8*
183	Prostatic Acid Phosphatase	*ACPP*
184	Protein S100-A4	*S100A4*
185	Protein S100-A6	*S100A6*
186	Pulmonary and Activation-Regulated Chemokine	*CCL18*
187	Receptor for advanced glycosylation end products	*AGER*
188	Receptor tyrosine-protein kinase erbB-3	*ERBB3*
189	Resistin	*RETN*
190	Retinol-binding protein 4	*RBP4*
191	Secreted frizzled-related protein 4	*SFRP4*
192	Selenoprotein P	*SEPP1*
193	Serotransferrin	*TF*
194	Serum Amyloid A Protein	*SAA1,SAA2,SAA4,SAA3P*
195	Serum Amyloid P-Component	*APCS*
196	Serum Glutamic Oxaloacetic Transaminase	*GOT1,GOT2*
197	Sex Hormone-Binding Globulin	*SHBG*
198	Sortilin	*SORT1*
199	ST2	*IL1RL1*
200	Stem Cell Factor	*KITLG*
201	Stromal cell-derived factor-1	*CXCL12*
202	Superoxide Dismutase 1 soluble	*SOD1*
203	T Lymphocyte-Secreted Protein I-309	*CCL1*
204	T-Cell-Specific Protein RANTES	*CCL5*
205	Tamm-Horsfall Urinary Glycoprotein	*UMOD*
206	Tenascin-C	*TNC*
207	Testosterone Total	*NA*
208	Tetranectin	*CLEC3B*
209	Thrombin-activable fibrinolysis inhibitor	*CPB2*
210	Thrombomodulin	*THBD*
211	Thrombospondin-1	*THBS1*
212	Thyroid-Stimulating Hormone	*TSHB,CGA*

213	Thyroxine-Binding Globulin	*SERPINA7*
214	Tissue Inhibitor of Metalloproteinases 1	*TIMP1*
215	Tissue type Plasminogen activator	*PLAT*
216	TNF-Related Apoptosis-Inducing Ligand Receptor 3	*TNFRSF10C*
217	Transthyretin	*TTR*
218	Trefoil Factor 3	*TFF3*
219	Tumor Necrosis Factor alpha	*TNF*
220	Tumor necrosis factor receptor 2	*TNFRSF1B*
221	Tumor Necrosis Factor Receptor I	*TNFRSF1A*
222	Tyrosine kinase with Ig and EGF homology domains 2	*TIE1*
223	Urokinase-type Plasminogen Activator	*PLAU*
224	Urokinase-type plasminogen activator receptor	*PLAUR*
225	Vascular Cell Adhesion Molecule-1	*VCAM1*
226	Vascular Endothelial Growth Factor	*VEGFA*
227	Vascular Endothelial Growth Factor C	*VEGFC*
228	Vascular endothelial growth factor D	*FIGF*
229	Vascular Endothelial Growth Factor Receptor 2	*FLT1*
230	Vascular endothelial growth factor receptor 3	*FLT4*
231	Visceral adipose tissue derived serpin A12	*SERPINA12*
232	Visfatin	*NAMPT*
233	Vitamin D-Binding Protein	*GC*
234	Vitamin K-Dependent Protein S	*PROS1*
235	Vitronectin	*VTN*
236	von Willebrand Factor	*VWF*
237	YKL-40	*CHI3L1*

Supplementary Table 5: Summary of consortia used in MR analyses.

Consortia	Variable(s) used	Study description	Sample	URL
CARDIoGRAM(11)	CAD, MI	Meta-analysis of 48 GWAS, case status defined as CAD diagnosis (i.e. MI, acute coronary syndrome, chronic stable angina or coronary stenosis of >50%)	60,801 CAD cases (approximately 70% MI) and 123,504 controls in predominantly Europeans (77%)	http://www.cardiogramplusc4d.org/
Diabetes Genetics Replication and Meta-Analysis (DIAGRAM)(12)	Type 2 diabetes (T2D)	Meta-analysis of 18 GWAS, reported as Stage 1 GWAS, case status defined as T2D diagnosis	26,676 T2D cases and 132,532 controls in Europeans	http://diagram-consortium.org/index.html
Genetic Investigation of Anthropometric Traits (GIANT)(13)	BMI	Meta-analysis of 114 GWAS and metabochip studies	322,154 European individuals	https://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files
CKDGEN(14)	CKD	Meta-analysis of 43 GWAS, CKD defined as eGFRcrea < 60ml/min per 1.73 m^2	12,385 cases and 104,780 controls in Europeans	https://www.nhlbi.nih.gov/research/intramural/researchers/pi/fox-caroline/datasets
Meta-Analysis of Glucose and Insulin-Related Traits Consortium (MAGIC)(15, 16)	HbA_1c, fasting glucose	Meta-analysis of 23 GWAS for HbA_1c and 21 GWAS for fasting glucose	46,368 in HbA_1c and 46,186 in fasting glucose, non-diabetic, Europeans	http://www.magicinvestigators.org/
Global Lipids Genetics Consortium(17)	LDL-cholesterol, HDL-cholesterol, triglycerides	Meta-analysis of 60 GWAS	188,577 European individuals	http://csg.sph.umich.edu//abecasis/public/lipids2013/
International Consortium for Blood Pressure (ICBP)(18)	SBP, DBP	Meta-analysis of 29 GWAS	69,395 European individuals	http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000585.v1.p1

Supplementary Table 6: Summary of 7 independent (R^2<0.1) SNP associations at the *CSF1* locus (+/- 300KB) located on chromosome 1. Association statistics provided for serum CSF1 (in ORIGIN) and CAD (from CARDIoGRAM).

SNP	Position	MAF	Effect allele	Other allele	CSF1 Beta	CSF1 SE	CSF1 p-value	CAD Beta	CAD SE	CAD p-value
rs9429558	110495380	0.24754	A	G	0.0717	0.0253	0.0046	0.007441	0.010906	0.495068
rs875902	110425963	0.329648	T	C	-0.0641	0.0231	0.0055	-0.00819	0.009793	0.40297
rs3768471	110612436	0.065892	A	G	0.154	0.0445	0.00055	0.028094	0.019986	0.159827
rs2050462	110472956	0.382667	G	T	0.0836	0.023	0.00028	0.01877	0.009739	0.053929
rs12089727	110461748	0.121566	T	C	0.0815	0.0314	0.0095	0.01309	0.013877	0.345541
rs11588387	110537864	0.076395	T	C	-0.1375	0.0455	0.0025	-0.0227	0.021329	0.287183
rs11579145	110501597	0.469632	A	G	-0.1275	0.0211	1.67E-09	-0.02202	0.009439	0.019662

MAF: minor allele frequency, SE: standard error. Beta coefficient corresponds to the risk coefficient for each SD increase in effect allele.

Supplementary Table 7: Summary of 2 independent (R^2<0.1) SNP associations at the *CXCL12* locus (+/- 300KB) located on chromosome 10. Association statistics provided for serum CXCL12 (in ORIGIN) and CAD (from CARDIoGRAM).

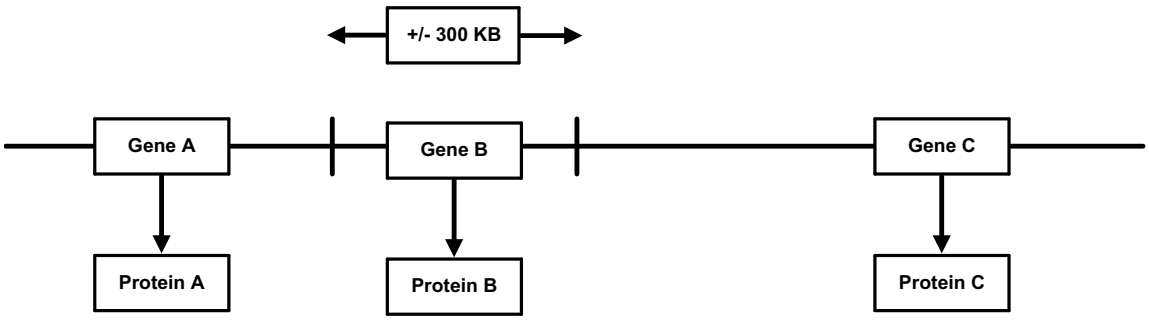
SNP	Position	MAF	Effect allele	Other allele	CXCL12 Beta	CXCL12 SE	CXCL12 p-value	CAD Beta	CAD SE	CAD p-value
rs880175	44854194	0.158794	T	C	-0.1107	0.0297	0.000198	-0.02919	0.014378	0.042309
rs1482478	44596130	0.549291	A	G	-0.0672	0.0225	0.00290	-0.06117	0.009459	1.00E-10

MAF: minor allele frequency, SE: standard error. Beta coefficient corresponds to the risk coefficient for each unit increase in effect allele.

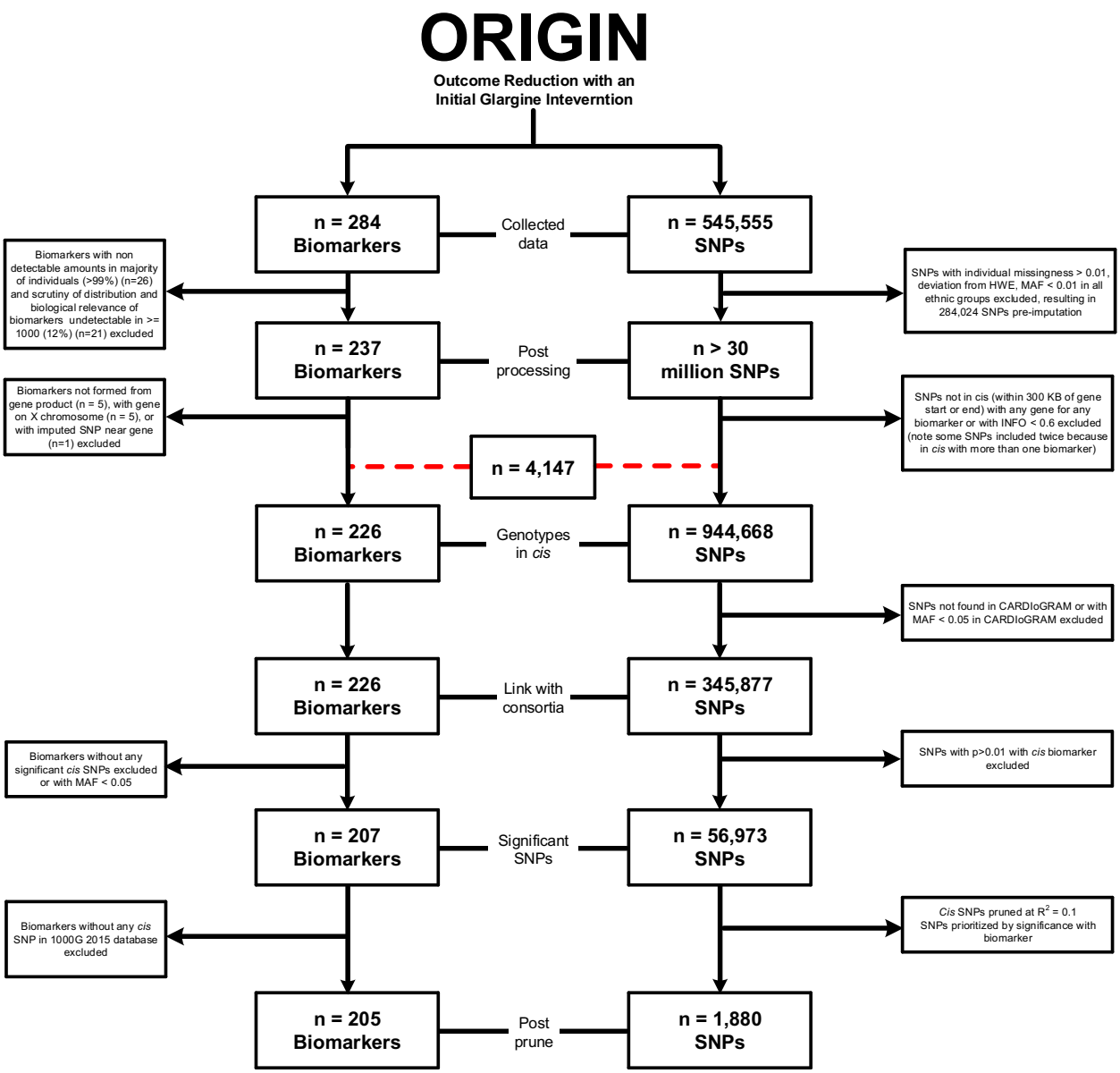
Supplementary Table 8: Summary of leave-one-out sensitivity MR analysis for CSF1.

SNP excluded	OR	95% CI	P-value
rs9429558	1.18	(1.06, 1.33)	$3.8x10^{-3}$
rs875902	1.19	(1.08, 1.3)	$3.7x10^{-4}$
rs3768471	1.19	(1.08, 1.30)	$3.2x10^{-4}$
rs2050462	1.17	(1.06, 1.29)	$1.3x10^{-3}$
rs12089727	1.18	(1.08, 1.30)	$5.8x10^{-4}$
rs11588387	1.19	(1.08, 1.31)	$2.8x10^{-4}$
rs11579145	1.19	(1.09, 1.31)	$2.3x10^{-4}$
Overall	1.19	(1.08, 1.30)	$2.0x10^{-4}$

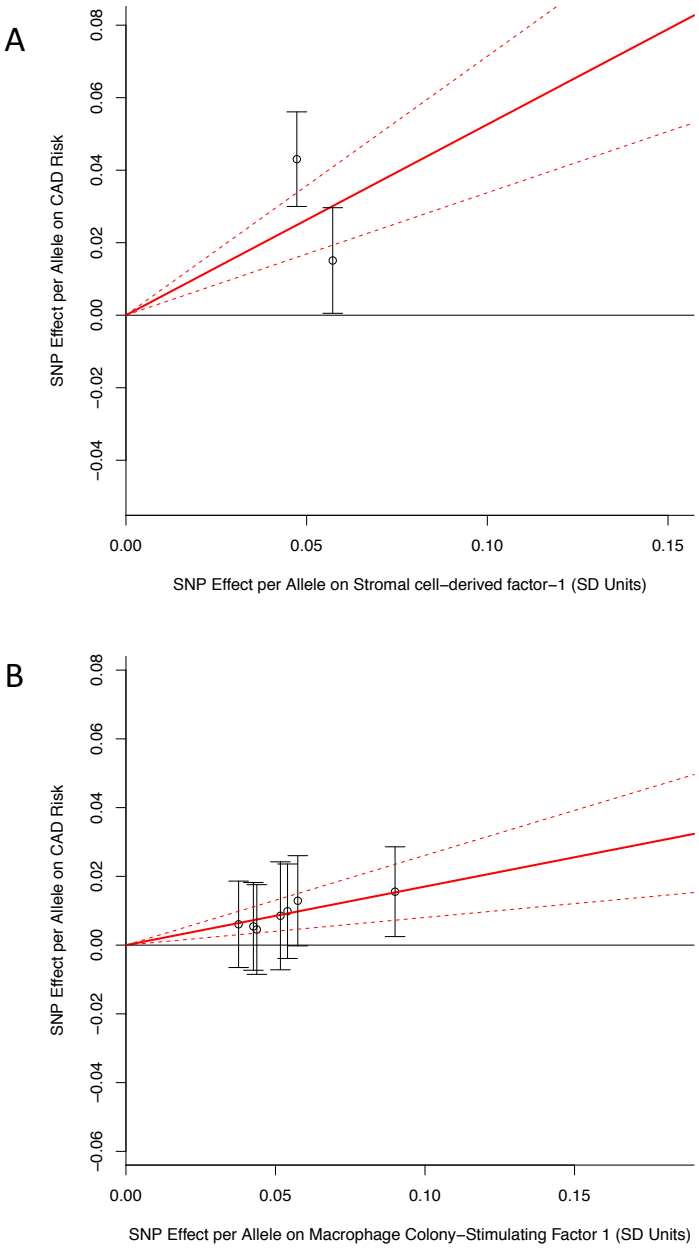
Supplementary Figure 1: Schematic representation of *cis* association.



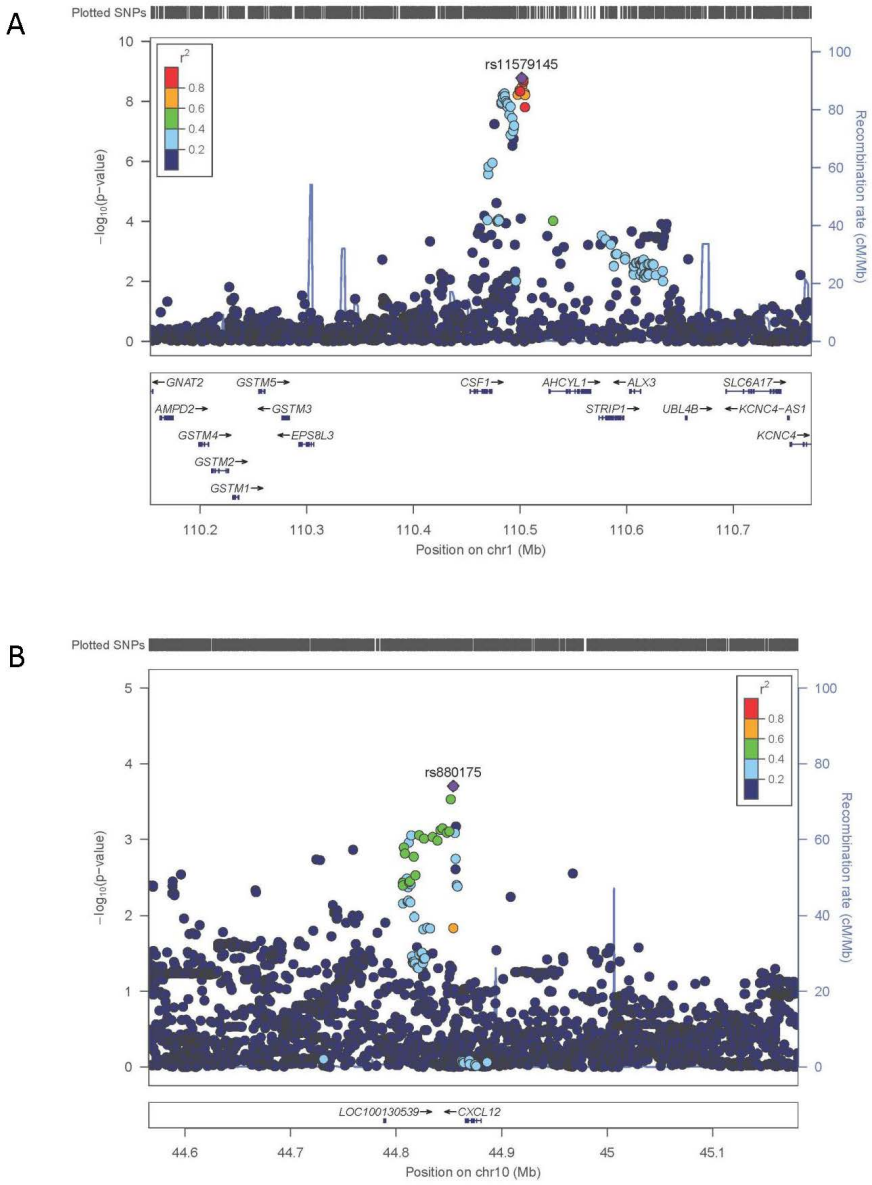
Supplementary Figure 2: Overview of SNP and biomarker selection.



Supplementary Figure 3: Illustration of Mendelian randomization association between CXCL12 and CAD (A) and CSF1 and CAD (2).



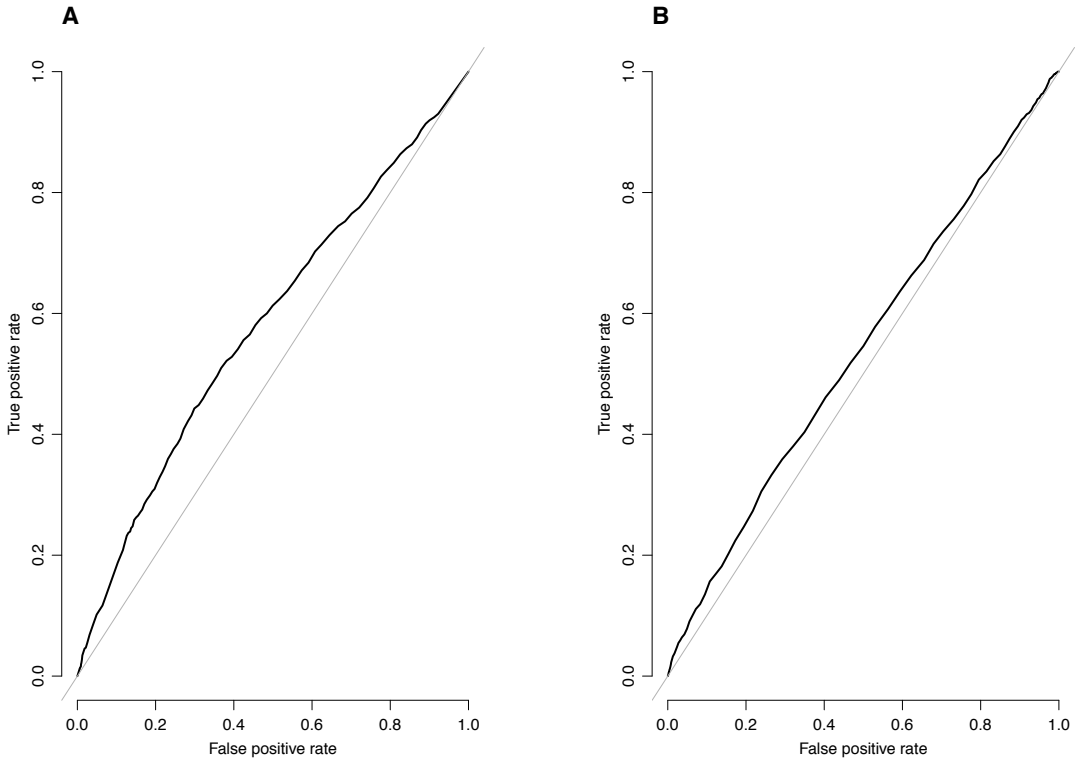
Scatterplot showing the effect estimates of SNP-biomarker associations on the x-axis and SNP-CAD associations (95% CI, from CARDIoGRAM) on the y-axis for all SNPs used in the MR analysis. The continuous black line represents the MR fixed-effects IVW estimate (dashed lines represent corresponding 95% CI).



Supplementary Figure 4: Regional plots for CXCL12 and CSF1 associations.

Plots show association of SNPs with serum CSF1 levels at the *CSF1* locus (A) and serum CXCL12 levels at the *CXCL12* locus (B) +/- 300 KB along with recombination rates. The region defined as *cis,* +/- 300 KB of the *CSF1/CXCL12* genes, is highlighted. $-\log_{10}$ *P* values (y axis) of the SNPs are shown according to their chromosomal positions (x axis). The most significant SNP in the analysis is labeled as a purple triangle. The color intensity of each symbol reflects the extent of LD with the top SNP, colored red ($r^2 > 0{\cdot}8$) through to blue ($r^2 < 0{\cdot}2$). SNPs with missing LD information are labeled grey. Genetic recombination rates (cM/Mb), estimated using 1000 Genomes European samples, are shown with a light blue line. Physical positions are based on build hg19 of the human genome. Also shown are the relative positions of genes mapping to the region of association.

Supplementary Figure 5: ROC curve of (A) CSF1 and (B) CXCL12 for MACE outcomes.



REFERENCES

1. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 2007;81:559–575.

2. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: A tool for genome-wide complex trait analysis. Am. J. Hum. Genet. 2011;88:76–82.

3. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. Nature 2015;526:68–74.

4. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. 2012;44:955–959.

5. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009;5:e1000529.

6. Paré G, Asma S, Deng WQ. Contribution of large region joint associations to complex traits genetics. PLoS Genet. 2015;11:e1005103.

7. Hsu F, Kent JW, Clawson H, Kuhn RM, Diekhans M, Haussler D. The UCSC known genes. Bioinformatics 2006;22:1036–1046.

8. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. Bioinformatics 1998;14:656–664.

9. Gerstein HC, Yusuf S, Riddle MC, Ryden L, Bosch J. Rationale, design, and baseline characteristics for a large international trial of cardiovascular disease prevention in people with dysglycemia: The ORIGIN Trial (Outcome Reduction with an Initial Glargine Intervention). Am. Heart J. 2008;155:26. e1-26. e13.

10. Gerstein HC, Paré G, McQueen MJ, et al. Identifying novel biomarkers for cardiovascular events or death in people with dysglycemia. Circulation 2015;132:2297–2304.

11. Nikpay M, Goel A, Won H-H, et al. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nat. Genet. 2015;47:1121–30.

12. Scott RA, Scott LJ, Mägi R, et al. An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. Diabetes 2017;66:db161253.

13. Locke AE, Kahali B, Berndt SI, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature 2015;518:197–206.

14. Pattaro C, Teumer A, Gorski M, et al. Genetic associations at 53 loci highlight cell types and biological pathways relevant for kidney function. Nat. Commun. 2016;7:10023.

15. Dupuis J, Langenberg C, Prokopenko I, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat. Genet. 2010;42:105–116.

16. Soranzo N, Sanna S, Wheeler E, et al. Common variants at 10 genomic loci influence hemoglobin A1C levels via glycemic and nonglycemic pathways. Diabetes 2010;59:3229–3239.

17. Willer CJ, Schmidt EM, Sengupta S, et al. Discovery and refinement of loci associated with lipid levels. Nat. Genet. 2013;45:1274–83.

18. Ehret GB, Munroe PB, Rice KM, et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature 2011;478:103–109.

Supplementary Appendix for Chapter 4

Blood HER2 and Uromodulin as causal mediators of CKD

Biomarker Assay Methodology in ORIGIN

At Myriad RBM Inc., the samples were thawed at room temperature (RT), vortexed, spun at 13,000g for 5 minutes for clarification. An aliquot was removed into a master microtiter plate for analysis. Using automated pipetting, an aliquot of each sample was introduced into one of the capture microsphere multiplexes of the Human DiscoveryMAP. The mixtures of sample and capture microspheres were thoroughly mixed and incubated at RT for 1 hour. Next, multiplexed cocktails of biotinylated reporter antibodies for each multiplex were added robotically. After thorough mixing, they were incubated for an additional hour at RT. Multiplexes were developed using an excess of streptavidin-phycoerythrin solution which was thoroughly mixed into each multiplex and incubated for 1 hour at RT. The volume of each multiplexed reaction was reduced by vacuum filtration and then increased by dilution into matrix buffer for analysis. Analysis was performed in Luminex 100 and 200 instruments and the resulting data stream was interpreted using proprietary data analysis software developed at RBM. For each multiplex, both calibrators and controls were included on each microtiter plate. Eight-point calibrators were run in the first and last column of each plate and 3-level quality controls were included in duplicate. Testing results were determined first for the high, medium and low controls for each multiplex to ensure proper assay performance. Unknown values for each of the analytes localized in a specific multiplex were determined using 4 and 5 parameter, weighted and non-weighted curve fitting algorithms included in the data analysis package.

Determining the Distribution of Each Biomarker in ORIGIN

Biomarkers were scrutinized in 5 steps. First, 26 biomarkers with undetectable levels in > 8409 (i.e. 99%) participants were excluded from further analyses. Second, scrutiny of the mean, median, and distribution of results and biologic literature pertaining to another 64 biomarkers with undetectable levels in > 1000 (i.e. 12%) participants led to exclusion of a further 21, leaving 237 biomarkers for analysis. Third, those biomarkers with levels below the level of quantification in < 10% of participants (n=850) were assigned a level corresponding to the lower limit of quantification. Fourth, biomarkers with levels below the level of quantification in > 10% of participants were identified and analyzed as ordinal variables as follows. A level of 1 was assigned to the participants with unquantifiable levels, dividing the remaining participants into 4 groups using quartiles. Values of 2, 3, 4 and 5 were assigned to participants within each progressively higher group. This approach was used to manage skewed biomarker distributions. Fifth, biomarkers with levels above the level of quantification were identified and those affected were assigned a level 1% above the upper limit of quantification. This approach led to 192 biomarkers for analysis as continuous variables and 45 biomarkers for analysis as 5-level ordinal variables.

The distributions of each of the 192 continuous biomarkers' levels were then scrutinized to identify extreme outliers with levels more than 4 standard deviations above or below the mean; levels that met those criteria were assigned the value corresponding to the mean plus or minus the 4th standard deviation respectively. Subsequently, the levels of 125 biomarkers with distributions that were not normally distributed were log-transformed using the natural logarithm. Biomarkers analyzed as continuous variables were then standardized to have mean 0 and standard deviation of 1. Finally, data from 93 participants in whom all 237 biomarkers were not analyzed due to insufficient volume of serum were excluded.

Genotyping Quality Control and Imputation in ORIGIN

SNPs were excluded on the basis of low call rate (<99%), deviation from Hardy-Weinberg ($p<10^{-6}$), and low minor allele frequency (MAF<0.01). Samples with low call rates (<99%), sex or ethnicity mismatches, or cryptic relatedness were also removed. We also removed ethnicities with small sample sizes (n<500). All quality control steps were performed using PLINK[1] and GCTA.[2] The 1000 Genomes Project[3] was used as the reference panel for ORIGIN imputation and was performed using the software IMPUTE2.[4,5] We removed SNPs imputed with low certainty (info<0.6, as defined by IMPUTE2).[5]

Biomarker Gene Identification

A distance of 300Kb was chosen based on observation of regional associations extending several hundred kilobases (Kb) away from known loci.[6] Genes falling within 300 Kb of each SNP were identified using the Reference Sequence gene list compiled by the UCSC (University of California Santa Cruz) Genome Table Browser.[7] Gene names were identified through the GeneCards Encyclopedia.[8]

Measurement of Biomarkers in Sample of Healthy Kidney Donors

To further explore the relationship of the novel markers found in the MR analysis, we assessed their concentration in ten healthy people before and after donor nephrectomy for renal transplantation at St. Joseph's Healthcare, Hamilton, ON. Blood and urine samples were collected within 3-months prior to nephrectomy (pre-op) and 3 months after the operation (post-op), once kidney function had returned to normal according to eGFR. Participant samples were stored at -80 °C until required for further analyses. Plasma samples were thawed at room temperature and biomarkers were measured using a solid-phase sandwich enzyme-linked immunosorbent assay (ELISA). 90mL urine collection containers were thawed completely and mixed thoroughly. Approximately 1mL of urine was aliquoted from the collection container into Eppendorf tubes. The Eppendorf tubes were centrifuged at room temp for 7 minutes at 16,100 xg. 10uL from each Eppendorf was aliquoted into a 2nd tube with 190uL Calibrator diluent (referred to as dilution #1). All "dilution #1" tubes were vortexed thoroughly. These tubes were not centrifuged. 10uL from each "dilution #1" Eppendorf was aliquoted into a 3rd tube with 490uL Calibrator diluent (referred to as dilution #2). All "dilution #2" tubes were vortexed thoroughly, then centrifuged at room temp for 7 minutes @ 16,100 xg. Urinary uromodulin was indexed to creatinine as uromodulin-to-creatinine ratio in order to account for differences in urine concentration. ELISA was performed using the Human Magnetic Luminex Screening Assay for Uromodulin (BR64) from R&D according to manufacturer's specifications. The biomarker concentration was determined using Bio-Rad Bio-Plex 200 system. The optical density was analyzed and exported for further analysis using BioPlex Manager 6.1.

Association of UMOD and HER2 concentration with CKD in ORIGIN

In addition to a minimally adjusted epidemiological model, we also tested models further adjusting for traditional CKD risk factors as a sensitivity analysis, namely prior type 2 diabetes, prior renal disease, BMI, estimate glomerular filtration rate (eGFR), current smoker, diagnosis of hypertension, and LDL (mmol/L). Consistent with our minimally adjusted models, we found that increased levels of blood HER2 was associated with an increased risk of incident CKD, while blood UMOD was associated with a decreased risk of incident CKD (HER2: odds ratio (OR)=1.07 per SD; 95% CI, 1.01 to 1.13; p=0.026 and UMOD: OR=0.86 per SD; 95% CI, 0.81 to 0.92; $p<9x10^{-7}$).

Additional MR analyses

In addition to the bootstrap method employed, we have also tested UMOD and HER2 in an IVW fixed-effect model using the 'MR-base' package in R, where consistent estimates were obtained (UMOD: OR=1.32 per SD; 95% CI 1.28 to 1.37; p=4.1x10-44; HER2: OR=1.30 per SD; 95% CI 1.15 to 1.48; p=4.4x10-5). We also performed a leave-one-out sensitivity analysis to ensure that no single variant was driving the observed causal effect and found consistent findings for each SNP excluded (Supplementary Table 3 and 4)

Supplementary Table 1: Summary of 5 independent (R^2<0.1) SNP associations at the *ERBB2* locus (+/- 300KB) located on chromosome 17. Association statistics provided for serum HER2 (in ORIGIN) and CKD (from CKDGen).

SNP	Position	MAF	Effect allele	Other allele	HER2 Beta	HER2 SE	HER2 p-value	CKD Beta	CKD SE	CKD p-value
rs11655584	38139024	0.080	T	C	0.1023	0.0393	0.00925	0.041	0.028	0.15
rs1565922	37831035	0.319	G	A	0.1386	0.0226	9.42E-10	0.022	0.016	0.19
rs17608925	38082831	0.148	C	T	0.1443	0.036	6.08E-05	0.066	0.024	0.0068
rs9892427	37804858	0.066	C	T	0.1949	0.0444	1.15E-05	0.067	0.028	0.018
rs9896218	37894463	0.085	C	A	0.2731	0.0484	1.83E-08	0.052	0.033	0.11

MAF: minor allele frequency (from CKDGen), SE: standard error. Beta coefficient corresponds to the risk coefficient for each SD increase in effect allele.

Supplementary Table 2: Summary of 17 independent (R^2<0.1) SNP associations at the *UMOD* locus (+/- 300KB) located on chromosome 16. Association statistics provided for serum UMOD (in ORIGIN) and CKD (from CKDGen).

SNP	Position	MAF	Effect allele	Other allele	UMOD Beta	UMOD SE	UMOD p-value	CKD Beta	CKD SE	CKD p-value
rs12917707	20367690	0.151	T	G	-0.7709	0.028	8.35E-154	-0.22	0.021	2.10E-25
rs1155876	20522779	0.084	C	T	-0.128	0.0331	0.000110	-0.088	0.027	0.0012
rs12446492	20408377	0.389	A	T	-0.3293	0.0235	1.05E-43	-0.100	0.016	3.20E-10
rs12708631	20365697	0.075	A	T	-0.1312	0.0357	0.000241	-0.130	0.034	7.00E-05
rs12920537	20421556	0.310	G	A	-0.236	0.0293	1.16E-15	-0.066	0.021	0.0017
rs12920708	20407237	0.230	C	A	-0.3083	0.0257	1.50E-32	-0.067	0.019	4.00E-04
rs12930599	20335325	0.219	A	G	-0.2183	0.0309	1.77E-12	-0.036	0.029	0.21
rs4380062	20252959	0.086	T	C	-0.1215	0.0387	0.00172	-0.012	0.028	0.67
rs4462596	20232265	0.261	G	A	-0.0935	0.0296	0.00159	-0.037	0.017	0.027
rs4558425	20309215	0.058	A	G	-0.1138	0.0342	0.000876	-0.046	0.034	0.17
rs6497445	20216617	0.133	C	T	0.0886	0.0302	0.00331	0.020	0.022	0.37
rs7187470	20419775	0.451	A	G	0.1134	0.0293	0.000110	0.0022	0.022	0.92
rs7192921	20129595	0.406	G	T	0.0665	0.0222	0.00274	0.017	0.015	0.28
rs7198000	20351937	0.124	A	G	0.1804	0.0324	2.82E-08	0.062	0.024	0.0088
rs7498776	20611149	0.080	C	T	0.104	0.0325	0.00140	-0.034	0.027	0.21
rs7499304	20564390	0.288	T	G	-0.1043	0.0284	0.000249	0.012	0.019	0.54
rs8060932	20344077	0.315	G	A	0.1586	0.0274	7.23E-09	0.044	0.019	0.017

MAF: minor allele frequency (from CKDGen), SE: standard error. Beta coefficient corresponds to the risk coefficient for each unit increase in effect allele.

Supplementary Table 3: Summary of leave-one-out sensitivity MR analysis for UMOD.

SNP excluded	OR	95% CI	P-value
rs1155876	1.31	(1.26, 1.37)	1.24E-42
rs12446492	1.31	(1.26, 1.37)	1.15E-35
rs12708631	1.31	(1.26, 1.36)	1.36E-42
rs12917707	1.30	(1.23, 1.38)	3.70E-20
rs12920537	1.32	(1.27, 1.37)	5.89E-42
rs12920708	1.33	(1.27, 1.38)	1.28E-41
rs12930599	1.32	(1.27, 1.37)	6.23E-44
rs4380062	1.32	(1.27, 1.37)	3.35E-44
rs4462596	1.32	(1.27, 1.37)	3.57E-43
rs4558425	1.32	(1.27, 1.37)	9.41E-44
rs6497445	1.32	(1.27, 1.37)	6.11E-44
rs7187470	1.32	(1.27, 1.37)	1.70E-44
rs7192921	1.32	(1.27, 1.37)	7.82E-44
rs7198000	1.32	(1.27, 1.37)	1.03E-42
rs7498776	1.32	(1.27, 1.37)	5.98E-45
rs7499304	1.32	(1.27, 1.38)	4.83E-45
rs8060932	1.32	(1.27, 1.37)	6.09E-43
Overall	1.32	(1.27, 1.37)	4.12E-44

Supplementary Table 4: Summary of leave-one-out sensitivity MR analysis for HER2.

SNP excluded	OR	95% CI	P-value
rs11655584	1.29	(1.13, 1.47)	0.00012
rs1565922	1.36	(1.17, 1.59)	6.38E-05
rs17608925	1.26	(1.1, 1.44)	0.0010
rs9892427	1.27	(1.11, 1.47)	0.00074
rs9896218	1.34	(1.15, 1.55)	0.00012
Overall	1.30	(1.15, 1.48)	4.39E-05

Supplementary Table 5: List of all biomarkers tested and their corresponding genes.

	Biomarker	Gene
1	6Ckine	*CCL21*
2	Adiponectin	*ADIPOQ*
3	Adrenomedullin	*ADM*
4	Agouti-Related Protein	*AGRP*
5	Aldose Reductase	*AKR1B1*
6	Alpha-1-acid glycoprotein 1	*ORM1*
7	Alpha-1-Antichymotrypsin	*SERPINA3*
8	Alpha-1-Antitrypsin	*SERPINA1*
9	Alpha-1-Microglobulin	*AMBP*
10	Alpha-2-Macroglobulin	*A2M*
11	Angiogenin	*ANG*
12	Angiopoietin-2	*ANGPT2*
13	Angiopoietin-related protein 3	*ANGPTL3*
14	Angiotensin-Converting Enzyme	*ACE*
15	Angiotensinogen	*AGT*
16	Antithrombin-III	*SERPINC1*
17	Apolipoprotein A-I	*APOA1*
18	Apolipoprotein A-II	*APOA2*
19	Apolipoprotein A-IV	*APOA4*
20	Apolipoprotein B	*APOB*
21	Apolipoprotein C-I	*APOC1*
22	Apolipoprotein C-III	*APOC3*
23	Apolipoprotein D	*APOD*
24	Apolipoprotein E	*APOE*
25	Apolipoprotein H	*APOH*
26	Apolipoprotein(a)	*LPA*
27	AXL Receptor Tyrosine Kinase	*AXL*
28	B cell-activating factor	*TNFSF13B*
29	B Lymphocyte Chemoattractant	*CXCL13*
30	Beta Amyloid 1-40	*APP*
31	Beta-2-Microglobulin	*B2M*
32	Brain-Derived Neurotrophic Factor	*BDNF*
33	C-Peptide	*INS*
34	C-Reactive Protein	*CRP*
35	Cathepsin D	*CTSD*
36	CD 40 antigen	*CD40*
37	CD163	*CD163*
38	CD40 Ligand	*CD40LG*
39	CD5 Antigen-like	*CD5L*
40	Cellular Fibronectin	*FN1*
41	Chemerin	*RARRES2*
42	Chemokine CC-4	*CCR4*
43	Chromogranin-A	*CHGA*
44	Clusterin	*CLU*
45	Collagen IV	*COL4A1,COL4A2,COL4A3,COL4A4,COL4A5,COL4A6*
46	Complement C3	*C3*
47	Complement Factor H Related Protein 1	*CFHR1*
48	Cortisol	*NA*
49	Creatine Kinase-MB	*CKM,CKB*
50	Cystatin-C	*CST3*
51	E-Selectin	*SELE*
52	EN-RAGE	*S100A12*
53	Endoglin	*ENG*
54	Endostatin	*COL18A1*
55	Eotaxin-1	*CCL11*
56	Eotaxin-2	*CCL24*
57	Eotaxin-3	*CCL26*
58	Epithelial-Derived Neutrophil-Activating Protein 78	*CXCL5*
59	Erythropoietin	*EPO*
60	Ezrin	*EZR*
61	Factor VII	*F7*
62	Fas Ligand	*FASLG*
63	FASLG Receptor	*TNFRSF6B*
64	Fatty Acid-Binding Protein adipocyte	*FABP4*

65	Fatty Acid-Binding Protein liver	*FABP1*
66	Ferritin	*FTL,FTH1*
67	Fetuin-A	*AHSG*
68	Fibroblast Growth Factor 21	*FGF21*
69	Fibroblast growth factor 23	*FGF23*
70	Fibulin-1C	*FBLN1*
71	Ficolin-3	*FCN3*
72	Follicle-Stimulating Hormone	*FSHB,CGA*
73	Galectin-3	*LGALS3*
74	Gastric inhibitory polypeptide	*GIP*
75	Gelsolin	*GSN*
76	Glucagon-like Peptide 1 total	*GCG*
77	Glucose-6-phosphate Isomerase	*GPI*
78	Glutathione S-Transferase alpha	*GSTA1,GSTA2,GSTA3,GSTA4,GSTA5*
79	Glycogen phosphorylase isoenzyme BB	*PYGB*
80	Granulocyte Colony-Stimulating Factor	*CSF3*
81	Growth differentiation factor 15	*GDF15*
82	Growth Hormone	*GH1,GH2*
83	Growth-Regulated alpha protein	*CXCL1*
84	Haptoglobin	*HP*
85	Heat-Shock protein 70	*HSPA1A,HSPA1B,HSPA1L,HSPA2,HSPA 4,HSPA4L,HSPA5,HSPA6,HSPA8,HSPA9, HSPA12A,HSPA12B,HSPA13,HSPA14*
86	Hemopexin	*HPX*
87	Hepatocyte Growth Factor	*HGF*
88	Hepatocyte Growth Factor receptor	*MET*
89	Hepsin	*HPN*
90	Human Epidermal Growth Factor Receptor 2	*ERBB2*
91	Immunoglobulin A	*IGH*
92	Immunoglobulin E	*IGH*
93	Immunoglobulin M	*IGH*
94	Insulin	*INS*
95	Insulin-like Growth Factor Binding Protein 4	*IGFBP4*
96	Insulin-like Growth Factor Binding Protein 5	*IGFBP5*
97	Insulin-like Growth Factor Binding Protein 6	*IGFBP6*
98	Insulin-like Growth Factor I	*IGF1*
99	Insulin-like Growth Factor-Binding Protein 1	*IGFBP1*
100	Insulin-like Growth Factor-Binding Protein 2	*IGFBP2*
101	Insulin-like Growth Factor-Binding Protein 3	*IGFBP3*
102	Intercellular Adhesion Molecule 1	*ICAM1*
103	Interferon gamma	*IFNG*
104	Interferon gamma Induced Protein 10	*CXCL10*
105	Interferon-inducible T-cell alpha chemoattractant	*CXCL11*
106	Interleukin-1 beta	*IL1B*
107	Interleukin-1 receptor antagonist	*IL1RN*
108	Interleukin-10	*IL10*
109	Interleukin-12 Subunit p40	*IL12B*
110	Interleukin-16	*IL16*
111	Interleukin-17	*IL17A*
112	Interleukin-18	*IL18*
113	Interleukin-2	*IL2*
114	Interleukin-2 receptor alpha	*IL2RA*
115	Interleukin-23	*IL23A,IL12B*
116	Interleukin-6	*IL6*
117	Interleukin-6 receptor	*IL6R*
118	Interleukin-6 receptor subunit beta	*IL6ST*
119	Interleukin-7	*IL7*
120	Interleukin-8	*CXCL8*
121	Kallikrein 5	*KLK5*
122	Kidney Injury Molecule-1	*HAVCR1*
123	Lactoferrin	*LTF*
124	Lactoylglutathione lyase	*GLO1*
125	Latency-Associated Peptide of Transforming Growth Factor beta 1	*LTBP1*
126	Lectin-Like Oxidized LDL Receptor 1	*OLR1*
127	Leptin	*LEP*
128	Leptin Receptor	*LEPR*
129	Leucine-rich alpha-2-glycoprotein	*LRG1*
130	Luteinizing Hormone	*LHB,CGA*
131	Macrophage Colony-Stimulating Factor 1	*CSF1*

132	Macrophage inflammatory protein 3 beta	*CCL19*
133	Macrophage Inflammatory Protein-1 alpha	*CCL3*
134	Macrophage Inflammatory Protein-1 beta	*CCL4*
135	Macrophage Inflammatory Protein-3 alpha	*CCL20*
136	Macrophage Migration Inhibitory Factor	*MIF*
137	Macrophage-Derived Chemokine	*CCL22*
138	Macrophage-Stimulating Protein	*MST1*
139	Matrix Metalloproteinase-1	*MMP1*
140	Matrix Metalloproteinase-10	*MMP10*
141	Matrix Metalloproteinase-3	*MMP3*
142	Matrix Metalloproteinase-7	*MMP7*
143	Matrix Metalloproteinase-9	*MMP9*
144	Matrix Metalloproteinase-9 total	*MMP9*
145	Mesothelin	*MSLN*
146	Methylglyoxal	*NA*
147	MHC class I chain-related protein A	*MICA*
148	Monocyte Chemotactic Protein 1	*CCL2*
149	Monocyte Chemotactic Protein 2	*CCL8*
150	Monocyte Chemotactic Protein 3	*CCL7*
151	Monocyte Chemotactic Protein 4	*CCL13*
152	Monokine Induced by Gamma Interferon	*CXCL9*
153	Myeloid Progenitor Inhibitory Factor 1	*CCL23*
154	Myeloperoxidase	*MPO*
155	Myoglobin	*MB*
156	N-terminal prohormone of brain natriuretic peptide	*NPPB*
157	Neuronal Cell Adhesion Molecule	*NRCAM*
158	Neuropilin-1	*NRP1*
159	Neutrophil Activating Peptide 2	*PPBP*
160	Neutrophil Gelatinase-Associated Lipocalin	*LCN2*
161	Omentin	*ITLN1*
162	Osteocalcin	*BGLAP*
163	Osteopontin	*SPP1*
164	Osteoprotegerin	*TNFRSF11B*
165	P-Selectin	*SELP*
166	Pancreatic Polypeptide	*PPY*
167	Paraoxanase-1	*PON1*
168	Pentraxin-3	*PTX3*
169	Pepsinogen I	*NA*
170	Peptide YY	*PYY*
171	Periostin	*POSTN*
172	Peroxiredoxin-4	*PRDX4*
173	Phosphoserine Aminotransferase	*PSAT1*
174	Pigment Epithelium Derived Factor	*SERPINF1*
175	Plasminogen Activator Inhibitor 1	*SERPINE1*
176	Platelet-Derived Growth Factor BB	*PDGFB*
177	Progesterone	*NA*
178	Progranulin	*GRN*
179	Proinsulin Intact	*INS*
180	Proinsulin Total	*INS*
181	Prolactin	*PRL*
182	Prostasin	*PRSS8*
183	Prostatic Acid Phosphatase	*ACPP*
184	Protein S100-A4	*S100A4*
185	Protein S100-A6	*S100A6*
186	Pulmonary and Activation-Regulated Chemokine	*CCL18*
187	Receptor for advanced glycosylation end products	*AGER*
188	Receptor tyrosine-protein kinase erbB-3	*ERBB3*
189	Resistin	*RETN*
190	Retinol-binding protein 4	*RBP4*
191	Secreted frizzled-related protein 4	*SFRP4*
192	Selenoprotein P	*SEPP1*
193	Serotransferrin	*TF*
194	Serum Amyloid A Protein	*SAA1,SAA2,SAA4,SAA3P*
195	Serum Amyloid P-Component	*APCS*
196	Serum Glutamic Oxaloacetic Transaminase	*GOT1,GOT2*
197	Sex Hormone-Binding Globulin	*SHBG*
198	Sortilin	*SORT1*
199	ST2	*IL1RL1*
200	Stem Cell Factor	*KITLG*

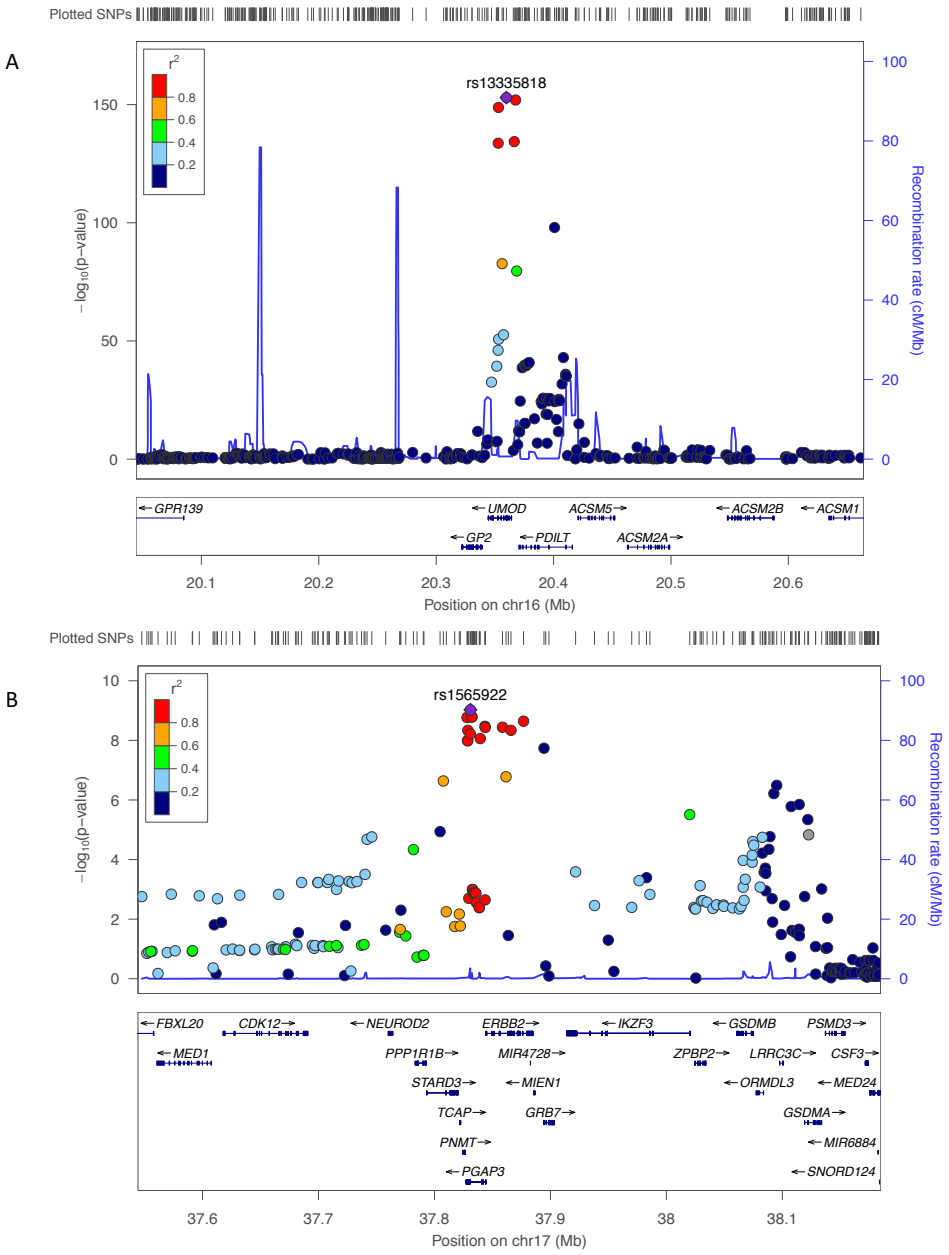
201	Stromal cell-derived factor-1	*CXCL12*
202	Superoxide Dismutase 1 soluble	*SOD1*
203	T Lymphocyte-Secreted Protein I-309	*CCL1*
204	T-Cell-Specific Protein RANTES	*CCL5*
205	Tamm-Horsfall Urinary Glycoprotein (Uromodoulin)	*UMOD*
206	Tenascin-C	*TNC*
207	Testosterone Total	*NA*
208	Tetranectin	*CLEC3B*
209	Thrombin-activable fibrinolysis inhibitor	*CPB2*
210	Thrombomodulin	*THBD*
211	Thrombospondin-1	*THBS1*
212	Thyroid-Stimulating Hormone	*TSHB,CGA*
213	Thyroxine-Binding Globulin	*SERPINA7*
214	Tissue Inhibitor of Metalloproteinases 1	*TIMP1*
215	Tissue type Plasminogen activator	*PLAT*
216	TNF-Related Apoptosis-Inducing Ligand Receptor 3	*TNFRSF10C*
217	Transthyretin	*TTR*
218	Trefoil Factor 3	*TFF3*
219	Tumor Necrosis Factor alpha	*TNF*
220	Tumor necrosis factor receptor 2	*TNFRSF1B*
221	Tumor Necrosis Factor Receptor I	*TNFRSF1A*
222	Tyrosine kinase with Ig and EGF homology domains 2	*TIE1*
223	Urokinase-type Plasminogen Activator	*PLAU*
224	Urokinase-type plasminogen activator receptor	*PLAUR*
225	Vascular Cell Adhesion Molecule-1	*VCAM1*
226	Vascular Endothelial Growth Factor	*VEGFA*
227	Vascular Endothelial Growth Factor C	*VEGFC*
228	Vascular endothelial growth factor D	*FIGF*
229	Vascular Endothelial Growth Factor Receptor 2	*FLT1*
230	Vascular endothelial growth factor receptor 3	*FLT4*
231	Visceral adipose tissue derived serpin A12	*SERPINA12*
232	Visfatin	*NAMPT*
233	Vitamin D-Binding Protein	*GC*
234	Vitamin K-Dependent Protein S	*PROS1*
235	Vitronectin	*VTN*
236	von Willebrand Factor	*VWF*
237	YKL-40	*CHI3L1*

Supplementary Table 6: Participant characteristics for the genetic and biomarker sub-study subsets of the ORIGIN study.

Variable	Genetic Study participants (n=4,147)	All biomarker participants (n=8,197)
Age (years), mean (SD)	63.45 (7.98)	63.72 (7.94)
Gender (% male)	64.14	66.11
Ethnicity (%)		
European	46.56	55.41
Black	0	4.36
South Asian	0	5.49
South East Asian	0	0.46
Latin	53.44	34.28
Current smoker (% yes)	55.79	57.79
LDL (mmol/L)	3.07 (1.05)	2.89 (1.03)
HDL (mmol/L)	1.17 (0.32)	1.18 (0.32)
Fasting plasma glucose (mmol/L)	7.58 (2.17)	7.33 (2.02)
Hypertension (% yes)	82.90	78.91
Hypercholesterolemia (% yes)[a]	8.85	6.8
Body mass index (kg/m^2)	30.45 (5.33)	30.04 (5.27)
Prior diabetes (% yes)	87.56	81.66
EGFR (mL/min/1.73 m^2)	75.91 (21.07)	77.51 (21.86)
Prior renal disease (% yes)	6.75	5.83
Prospective CKD (% with event)	21.41	20.47

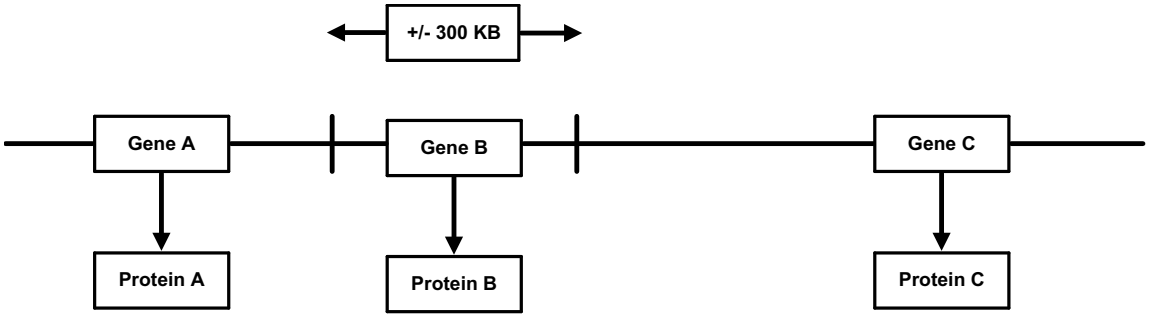
Data are presented as mean (SD) unless stated otherwise.

Supplementary Figure 1: Regional plots for UMOD and HER2 associations.

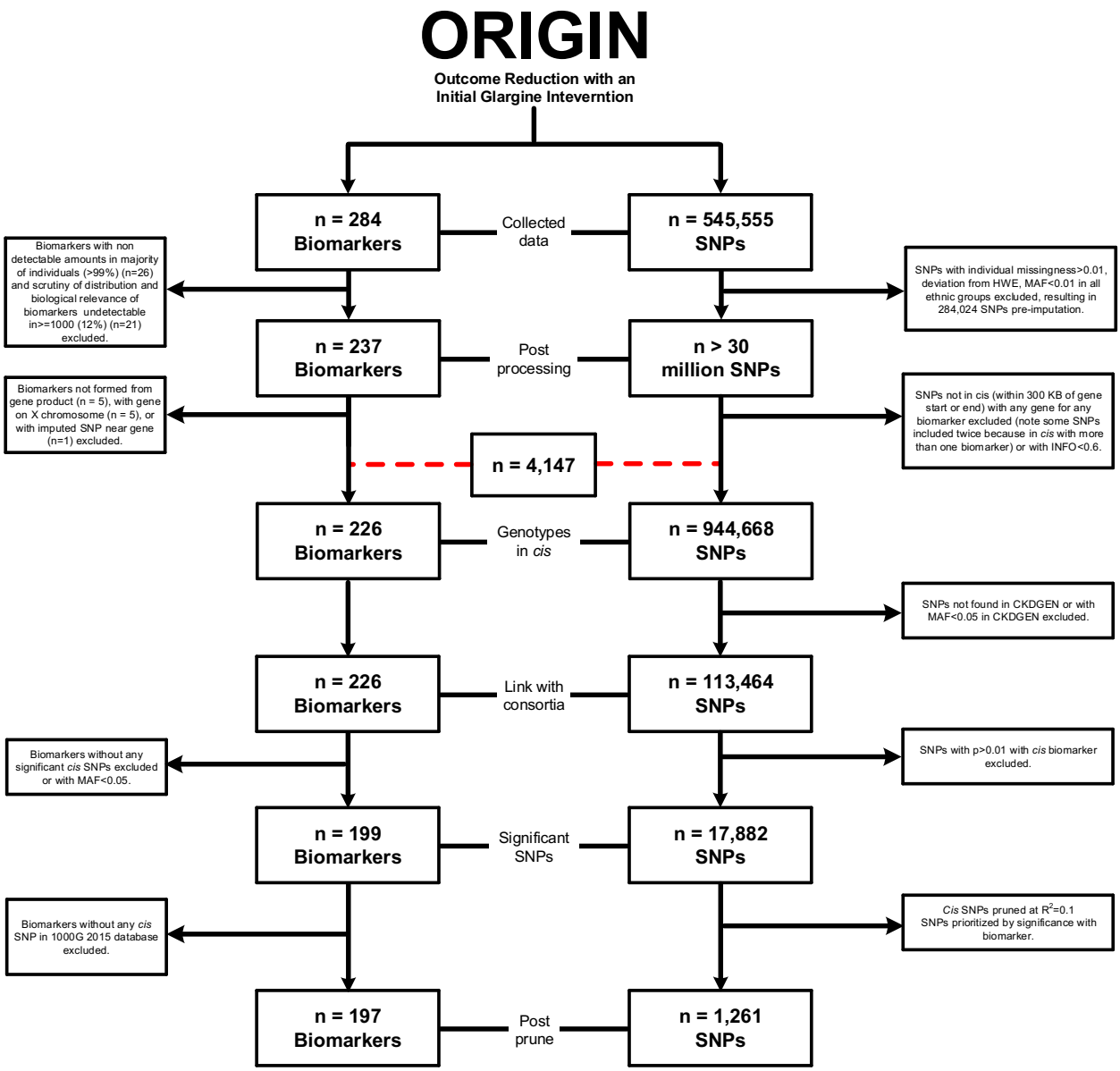


Plots show association of SNPs prior to MR instrument selection with serum UMOD levels at the *UMOD* locus (A) and serum HER2 levels at the *ERBB2* locus (B) +/- 300 KB along with recombination rates. $-\log_{10}$ *P* values (y axis) of the SNPs are shown according to their chromosomal positions (x axis). The most significant SNP in the analysis is labeled as a purple triangle. The color intensity of each symbol reflects the extent of LD with the top SNP, colored red ($r^2 > 0.8$) through to blue ($r^2 < 0.2$). SNPs with missing LD information are labeled grey. Genetic recombination rates (cM/Mb), estimated using 1000 Genomes European samples, are shown with a light blue line. Physical positions are based on build hg19 of the human genome. Also shown are the relative positions of genes mapping to the region of association.

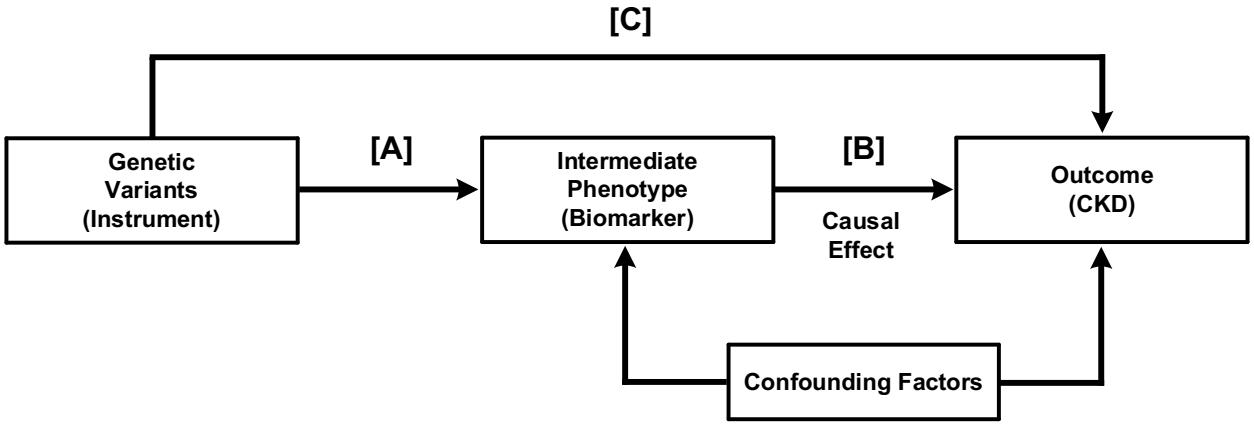
Supplementary Figure 2: Schematic representation of *cis* association.



Supplementary Figure 3: Overview of SNP and biomarker selection.



Supplementary Figure 4: Schematic representation of the instrumental variable assumptions of Mendelian randomization study. Instrumental variable analyses use associations of A and B to estimate the causal effect of an intermediate phenotype (biomarker) on an outcome (CKD), represented by association C.



REFERENCES

1. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M a R, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC: PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81: 559–575, 2007
2. Yang J, Lee SH, Goddard ME, Visscher PM: GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* 88: 76–82, 2011
3. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, Gabriel SB, Gibbs RA, Green ED, Hurles ME, Knoppers BM, Korbel JO, Lander ES, Lee C, Lehrach H, Mardis ER, Marth GT, McVean GA, Nickerson DA, Schmidt JP, Sherry ST, Wang J, Wilson RK, Boerwinkle E, Doddapaneni H, Han Y, Korchina V, Kovar C, Lee S, Muzny D, Reid JG, Zhu Y, Chang Y, Feng Q, Fang X, Guo X, Jian M, Jiang H, Jin X, Lan T, Li G, Li J, Li Y, Liu S, Liu X, Lu Y, Ma X, Tang M, Wang B, Wang G, Wu H, Wu R, Xu X, Yin Y, Zhang D, Zhang W, Zhao J, Zhao M, Zheng X, Gupta N, Gharani N, Toji LH, Gerry NP, Resch AM, Barker J, Clarke L, Gil L, Hunt SE, Kelman G, Kulesha E, Leinonen R, McLaren WM, Radhakrishnan R, Roa A, Smirnov D, Smith RE, Streeter I, Thormann A, Toneva I, Vaughan B, Zheng-Bradley X, Grocock R, Humphray S, James T, Kingsbury Z, Sudbrak R, Albrecht MW, Amstislavskiy VS, Borodina TA, Lienhard M, Mertes F, Sultan M, Timmermann B, Yaspo M-L, Fulton L, Fulton R, Ananiev V, Belaia Z, Beloslyudtsev D, Bouk N, Chen C, Church D, Cohen R, Cook C, Garner J, Hefferon T, Kimelman M, Liu C, Lopez J, Meric P, O'Sullivan C, Ostapchuk Y, Phan L, Ponomarov S, Schneider V, Shekhtman E, Sirotkin K, Slotta D, Zhang H, Balasubramaniam S, Burton J, Danecek P, Keane TM, Kolb-Kokocinski A, McCarthy S, Stalker J, Quail M, Davies CJ, Gollub J, Webster T, Wong B, Zhan Y, Campbell CL, Kong Y, Marcketta A, Yu F, Antunes L, Bainbridge M, Sabo A, Huang Z, Coin LJM, Fang L, Li Q, Li Z, Lin H, Liu B, Luo R, Shao H, Xie Y, Ye C, Yu C, Zhang F, Zheng H, Zhu H, Alkan C, Dal E, Kahveci F, Garrison EP, Kural D, Lee W-P, Fung Leong W, Stromberg M, Ward AN, Wu J, Zhang M, Daly MJ, DePristo MA, Handsaker RE, Banks E, Bhatia G, del Angel G, Genovese G, Li H, Kashin S, McCarroll SA, Nemesh JC, Poplin RE, Yoon SC, Lihm J, Makarov V, Gottipati S, Keinan A, Rodriguez-Flores JL, Rausch T, Fritz MH, Stütz AM, Beal K, Datta A, Herrero J, Ritchie GRS, Zerbino D, Sabeti PC, Shlyakhter I, Schaffner SF, Vitti J, Cooper DN, Ball E V., Stenson PD, Barnes B, Bauer M, Keira Cheetham R, Cox A, Eberle M, Kahn S, Murray L, Peden J, Shaw R, Kenny EE, Batzer MA, Konkel MK, Walker JA, MacArthur DG, Lek M, Herwig R, Ding L, Koboldt DC, Larson D, Ye K, Gravel S, Swaroop A, Chew E, Lappalainen T, Erlich Y, Gymrek M, Frederick Willems T, Simpson JT, Shriver MD, Rosenfeld JA, Bustamante CD, Montgomery SB, De La Vega FM, Byrnes JK, Carroll AW, DeGorter MK, Lacroute P, Maples BK, Martin AR, Moreno-Estrada A, Shringarpure SS, Zakharia F, Halperin E, Baran Y, Cerveira E, Hwang J, Malhotra A, Plewczynski D, Radew K, Romanovitch M, Zhang C, Hyland FCL, Craig DW, Christoforides A, Homer N, Izatt T, Kurdoglu AA, Sinari SA, Squire K, Xiao C, Sebat J, Antaki D, Gujral M, Noor A, Ye K, Burchard EG, Hernandez RD, Gignoux CR, Haussler D, Katzman SJ, James Kent W, Howie B, Ruiz-Linares A, Dermitzakis ET, Devine SE, Min Kang H, Kidd JM, Blackwell T, Caron S, Chen W, Emery S, Fritsche L, Fuchsberger C, Jun G, Li B, Lyons R, Scheller C, Sidore C, Song S, Sliwerska E, Taliun D, Tan A, Welch R, Kate Wing M, Zhan X, Awadalla P, Hodgkinson A, Li Y, Shi X, Quitadamo A, Lunter G, Marchini JL, Myers S, Churchhouse C, Delaneau O, Gupta-Hinch A, Kretzschmar W, Iqbal Z, Mathieson I, Menelaou A, Rimmer A, Xifara DK, Oleksyk TK, Fu Y, Liu X, Xiong M, Jorde L, Witherspoon D, Xing J, Browning BL, Browning SR, Hormozdiari F, Sudmant PH, Khurana E, Tyler-Smith C, Albers CA, Ayub Q, Chen Y, Colonna V, Jostins L, Walter K, Xue Y, Gerstein MB, Abyzov A, Balasubramanian S, Chen J, Clarke D, Fu Y, Harmanci AO, Jin M, Lee D, Liu J, Jasmine Mu X, Zhang J, Zhang Y, Hartl C, Shakir K, Degenhardt J, Meiers S, Raeder B, Paolo Casale F, Stegle O, Lameijer E-W, Hall I, Bafna V, Michaelson J, Gardner EJ, Mills RE, Dayama G, Chen K, Fan X, Chong Z, Chen T, Chaisson MJ, Huddleston J, Malig M, Nelson BJ, Parrish NF, Blackburne B, Lindsay SJ, Ning Z, Zhang Y, Lam H, Sisu C, Challis D, Evani US, Lu J, Nagaswamy U, Yu J, Li W, Habegger L, Yu H, Cunningham F, Dunham I, Lage K, Berg Jespersen J, Horn H, Kim D, Desalle R, Narechania A, Wilson Sayres MA, Mendez FL, David Poznik G, Underhill PA, Coin L, Mittelman D, Banerjee R, Cerezo M, Fitzgerald TW, Louzada S, Massaia A, Ritchie GR, Yang F, Kalra D, Hale W, Dan X, Barnes KC, Beiswanger C, Cai H, Cao H, Henn B, Jones D, Kaye JS, Kent A, Kerasidou A, Mathias R, Ossorio PN, Parker M, Rotimi CN, Royal CD, Sandoval K, Su Y, Tian Z, Tishkoff S, Via M, Wang Y, Yang H, Yang L, Zhu J, Bodmer W, Bedoya G, Cai Z, Gao Y, Chu J, Peltonen L, Garcia-Montero A, Orfao A, Dutil J, Martinez-Cruzado JC, Mathias RA, Hennis A, Watson H, McKenzie C, Qadri F, LaRocque R, Deng X, Asogun D, Folarin O, Happi C, Omoniwa O, Stremlau M, Tariyal R, Jallow M, Sisay Joof F, Corrah T, Rockett K, Kwiatkowski D, Kooner J, Tịnh Hiê`n T, Dunstan SJ, Thuy Hang N, Fonnie R, Garry R, Kanneh L, Moses L, Schieffelin J, Grant DS, Gallo C, Poletti G, Saleheen D, Rasheed A, Brooks LD, Felsenfeld AL,

McEwen JE, Vaydylevich Y, Duncanson A, Dunn M, Schloss JA: A global reference for human genetic variation. *Nature* 526: 68–74, 2015

4. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR: Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* 44: 955–959, 2012
5. Howie BN, Donnelly P, Marchini J: A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* 5: e1000529, 2009
6. Paré G, Asma S, Deng WQ: Contribution of large region joint associations to complex traits genetics. *PLoS Genet.* 11: e1005103, 2015
7. Hsu F, Kent JW, Clawson H, Kuhn RM, Diekhans M, Haussler D: The UCSC known genes. *Bioinformatics* 22: 1036–1046, 2006
8. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D: GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics* 14: 656–664, 1998

Supplementary Appendix for Chapter 5

Influence of Genetic Ancestry on Human Serum Proteome

Biomarker Assay Methodology in ORIGIN

At Myriad RBM Inc., the samples were thawed at room temperature (RT), vortexed, spun at 13,000g for 5 minutes for clarification. An aliquot was removed into a master microtiter plate for analysis. Using automated pipetting, an aliquot of each sample was introduced into one of the capture microsphere multiplexes of the Human DiscoveryMAP. The mixtures of sample and capture microspheres were thoroughly mixed and incubated at RT for 1 hour. Next, multiplexed cocktails of biotinylated reporter antibodies for each multiplex were added robotically. After thorough mixing, they were incubated for an additional hour at RT. Multiplexes were developed using an excess of streptavidin-phycoerythrin solution which was thoroughly mixed into each multiplex and incubated for 1 hour at RT. The volume of each multiplexed reaction was reduced by vacuum filtration and then increased by dilution into matrix buffer for analysis. Analysis was performed in Luminex 100 and 200 instruments and the resulting data stream was interpreted using proprietary data analysis software developed at RBM. For each multiplex, both calibrators and controls were included on each microtiter plate. Eight-point calibrators were run in the first and last column of each plate and 3-level quality controls were included in duplicate. Testing results were determined first for the high, medium and low controls for each multiplex to ensure proper assay performance. Unknown values for each of the analytes localized in a specific multiplex were determined using 4 and 5 parameter, weighted and non-weighted curve fitting algorithms included in the data analysis package.

Determining the Distribution of Each Biomarker in ORIGIN

Biomarkers were scrutinized in 5 steps. First, 26 biomarkers with undetectable levels in > 8409 (i.e. 99%) participants were excluded from further analyses. Second, scrutiny of the mean, median, and distribution of results and biologic literature pertaining to another 64 biomarkers with undetectable levels in > 1000 (i.e. 12%) participants led to exclusion of a further 21, leaving 237 biomarkers for analysis. Third, those biomarkers with levels below the level of quantification in < 10% of participants (n=850) were assigned a level corresponding to the lower limit of quantification. Fourth, biomarkers with levels below the level of quantification in > 10% of participants were identified and analyzed as ordinal variables as follows. A level of 1 was assigned to the participants with unquantifiable levels, dividing the remaining participants into 4 groups using quartiles. Values of 2, 3, 4 and 5 were assigned to participants within each progressively higher group. This approach was used to manage skewed biomarker distributions. Fifth, biomarkers with levels above the level of quantification were identified and those affected were assigned a level 1% above the upper limit of quantification. This approach led to 192 biomarkers for analysis as continuous variables and 45 biomarkers for analysis as 5-level ordinal variables.

The distributions of each of the 192 continuous biomarkers' levels were then scrutinized to identify extreme outliers with levels more than 4 standard deviations above or below the mean; levels that met those criteria were assigned the value corresponding to the mean plus or minus the 4th standard deviation respectively. Subsequently, the levels of 125 biomarkers with distributions that were not normally distributed were log-transformed using the natural logarithm. Biomarkers analyzed as continuous variables were then standardized to have mean 0 and standard deviation of 1. Finally, data from 93 participants in whom all 237 biomarkers were not analyzed due to insufficient volume of serum were excluded.

Biomarker Gene Identification

A distance of 300Kb was chosen based on observation of regional associations extending several hundred kilobases (Kb) away from known loci.[1] Genes falling within 300 Kb of each SNP were identified using the Reference Sequence gene list compiled by the UCSC (University of California Santa Cruz) Genome Table Browser.[2] Gene names were identified through the GeneCards Encyclopedia.[3]

Supplementary Table 1: List of all biomarkers tested and their corresponding genes.

	Biomarker	Gene
1	6Ckine	*CCL21*
2	Adiponectin	*ADIPOQ*
3	Adrenomedullin	*ADM*
4	Agouti-Related Protein	*AGRP*
5	Aldose Reductase	*AKR1B1*
6	Alpha-1-acid glycoprotein 1	*ORM1*
7	Alpha-1-Antichymotrypsin	*SERPINA3*
8	Alpha-1-Antitrypsin	*SERPINA1*
9	Alpha-1-Microglobulin	*AMBP*
10	Alpha-2-Macroglobulin	*A2M*
11	Angiogenin	*ANG*
12	Angiopoietin-2	*ANGPT2*
13	Angiopoietin-related protein 3	*ANGPTL3*
14	Angiotensin-Converting Enzyme	*ACE*
15	Angiotensinogen	*AGT*
16	Antithrombin-III	*SERPINC1*
17	Apolipoprotein A-I	*APOA1*
18	Apolipoprotein A-II	*APOA2*
19	Apolipoprotein A-IV	*APOA4*
20	Apolipoprotein B	*APOB*
21	Apolipoprotein C-I	*APOC1*
22	Apolipoprotein C-III	*APOC3*
23	Apolipoprotein D	*APOD*
24	Apolipoprotein E	*APOE*
25	Apolipoprotein H	*APOH*
26	Apolipoprotein(a)	*LPA*
27	AXL Receptor Tyrosine Kinase	*AXL*
28	B cell-activating factor	*TNFSF13B*
29	B Lymphocyte Chemoattractant	*CXCL13*
30	Beta Amyloid 1-40	*APP*
31	Beta-2-Microglobulin	*B2M*
32	Brain-Derived Neurotrophic Factor	*BDNF*
33	C-Peptide	*INS*
34	C-Reactive Protein	*CRP*
35	Cathepsin D	*CTSD*
36	CD 40 antigen	*CD40*
37	CD163	*CD163*
38	CD40 Ligand	*CD40LG*
39	CD5 Antigen-like	*CD5L*
40	Cellular Fibronectin	*FN1*
41	Chemerin	*RARRES2*
42	Chemokine CC-4	*CCR4*
43	Chromogranin-A	*CHGA*
44	Clusterin	*CLU*
45	Collagen IV	*COL4A1,COL4A2,COL4A3,COL4A4,COL4A5,COL4A6*
46	Complement C3	*C3*
47	Complement Factor H Related Protein 1	*CFHR1*
48	Cortisol	*NA*
49	Creatine Kinase-MB	*CKM,CKB*
50	Cystatin-C	*CST3*
51	E-Selectin	*SELE*
52	EN-RAGE	*S100A12*
53	Endoglin	*ENG*
54	Endostatin	*COL18A1*
55	Eotaxin-1	*CCL11*
56	Eotaxin-2	*CCL24*
57	Eotaxin-3	*CCL26*
58	Epithelial-Derived Neutrophil-Activating Protein 78	*CXCL5*
59	Erythropoietin	*EPO*
60	Ezrin	*EZR*
61	Factor VII	*F7*
62	Fas Ligand	*FASLG*
63	FASLG Receptor	*TNFRSF6B*
64	Fatty Acid-Binding Protein adipocyte	*FABP4*

65	Fatty Acid-Binding Protein liver	*FABP1*
66	Ferritin	*FTL,FTH1*
67	Fetuin-A	*AHSG*
68	Fibroblast Growth Factor 21	*FGF21*
69	Fibroblast growth factor 23	*FGF23*
70	Fibulin-1C	*FBLN1*
71	Ficolin-3	*FCN3*
72	Follicle-Stimulating Hormone	*FSHB,CGA*
73	Galectin-3	*LGALS3*
74	Gastric inhibitory polypeptide	*GIP*
75	Gelsolin	*GSN*
76	Glucagon-like Peptide 1 total	*GCG*
77	Glucose-6-phosphate Isomerase	*GPI*
78	Glutathione S-Transferase alpha	*GSTA1,GSTA2,GSTA3,GSTA4,GSTA5*
79	Glycogen phosphorylase isoenzyme BB	*PYGB*
80	Granulocyte Colony-Stimulating Factor	*CSF3*
81	Growth differentiation factor 15	*GDF15*
82	Growth Hormone	*GH1,GH2*
83	Growth-Regulated alpha protein	*CXCL1*
84	Haptoglobin	*HP*
85	Heat-Shock protein 70	*HSPA1A,HSPA1B,HSPA1L,HSPA2,HSPA 4,HSPA4L,HSPA5,HSPA6,HSPA8,HSPA9, HSPA12A,HSPA12B,HSPA13,HSPA14*
86	Hemopexin	*HPX*
87	Hepatocyte Growth Factor	*HGF*
88	Hepatocyte Growth Factor receptor	*MET*
89	Hepsin	*HPN*
90	Human Epidermal Growth Factor Receptor 2	*ERBB2*
91	Immunoglobulin A	*IGH*
92	Immunoglobulin E	*IGH*
93	Immunoglobulin M	*IGH*
94	Insulin	*INS*
95	Insulin-like Growth Factor Binding Protein 4	*IGFBP4*
96	Insulin-like Growth Factor Binding Protein 5	*IGFBP5*
97	Insulin-like Growth Factor Binding Protein 6	*IGFBP6*
98	Insulin-like Growth Factor I	*IGF1*
99	Insulin-like Growth Factor-Binding Protein 1	*IGFBP1*
100	Insulin-like Growth Factor-Binding Protein 2	*IGFBP2*
101	Insulin-like Growth Factor-Binding Protein 3	*IGFBP3*
102	Intercellular Adhesion Molecule 1	*ICAM1*
103	Interferon gamma	*IFNG*
104	Interferon gamma Induced Protein 10	*CXCL10*
105	Interferon-inducible T-cell alpha chemoattractant	*CXCL11*
106	Interleukin-1 beta	*IL1B*
107	Interleukin-1 receptor antagonist	*IL1RN*
108	Interleukin-10	*IL10*
109	Interleukin-12 Subunit p40	*IL12B*
110	Interleukin-16	*IL16*
111	Interleukin-17	*IL17A*
112	Interleukin-18	*IL18*
113	Interleukin-2	*IL2*
114	Interleukin-2 receptor alpha	*IL2RA*
115	Interleukin-23	*IL23A,IL12B*
116	Interleukin-6	*IL6*
117	Interleukin-6 receptor	*IL6R*
118	Interleukin-6 receptor subunit beta	*IL6ST*
119	Interleukin-7	*IL7*
120	Interleukin-8	*CXCL8*
121	Kallikrein 5	*KLK5*
122	Kidney Injury Molecule-1	*HAVCR1*
123	Lactoferrin	*LTF*
124	Lactoylglutathione lyase	*GLO1*
125	Latency-Associated Peptide of Transforming Growth Factor beta 1	*LTBP1*
126	Lectin-Like Oxidized LDL Receptor 1	*OLR1*
127	Leptin	*LEP*
128	Leptin Receptor	*LEPR*
129	Leucine-rich alpha-2-glycoprotein	*LRG1*
130	Luteinizing Hormone	*LHB,CGA*
131	Macrophage Colony-Stimulating Factor 1	*CSF1*

132	Macrophage inflammatory protein 3 beta	*CCL19*
133	Macrophage Inflammatory Protein-1 alpha	*CCL3*
134	Macrophage Inflammatory Protein-1 beta	*CCL4*
135	Macrophage Inflammatory Protein-3 alpha	*CCL20*
136	Macrophage Migration Inhibitory Factor	*MIF*
137	Macrophage-Derived Chemokine	*CCL22*
138	Macrophage-Stimulating Protein	*MST1*
139	Matrix Metalloproteinase-1	*MMP1*
140	Matrix Metalloproteinase-10	*MMP10*
141	Matrix Metalloproteinase-3	*MMP3*
142	Matrix Metalloproteinase-7	*MMP7*
143	Matrix Metalloproteinase-9	*MMP9*
144	Matrix Metalloproteinase-9 total	*MMP9*
145	Mesothelin	*MSLN*
146	Methylglyoxal	*NA*
147	MHC class I chain-related protein A	*MICA*
148	Monocyte Chemotactic Protein 1	*CCL2*
149	Monocyte Chemotactic Protein 2	*CCL8*
150	Monocyte Chemotactic Protein 3	*CCL7*
151	Monocyte Chemotactic Protein 4	*CCL13*
152	Monokine Induced by Gamma Interferon	*CXCL9*
153	Myeloid Progenitor Inhibitory Factor 1	*CCL23*
154	Myeloperoxidase	*MPO*
155	Myoglobin	*MB*
156	N-terminal prohormone of brain natriuretic peptide	*NPPB*
157	Neuronal Cell Adhesion Molecule	*NRCAM*
158	Neuropilin-1	*NRP1*
159	Neutrophil Activating Peptide 2	*PPBP*
160	Neutrophil Gelatinase-Associated Lipocalin	*LCN2*
161	Omentin	*ITLN1*
162	Osteocalcin	*BGLAP*
163	Osteopontin	*SPP1*
164	Osteoprotegerin	*TNFRSF11B*
165	P-Selectin	*SELP*
166	Pancreatic Polypeptide	*PPY*
167	Paraoxanase-1	*PON1*
168	Pentraxin-3	*PTX3*
169	Pepsinogen I	*NA*
170	Peptide YY	*PYY*
171	Periostin	*POSTN*
172	Peroxiredoxin-4	*PRDX4*
173	Phosphoserine Aminotransferase	*PSAT1*
174	Pigment Epithelium Derived Factor	*SERPINF1*
175	Plasminogen Activator Inhibitor 1	*SERPINE1*
176	Platelet-Derived Growth Factor BB	*PDGFB*
177	Progesterone	*NA*
178	Progranulin	*GRN*
179	Proinsulin Intact	*INS*
180	Proinsulin Total	*INS*
181	Prolactin	*PRL*
182	Prostasin	*PRSS8*
183	Prostatic Acid Phosphatase	*ACPP*
184	Protein S100-A4	*S100A4*
185	Protein S100-A6	*S100A6*
186	Pulmonary and Activation-Regulated Chemokine	*CCL18*
187	Receptor for advanced glycosylation end products	*AGER*
188	Receptor tyrosine-protein kinase erbB-3	*ERBB3*
189	Resistin	*RETN*
190	Retinol-binding protein 4	*RBP4*
191	Secreted frizzled-related protein 4	*SFRP4*
192	Selenoprotein P	*SEPP1*
193	Serotransferrin	*TF*
194	Serum Amyloid A Protein	*SAA1,SAA2,SAA4,SAA3P*
195	Serum Amyloid P-Component	*APCS*
196	Serum Glutamic Oxaloacetic Transaminase	*GOT1,GOT2*
197	Sex Hormone-Binding Globulin	*SHBG*
198	Sortilin	*SORT1*
199	ST2	*IL1RL1*
200	Stem Cell Factor	*KITLG*

201	Stromal cell-derived factor-1	*CXCL12*
202	Superoxide Dismutase 1 soluble	*SOD1*
203	T Lymphocyte-Secreted Protein I-309	*CCL1*
204	T-Cell-Specific Protein RANTES	*CCL5*
205	Tamm-Horsfall Urinary Glycoprotein (Uromodoulin)	*UMOD*
206	Tenascin-C	*TNC*
207	Testosterone Total	*NA*
208	Tetranectin	*CLEC3B*
209	Thrombin-activable fibrinolysis inhibitor	*CPB2*
210	Thrombomodulin	*THBD*
211	Thrombospondin-1	*THBS1*
212	Thyroid-Stimulating Hormone	*TSHB,CGA*
213	Thyroxine-Binding Globulin	*SERPINA7*
214	Tissue Inhibitor of Metalloproteinases 1	*TIMP1*
215	Tissue type Plasminogen activator	*PLAT*
216	TNF-Related Apoptosis-Inducing Ligand Receptor 3	*TNFRSF10C*
217	Transthyretin	*TTR*
218	Trefoil Factor 3	*TFF3*
219	Tumor Necrosis Factor alpha	*TNF*
220	Tumor necrosis factor receptor 2	*TNFRSF1B*
221	Tumor Necrosis Factor Receptor I	*TNFRSF1A*
222	Tyrosine kinase with Ig and EGF homology domains 2	*TIE1*
223	Urokinase-type Plasminogen Activator	*PLAU*
224	Urokinase-type plasminogen activator receptor	*PLAUR*
225	Vascular Cell Adhesion Molecule-1	*VCAM1*
226	Vascular Endothelial Growth Factor	*VEGFA*
227	Vascular Endothelial Growth Factor C	*VEGFC*
228	Vascular endothelial growth factor D	*FIGF*
229	Vascular Endothelial Growth Factor Receptor 2	*FLT1*
230	Vascular endothelial growth factor receptor 3	*FLT4*
231	Visceral adipose tissue derived serpin A12	*SERPINA12*
232	Visfatin	*NAMPT*
233	Vitamin D-Binding Protein	*GC*
234	Vitamin K-Dependent Protein S	*PROS1*
235	Vitronectin	*VTN*
236	von Willebrand Factor	*VWF*
237	YKL-40	*CHI3L1*

Supplementary Table 2: Summary of 46 local admixture signals associated with residualized biomarker levels at p <1.13x10^{-6}.

Biomarker	Ethnicity	Chr	BP	Admixture region	Top Admixture SNP	*Cis*	Beta	SE	P-value
Angiotensin-Converting Enzyme	Asian	17	60896263	59277614-63476044	rs8077625	YES	-0.49	0.09	4.4E-08
Antithrombin-III	African	11	124489274	123485444-125255715	rs35569094	NO	0.82	0.15	1.3E-07
Apolipoprotein(a)	Asian	6	161136058	159822140-162157860	rs4252105	YES	-0.57	0.09	1.2E-10
C-Peptide	African	2	238769680	237573038-240208154	rs3769050	NO	-0.76	0.15	4.2E-07
C-Peptide	African	9	107678278	107045058-108862046	rs4149261	NO	0.78	0.15	3.9E-07
Cellular Fibronectin	Asian	2	216141237	213413231-217588017	rs10498034	YES	0.49	0.08	1.1E-08
Chemokine CC-4	Asian	17	34450463	32917274-35272788	rs9303700	NO	-0.91	0.12	3.3E-14
Clusterin	African	8	127543114	126385776-128617612	rs11990954	NO	0.77	0.15	1.9E-07
Complement Factor H Related Protein 1	African	1	194377156	192323753-198618707	rs4460614	YES	-0.71	0.14	8.7E-07
Complement Factor H Related Protein 1	Asian	1	199588458	195832997-201081943	rs4612653	YES	0.60	0.08	4.6E-14
Eotaxin-2	Asian	7	78544172	77614334-80648073	rs1799022	NO	-0.44	0.08	8.8E-08
Eotaxin-3	African	6	100642867	99480633-102642021	rs6570574	NO	0.93	0.15	1.3E-09
Epithelial-Derived Neutrophil-Activating Protein 78	African	1	159069211	157463366-160580549	rs1894043	NO	1.28	0.14	9.4E-19
Fetuin-A	Asian	3	185951648	184173466-186880956	rs4686753	YES	-0.54	0.08	1.7E-11
Galectin-3	Asian	14	56803764	55172429-57657201	rs17091300	YES	0.46	0.09	1.9E-07
Growth Hormone	Asian	7	7813890	6780231-8356809	rs11771147	NO	-0.48	0.09	2.4E-07
Growth-Regulated alpha protein	African	1	159069211	157463366-160580549	rs1894043	NO	1.29	0.14	7.1E-19
Insulin-like Growth Factor-Binding Protein 1	Asian	6	151327848	150727790-152328616	rs505358	NO	-0.49	0.10	8.4E-07
Intercellular Adhesion Molecule 1	African	19	10219076	10000306-11456271	rs11666402	YES	-1.36	0.14	1.3E-21

Interferon-inducible T-cell alpha chemoattractant	Asian	4	76769442	75170140-78073655	rs924937	YES	-0.42	0.08	6.1E-07
Interleukin-16	Asian	15	81582868	80622567-85437979	rs8031107	YES	-0.58	0.09	2.8E-11
Interleukin-2	African	9	36824086	36066385-38639318	rs7848675	NO	0.72	0.14	4.5E-07
Interleukin-6 receptor	Asian	1	154551032	151316324-156848946	rs3811450	YES	0.83	0.08	2.4E-27
Interleukin-6 receptor subunit beta	African	5	55422681	54038676-55892384	rs321768	YES	-0.92	0.16	1.2E-08
Kallikrein 5	African	19	51455389	51173218-51582163	rs2569522	YES	-0.60	0.10	2.2E-09
Lactoferrin	African	3	46658737	42917047-54407040	rs9832679	YES	0.77	0.14	2.3E-08
Lactoylglutathione lyase	African	22	27689956	27361465-27957965	rs8136576	NO	-0.78	0.15	4.4E-07
Matrix Metalloproteinase-1	Asian	11	102460777	100718626-103789461	rs1711399	YES	-0.62	0.09	5.6E-12
Methylglyoxal	Asian	14	105163532	103784918-107287663	rs7140154	NO	-0.46	0.09	7.5E-07
MHC class I chain-related protein A	Asian	6	30347833	25292214-32974268	rs9404964	YES	0.51	0.08	1.2E-09
Monocyte Chemotactic Protein 1	African	1	159069211	157463366-160580549	rs1894043	NO	-0.86	0.15	1.5E-08
Monocyte Chemotactic Protein 1	Asian	1	159069211	157463366-160580549	rs1894043	NO	-0.42	0.08	4.0E-07
Monocyte Chemotactic Protein 2	Asian	17	32375998	32000619-32944070	rs6505393	YES	-0.67	0.10	1.2E-11
Monocyte Chemotactic Protein 4	Asian	1	159069211	157463366-160580549	rs1894043	NO	-0.49	0.08	2.2E-09
Myeloid Progenitor Inhibitory Factor 1	African	17	34450463	32917274-35272788	rs9303700	YES	-1.01	0.13	1.2E-13
Omentin	African	1	160696304	159751104-162458412	rs503832	YES	0.90	0.14	4.4E-10
Pigment Epithelium Derived Factor	Asian	17	1673104	907315-2883607	rs11658342	YES	-0.49	0.08	7.6E-09
Sortilin	African	11	79803310	78804058-81223717	rs10897545	NO	-0.85	0.17	3.6E-07
T-Cell-Specific Protein RANTES	African	1	159069211	157463366-160580549	rs1894043	NO	1.12	0.15	5.0E-14
T-Cell-Specific Protein RANTES	African	5	17235998	16706530-18833633	rs298528	NO	0.79	0.16	5.6E-07
Tetranectin	Asian	3	44347833	42166169-53119703	rs11921568	YES	-0.73	0.08	6.8E-22
Vascular Endothelial Growth Factor	Asian	6	44235690	43823689-46045157	rs513688	YES	-0.63	0.09	1.8E-13

Visceral adipose tissue derived serpin A12	Asian	14	94940717	94417541-95360139	rs8021858	YES	0.56	0.10	8.6E-09
Vitamin D-Binding Protein	African	4	73821904	72405132-76254408	rs13102676	YES	-1.13	0.15	5.9E-14
Vitamin D-Binding Protein	Asian	4	72425863	71237421-74909180	rs1453458	YES	-2.48	0.24	4.9E-24
Vitamin D-Binding Protein	Asian	4	72801285	71509980-75639740	rs17774208	YES	1.88	0.21	1.1E-18

Beta per 1% increase in associated ethnicity (corresponding to ethnicity column). Biomarkers residualized for age and sex. Association for biomarkers with more than one local association correspond to models with all local components included. Admixture region defined as r^2 >0.8 with top admixture signal. Biomarker in *cis* if gene encoding biomarker is within admixture region. SE: standard error.

Supplementary Table 3: Summary of genotypic associations in ORIGIN Europeans with the identified 46 local associations.

Biomarker	Ethnicity	Top Admixture SNP	European Associated SNP	MAF	Gene	Effect Allele	Other Allele	Beta	SE	P-value
Angiotensin-Converting Enzyme	Asian	rs8077625	rs4359	0.45	*ACE*	C	T	-0.67	0.03	8.44E-96
Antithrombin-III	African	rs35569094	rs1939923	0.09	*OR6M1*	G	A	0.22	0.05	5.0E-05
Apolipoprotein(a)	Asian	rs4252105	rs55730499	0.06	*LPA*	T	C	1.14	0.07	4.2E-54
C-Peptide	African	rs3769050	rs744837	0.02	*PER2*	C	T	0.40	0.12	5.7E-04
C-Peptide	African	rs4149261	rs181981786	0.01	*SLC44A1*	A	G	-0.76	0.21	3.0E-04
Cellular Fibronectin	Asian	rs10498034	rs12468524	0.24	*FN1*	T	C	0.79	0.04	3.5E-100
Chemokine CC-4	Asian	rs9303700	rs112689088	0.08	*CCL16*	C	T	-1.41	0.07	8.0E-91
Clusterin	African	rs11990954	rs2735974	0.15	*FAM84B*	C	T	0.17	0.04	5.5E-05
Complement Factor H Related Protein 1	African	rs4460614	rs149369377	0.22	*SRGAP2D, FAM72C*	G	A	-1.34	0.04	5.3E-258
Complement Factor H Related Protein 1	Asian	rs4612653	rs149369377	0.22	*SRGAP2D, FAM72C*	G	A	-1.34	0.04	5.3E-258
Eotaxin-2	Asian	rs1799022	rs12666387	0.35	*SEMA3C*	G	A	0.13	0.03	1.8E-04
Eotaxin-3	African	rs6570574	rs62420667	0.02	*ASCC3*	T	C	1.64	0.34	1.0E-11
Epithelial-Derived Neutrophil-Activating Protein 78	African	rs1894043	rs12075	0.44	*SRGAP2D, FAM72C, ACKR1*	A	G	0.31	0.03	2.2E-25
Fetuin-A	Asian	rs4686753	rs1900618	0.32	*AHSG*	T	C	0.66	0.03	1.9E-88
Galectin-3	Asian	rs17091300	rs141460994	0.08	*DLGAP5*	G	GA	-1.22	0.06	1.7E-101
Growth Hormone	Asian	rs11771147	7:7958913	0.25	*RPA3OS*	A	ACT	0.13	0.04	9.3E-04
Growth-Regulated alpha protein	African	rs1894043	rs12075	0.44	*SRGAP2D, FAM72C, ACKR1*	A	G	0.37	0.03	3.4E-32
Insulin-like Growth Factor-Binding Protein 1	Asian	rs505358	rs9371198	0.22	*PLEKHG1*	C	T	0.11	0.04	2.7E-03
Intercellular Adhesion Molecule 1	African	rs11666402	rs17852402	0.15	*ICAM5*	A	G	0.24	0.05	1.3E-06
Interferon-inducible T-cell alpha chemoattractant	Asian	rs924937	rs11947481	0.20	*SOWAHB*	A	G	0.17	0.04	1.3E-05
Interleukin-16	Asian	rs8031107	rs11556218	0.08	*IL16*	G	T	-0.97	0.06	1.4E-61

Interleukin-2	African	rs7848675	rs138591782	0.02	*FAM95C*	A	G	5.46	0.67	1.1E-125
Interleukin-6 receptor	Asian	rs3811450	rs2228145	0.36	*SRGAP2D, IL6R, FAM72C*	C	A	0.91	0.03	9.5E-164
Interleukin-6 receptor subunit beta	African	rs321768	rs6875155	0.12	*IL6ST*	A	G	0.53	0.05	2.5E-28
Kallikrein 5	African	rs2569522	rs34227821	0.36	*KLK5*	A	C	0.51	0.03	6.1E-51
Lactoferrin	African	rs9832679	rs4683228	0.31	*LTF*	A	C	0.53	0.03	2.3E-55
Lactoylglutathione lyase	African	rs8136576	rs5997202	0.13	*LINC01422*	A	T	-0.20	0.05	2.1E-04
Matrix Metalloproteinase-1	Asian	rs1711399	rs484915	0.44	*WTAPP1*	T	A	0.49	0.03	6.5E-50
Methylglyoxal	Asian	rs7140154	rs77844682	0.07	*ZFYVE21*	G	C	-0.19	0.07	2.8E-03
MHC class I chain-related protein A	Asian	rs9404964	rs114528635	0.02	*HLA-B*	A	C	3.84	0.20	3.6E-82
Monocyte Chemotactic Protein 1	African	rs1894043	rs12075	0.44	*SRGAP2D, FAM72C, ACKR1*	A	G	0.52	0.03	5.6E-59
Monocyte Chemotactic Protein 1	Asian	rs1894043	rs12075	0.44	*SRGAP2D, FAM72C, ACKR1*	A	G	0.52	0.03	5.6E-59
Monocyte Chemotactic Protein 2	Asian	rs6505393	rs11342894	0.16	*CCL8*	C	CA	-0.69	0.05	1.2E-50
Monocyte Chemotactic Protein 4	Asian	rs1894043	rs12075	0.44	*SRGAP2D, FAM72C, ACKR1*	A	G	0.49	0.03	3.0E-52
Myeloid Progenitor Inhibitory Factor 1	African	rs9303700	rs8073184	0.19	*CCL23*	T	C	-0.68	0.04	8.8E-60
Omentin	African	rs503832	rs1333062	0.32	*SRGAP2D ,FAM72C*	G	T	-0.76	0.03	3.8E-116
Pigment Epithelium Derived Factor	Asian	rs11658342	rs4274475	0.30	*SERPINF1*	T	C	-0.31	0.03	3.1E-20
Sortilin	African	rs10897545	rs201759795	0.08	*TENM4*	AT	A	0.35	0.08	2.0E-05
T-Cell-Specific Protein RANTES	African	rs1894043	rs12075	0.44	*SRGAP2D, FAM72C, ACKR1*	A	G	0.18	0.03	5.5E-09
T-Cell-Specific Protein RANTES	African	rs298528	rs17549158	0.09	*CDH18*	A	C	-0.21	0.06	5.7E-04
Tetranectin	Asian	rs11921568	rs3765173	0.41	*CLEC3B*	T	C	-1.01	0.03	3.4E-211
Vascular Endothelial Growth Factor	Asian	rs513688	rs9472168	0.44	*C6orf223*	G	A	-0.75	0.03	1.6E-117
Visceral adipose tissue derived serpin A12	Asian	rs8021858	rs61978271	0.25	*SERPINA12*	A	T	0.79	0.04	7.0E-105
Vitamin D-Binding Protein	African	rs13102676	rs222047	0.43	*GC*	A	C	-1.26	0.03	<1E-250
Vitamin D-Binding Protein	Asian	rs1453458	rs222047	0.43	*GC*	A	C	-1.26	0.03	<1E-250

Vitamin D-Binding Protein	Asian	rs17774208	rs222047	0.43	*GC*	A	C	-1.26	0.03	<1E-250

European associated SNP represents the strongest associated SNP in Europeans (n=1,931) within the admixture window, defined as r^2 >0.8 with top admixture signal. MAF: minor allele frequency, SE: standard error. Beta coefficient corresponds to the risk coefficient for each unit increase in effect allele, adjusted for age and sex and the first five principal components.

Supplementary Table 4: Summary of genotypic associations in ORIGIN Native Latins with the identified 46 local associations.

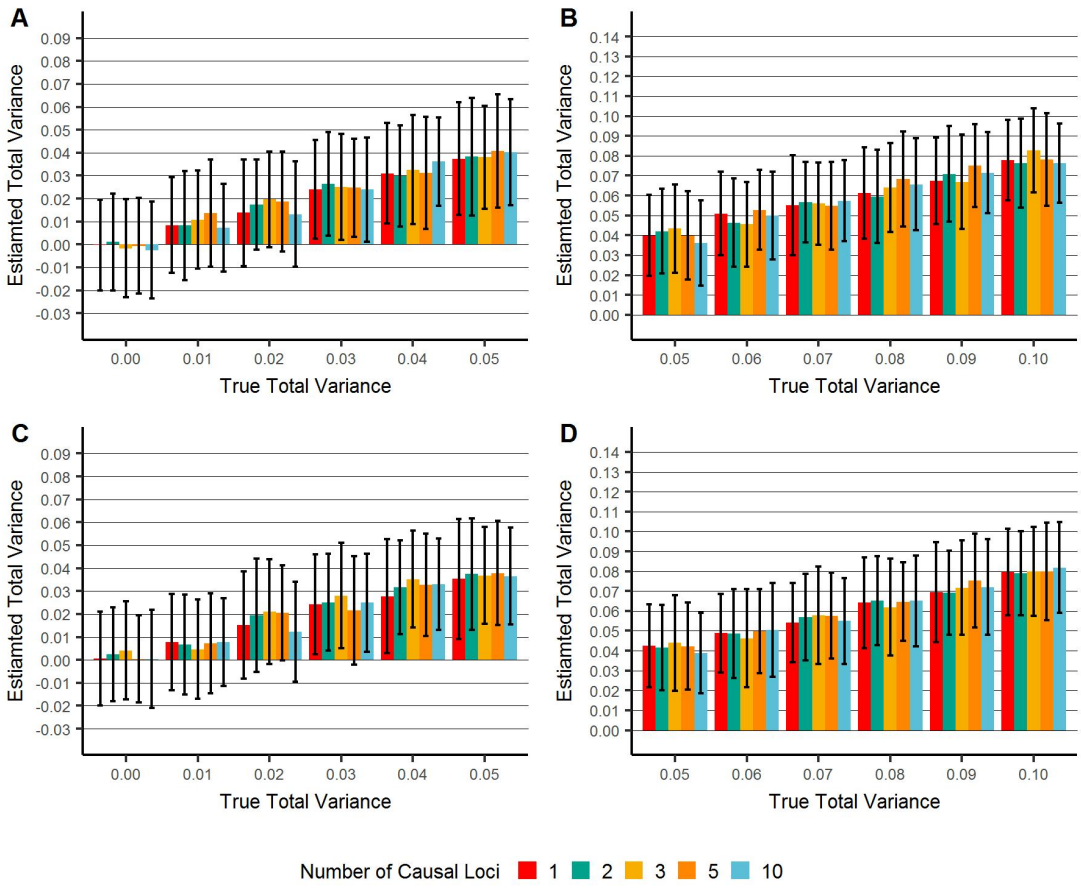
Biomarker	Ethnicity	SNP Admixture	GWAS SNP	MAF	Gene	Effect Allele	Other Allele	Beta	SE	P-value
Angiotensin-Converting Enzyme	Asian	rs8077625	rs4362	0.47	*ACE*	T	C	0.67	0.03	8.5E-108
Antithrombin-III	African	rs35569094	rs11605596	0.07	*PKNOX2*	A	G	-0.23	0.06	0.00031
Apolipoprotein(a)	Asian	rs4252105	rs6935921	0.43	*PLG*	C	T	-0.34	0.03	1.4E-25
C-Peptide	African	rs3769050	rs34371548	0.03	*ILKAP*	G	A	-0.26	0.09	0.0037
C-Peptide	African	rs4149261	rs111292742	0.01	*ABCA1*	C	G	0.35	0.12	0.0031
Cellular Fibronectin	Asian	rs10498034	rs2304573	0.36	*FN1*	G	A	0.75	0.04	5.2E-77
Chemokine CC-4	Asian	rs9303700	rs79254649	0.03	*CCL16*	A	C	-0.97	0.09	7.4E-26
Clusterin	African	rs11990954	rs6983561	0.05	*PRNCR1*	C	A	0.23	0.08	0.0030
Complement Factor H Related Protein 1	African	rs4460614	rs7519501	0.13	*SRGAP2D,FAM72C*	G	A	-1.14	0.04	8.9E-184
Complement Factor H Related Protein 1	Asian	rs4612653	rs7519501	0.13	*SRGAP2D,FAM72C*	G	A	-1.14	0.04	8.89E-184
Eotaxin-2	Asian	rs1799022	rs757865	0.35	*MAGI2*	T	C	0.08	0.03	0.0075
Eotaxin-3	African	rs6570574	rs9375894	0.08	*FBXL4*	A	G	0.05	0.02	0.0094
Epithelial-Derived Neutrophil-Activating Protein 78	African	rs1894043	rs12075	0.44	*SRGAP2D,FAM72C,ACKR1*	G	A	-0.28	0.03	8.6E-25
Fetuin-A	Asian	rs4686753	rs4918	0.37	*AHSG*	G	C	-0.65	0.03	1.1E-87
Galectin-3	Asian	rs17091300	rs11125	0.06	*LGALS3*	T	A	-1.15	0.05	1.8E-94
Growth Hormone	Asian	rs11771147	rs4236406	0.39	*GLCCI1*	G	A	-0.12	0.04	0.0045
Growth-Regulated alpha protein	African	rs1894043	rs12075	0.44	*SRGAP2D,FAM72C,ACKR1*	G	A	-0.27	0.03	5.9E-21
Insulin-like Growth Factor-Binding Protein 1	Asian	rs505358	rs10484921	0.22	*ESR1*	A	C	-0.13	0.04	0.00026
Intercellular Adhesion Molecule 1	African	rs11666402	rs5491	0.02	*ICAM1*	T	A	-1.73	0.12	3.4E-48
Interferon-inducible T-cell alpha chemoattractant	Asian	rs924937	rs7685696	0.46	*FAM47E*	A	G	0.17	0.04	6.9E-05
Interleukin-16	Asian	rs8031107	rs11556218	0.13	*IL16*	C	A	-0.96	0.04	9.4E-95

Interleukin-2	African	rs7848675	rs1885492	0.11	*ALDH1B1*	C	T	0.06	0.02	0.00088
Interleukin-6 receptor	Asian	rs3811450	rs2228145	0.48	*SRGAP2D,IL6R,FAM72C*	C	A	0.88	0.03	4.3E-211
Interleukin-6 receptor subunit beta	African	rs321768	rs2228046	0.02	*IL6ST*	G	A	-1.48	0.15	1.9E-22
Kallikrein 5	African	rs2569522	rs2569522	0.44	*KLK5*	C	T	0.48	0.03	1.0E-54
Lactoferrin	African	rs9832679	rs1126478	0.42	*LTF*	G	A	0.48	0.03	8.6E-60
Lactoylglutathione lyase	African	rs8136576	rs79136	0.48	*LINC01422*	A	G	-0.11	0.03	0.00096
Matrix Metalloproteinase-1	Asian	rs1711399	rs479095	0.41	*WTAPP1*	T	C	-0.43	0.03	4.3E-41
Methylglyoxal	Asian	rs7140154	rs6575979	0.25	*MARK3*	A	G	-0.10	0.04	0.0051
MHC class I chain-related protein A	Asian	rs9404964	rs2256175	0.34	*MICA*	C	T	0.61	0.04	4.5E-60
Monocyte Chemotactic Protein 1	African	rs1894043	rs12075	0.44	*SRGAP2D,FAM72C,ACKR1*	G	A	-0.49	0.03	1.7E-55
Monocyte Chemotactic Protein 1	Asian	rs1894043	rs12075	0.44	*SRGAP2D,FAM72C,ACKR1*	G	A	-0.49	0.03	1.7E-55
Monocyte Chemotactic Protein 2	Asian	rs6505393	rs12602195	0.26	*CCL8*	G	A	-0.66	0.03	7.3E-79
Monocyte Chemotactic Protein 4	Asian	rs1894043	rs12075	0.44	*SRGAP2D,FAM72C,ACKR1*	G	A	-0.43	0.03	9.4E-44
Myeloid Progenitor Inhibitory Factor 1	African	rs9303700	rs1003645	0.27	*CCL23*	G	A	-0.53	0.03	3.1E-50
Omentin	African	rs503832	rs1333062	0.30	*SRGAP2D,FAM72C*	T	G	0.72	0.03	6.0E-127
Pigment Epithelium Derived Factor	Asian	rs11658342	rs8074840	0.33	*SERPINF1*	C	T	-0.29	0.03	2.1E-19
Sortilin	African	rs10897545	rs12289861	0.15	*TENM4*	C	T	-0.10	0.04	0.018
T-Cell-Specific Protein RANTES	African	rs1894043	rs34599082	0.02	*SRGAP2D,FAM72C,ACKR1*	A	G	0.50	0.10	1.7E-06
T-Cell-Specific Protein RANTES	African	rs298528	rs167214	0.32	*BASP1*	G	A	-0.09	0.03	0.0021
Tetranectin	Asian	rs11921568	rs13963	0.47	*CLEC3B*	A	G	-0.95	0.02	1.6E-302
Vascular Endothelial Growth Factor	Asian	rs513688	rs7767396	0.49	*C6orf223*	A	G	0.74	0.03	3.2E-140
Visceral adipose tissue derived serpin A12	Asian	rs8021858	rs4905214	0.18	*SERPINA12*	A	G	0.66	0.04	1.2E-58
Vitamin D-Binding Protein	African	rs13102676	rs705120	0.37	*GC*	A	C	-1.12	0.02	<1E-250
Vitamin D-Binding Protein	Asian	rs1453458	rs705120	0.37	*GC*	A	C	-1.12	0.02	<1E-250

Vitamin D-Binding Protein	Asian	rs17774208	rs705120	0.37	*GC*	A	C	-1.12	0.02	<1E-250

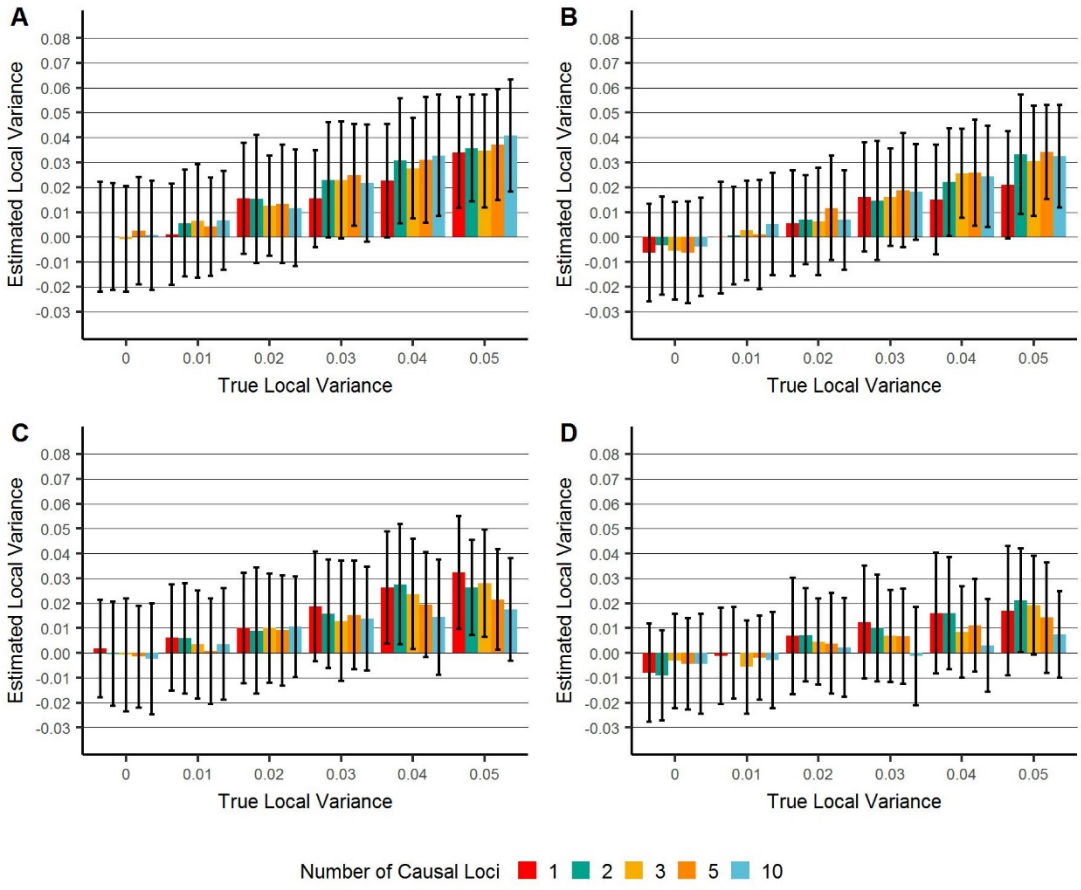
Native Latin associated SNP represents the strongest associated SNP in Europeans (n=2,216) within the admixture window, defined as r^2 >0.8 with top admixture signal. MAF: minor allele frequency, SE: standard error. Beta coefficient corresponds to the risk coefficient for each unit increase in effect allele, adjusted for age, sex and corresponding local Asian and African components.

Supplementary Figure 1: Estimated proportion of variance explained by local and global ancestry under various conditions with a simulated ordinal phenotype.



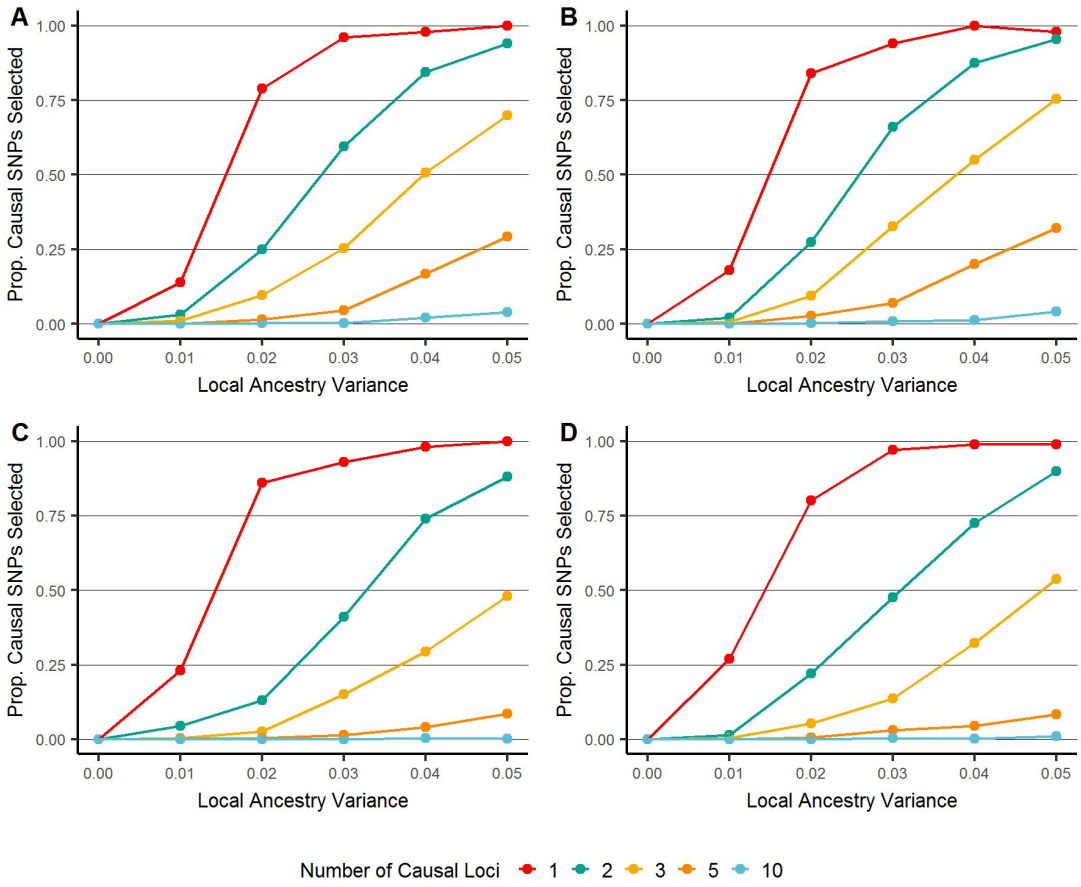
Average (±SD) estimated local and global variance explained under various simulated conditions. The sum of the true, unobserved local and global variances were pre-specified. Global variance was set at either 0 (panel A and C) or 0.05 (panel B and D) and local varied as 0.01, 0.02, 0.03, 0.04 and 0.05 as determined by 1, 2, 3, 5, or 10 causal SNPs. The sum of the two variances is shown on the x-axis. Each bar represents an average of 100 simulations, error bars show ±SD. Panels A and B illustrates simulated conditions with no directional condition, while panels C and D restrict local effects to be greater than 0. Panel A and C illustrate simulated conditions with no global effect, and Panel B and D have a pre-specified global effect.

Supplementary Figure 2: Estimated proportion of variance explained by local ancestry under various conditions with a simulated ordinal phenotype.



Average (±SD) estimated local variance explained under various simulated conditions. True, unobserved local variances were pre-specified at 0, 0.01, 0.02, 0.03, 0.04 and 0.05 (x-axis) as determined by 1, 2, 3, 5, or 10 causal SNPs. Each bar represents an average of 100 simulations, error bars show ±SD. Panels A and B illustrates simulated conditions with no directional condition, while panels C and D restrict local effects to be greater than 0. Panel A and C illustrate simulated conditions with no global effect, and Panel B and D have a pre-specified global effect.

Supplementary Figure 3: Proportion of causal SNPs selected under various conditions with a simulated ordinal phenotype.



Proportion of selected SNPs which were causal or regional selected by the forward selection algorithm under various simulated conditions. Proportion of causal SNPs (y-axis) was calculated as: (number of selected causal SNPs + number of selected regional SNPs) / (number of true, unobserved causal SNPs). True, unobserved local variances were pre-specified at 0, 0.01, 0.02, 0.03, 0.04 and 0.05 (x-axis) as determined by 1, 2, 3, 4, or 5 causal SNPs. Panels A and B illustrates simulated conditions with no directional condition, while panels C and D restrict local effects to be greater than 0. Panel A and C illustrate simulated conditions with no global effect, and Panel B and D have a pre-specified global effect.

REFERENCES

1. Paré G, Asma S, Deng WQ: Contribution of large region joint associations to complex traits genetics. *PLoS Genet.* 11: e1005103, 2015
2. Hsu F, Kent JW, Clawson H, Kuhn RM, Diekhans M, Haussler D: The UCSC known genes. *Bioinformatics* 22: 1036–1046, 2006
3. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D: GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics* 14: 656–664, 1998