

Genomic and epigenomic prediction of cardiovascular risk and modulation by diet

*A dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy*

Kenneth Westerman, M.S.

**Gerald J. and Dorothy R. Friedman School of Nutrition Science and Policy
at Tufts University**

**Jean Mayer-USDA Human Nutrition Research Center on Aging
Boston, MA**

August 2019

Thesis Advisor:

José M. Ordovás

Thesis Committee:

Dawn L. DeMeo, M.D., M.P.H.

Paola Sebastiani, Ph.D.

Paul Jacques, D.Sc.

Abstract

Personalized dietary approaches to cardiovascular disease (CVD) prevention will require sensitive biomarkers of multiple types. First, direct CVD risk biomarkers will enable the prioritization of dietary interventions for specific individuals. Second, biomarkers of diet response will allow more effective dietary recommendations for maintaining optimal cardiovascular risk factor (CRF) levels. Both genomic variants and epigenomic factors such as DNA methylation have the potential to enable improvements in both of these biomarker types.

Aim 1 of this work involved an investigation of mechanistic links between DNA methylation and incident CVD events. Differentially-methylated regions (DMRs) and the WGCNA module-based clustering approach were used to derive robust and interpretable epigenomic features before testing associations with CVD in the Women's Health Initiative and Framingham Heart Study Offspring Cohort. Multiple novel signals were uncovered, including differential methylation near immune function-related genes (*SLC1A5*, *SLC9A1*, and *TNRC6C*) and of development- and immune-related epigenomic modules. Module methylation was also linked to cumulative CRF exposure, providing proof-of-concept for the use of DNA methylation as a molecular recorder of cumulative risk factors over the life course.

Aim 2 expanded on these findings in seeking to directly optimize a predictor of CVD risk. Using time-to-CVD-event elastic net regression models, a methylation-based CVD risk score (MRS) was able to add predictive power after adjustment for traditional risk factors. A recently-introduced ensemble method for training predictive models in omic datasets showed comparable performance to a standard analysis pipeline. A preliminary analysis of interactions between the MRS and other CVD risk metrics (based on traditional risk factors or genomics) found that the MRS appears to perform better in lower-risk regions of these alternative metrics, suggesting its value in identifying individuals with "hidden" risk.

In Aim 3, genome-wide interaction study (GWIS) results were aggregated into genome-wide genetic scores for response to dietary fat for six CRFs. When testing these scores in the fat reduction-focused Dietary Modification trial from the Women's Health Initiative, only one of six scores (corresponding to LDL-cholesterol) showed meaningful prediction of one-year changes in the corresponding CRF, explaining a small proportion of the associated variance. Further longitudinal investigation of the LDL-FRS revealed suggestive interaction patterns between control and fat reduction arms for chronic disease outcomes, including support for a known divergence in LDL-C impact on different stroke subtypes.

An additional preliminary study from Aim 3 explored the epigenomic signature of response to vitamin K intake. After categorizing vitamin K responders and non-responders based on their changes in plasma phylloquinone during a supplementation trial, an epigenome-wide association study was performed to compare these groups. Differential DNA methylation was identified at known and novel loci including the *NPC1L1* gene, and was supported by a second EWAS for baseline plasma phylloquinone concentrations.

Taken together, these results build on and provide a more robust foundation for the development of improved omics-based biomarkers for CVD risk prediction and diet-centered risk reduction.

Table of Contents

Chapter 1	Introduction	1
1.1	Background and significance	2
1.2	Review of the literature	3
1.3	Central hypothesis and specific aims	11
1.4	References	12
Chapter 2	Epigenomics of CVD	21
2.1	DNA methylation modules associate with incident cardiovascular disease and cumulative risk factor exposure	22
Chapter 3	Epigenomic prediction of CVD risk	49
3.1	Epigenomic prediction of cardiovascular disease risk and interactions with traditional risk metrics	50
Chapter 4	Genomic and epigenomic biomarkers of diet response	72
4.1	A gene-diet interaction-based score predicts response to dietary fat in the Women's Health Initiative	73
4.2	Epigenome-wide association study reveals a molecular signature of response to vitamin K supplementation	91
Chapter 5	Summary	109
5.1	Summary of results	110
5.2	Implications	112
5.3	Methodological considerations and future directions	113
5.4	References	115

Tables and Figures

(Primary tables and figures only)

Table 2.1	FHS and WHI population characteristics.	24
Table 2.2	Modules associated with incident CVD at FDR < 0.02.	25
Table 2.3	Comb-p regions with multiple test-corrected $p < 0.05$ in WHI and Bonferroni $p < 0.05$ in FHS.	26
Table 2.4	Module-risk factor relationships (current and cumulative) after adjustment for covariates.	30
Table 2.5	CVD risk models using cumulative risk factor exposure and brown module activation.	31
Table 3.1	Baseline parameters of the populations used for model development.	52
Table 3.2	MRS performance in held-out FHS subset.	53
Table 3.3	Comparison of Cox regression coefficient estimates between the CSL and combined model options.	54
Table 3.4	Results from replication in REGICOR MI case-control	56
Table 4.1	WHI population description (European ancestry only).	75
Table 4.2	Responder score effects on CRF changes in DM trial participants.	76
Table 4.3	Phylloquinone supplementation trial population description.	95
Table 4.4	Top hits from responder EWAS.	96
Table 4.5	DMRs in responder/non-responder EWAS from Comb-p with uncorrected $p < 1e-5$.	97
Table 4.6	Baseline DMR results from Comb-p (regions with uncorrected $p < 1e-5$ shown).	99
Figure 1.1	Central hypothesis diagram.	12
Figure 2.1	DMRs identified by Comb-p in WHI and validated in FHS.	26
Figure 2.2	Genomic and epigenomic annotations of the brown and blue modules.	28
Figure 2.3	Risk factor-module relationships.	30
Figure 3.1	Computational workflow for MRS development and evaluation.	53
Figure 3.2	Kaplan-Meier survival curves in the held-out FHS-UM dataset.	54
Figure 3.3	Characterization of the final CSL model.	55
Figure 3.4	Interactions of MRS with other biomarkers of CVD risk.	57
Figure 4.1	LDL-FRS characterization.	77
Figure 4.2	LDL-FRS prediction of chronic disease development.	79
Figure 4.3	Genomic context of <i>NPC1L1</i> responder EWAS results.	97
Figure 4.4	Overlap between responder and baseline EWAS results.	100

Chapter 1

Introduction

1.1 Background and significance

Cardiovascular disease (CVD) is the leading cause of morbidity and mortality both domestically and globally. In 2016, CVD resulted in 2.7 million deaths in the US and 17.6 million deaths globally, with a prevalence (excluding hypertension) of 9% in US adults over the age of 20 (Benjamin et al. 2019). Thus, it is critically important that improvements be made in its identification, primary and secondary prevention, and treatment.

CVD risk prediction is excellent compared to many other diseases. Risk metrics such as the Framingham Risk Score (D'Agostino et al. 2008) and ACC/AHA ASCVD risk calculator (Goff et al. 2014) typically achieve c-statistics of 0.75-0.8 for predicting 10-year disease incidence in most populations. However, this risk score performance still leaves room for meaningful improvement if new data types can be incorporated into risk models, either reflecting complementary biological processes or demonstrating interactions between traditional risk factors and modifying factors such as behavioral inputs or genetic variants. Epigenomic factors such as DNA methylation are promising candidates, as they can reflect both genetic and environmental inputs (Dongen et al. 2016) and have already shown promise as clinical biomarkers for various cancers (DeVos et al. 2009).

Furthermore, the ability to predict risk at a given time point does not necessarily allow optimal preventative measures to be enacted during lower-risk periods earlier in life. For example, it has been demonstrated that cumulative exposure to high levels of LDL-cholesterol (LDL-C) over time is detrimental to cardiovascular health beyond simply correlating with high LDL-C immediately prior to disease onset. In order to optimize cardiovascular risk factors earlier in life and avoid compounding cumulative effects over time, it is important to resolve the contributors to interindividual variability in response to preventative behavioral choices.

As diet is a critical modifiable behavior influencing cardiovascular risk (Benjamin et al. 2019), much work has gone into identifying the source of differential diet responses across individuals. An abundance of studies has examined statistical interactions between dietary variables and genetic variants in predicting cardiometabolic risk factors (Parnell et al. 2014). However, in order to be effective in driving the practical implementation of personalized nutrition approaches, these interactions will need to be replicated, predict response to dietary interventions, and be combined into scores that can aggregate signal across many variants throughout the genome.

In order to enable personalized dietary approaches to CVD prevention, it will be necessary to develop sensitive biomarkers of multiple types. First, direct risk biomarkers will enable the prioritization of dietary interventions for specific individuals at the greatest risk for future disease. Epigenomic markers are well-positioned to complement current risk metrics in this regard, but would need to be validated more directly as prospective predictors of incident disease with grounding in plausible biological mechanisms. Second, biomarkers of diet response will allow increasingly personalized and effective dietary recommendations for maintaining optimal cardiovascular risk factor levels. Both genomic variants and epigenomic factors have the potential to make critical impacts on this less well-studied field. Integration of risk and response biomarkers will allow for preventive

dietary approaches to be undertaken earlier in life and in a more targeted and effective manner.

1.2 Review of the literature

1.2.1 Cardiovascular disease is complex and multifactorial

Cardiovascular disease etiology and risk factors

CVD encompasses a series of related illnesses of the heart and blood vessels. Specifically, it includes coronary artery diseases (most notably myocardial infarction and angina), heart failure, stroke, arrhythmias, and thrombosis. Linking many of these diseases is atherosclerosis, or a buildup of plaque leading to a narrowing of the artery. Atherosclerosis ultimately causes damage through the restriction of blood flow resulting in ischemia, or a decreased oxygen supply, in tissues such as the heart and brain.

Despite the mechanistic complexities and differences across various cardiovascular diseases, atherosclerotic and ischemic heart diseases share a series of core biological risk pathways. First, vascular health is an important risk modifier. Endothelial cells produce adhesion factors that become an upstream cause of atherosclerosis by recruiting immune cells. They also produce and respond to vasoactive substances, such as nitric oxide and endothelin-1, to regulate vascular tone, while their permeability affects the accumulation of atherosclerotic factors in the vessel wall (Cahill and Redmond 2016). Second, cholesterol metabolism and transport directly affect this cholesterol accumulation. Low-density lipoprotein (LDL) is the major source of the cholesterol that ultimately accumulates in artery walls, while high-density lipoprotein (HDL) functions in reverse cholesterol transport to oppose this process. Third, high blood pressure, or hypertension, results in hypertrophy, inflammation, and oxidative stress in the arterial wall (Alexander 1995). Finally, inflammation acts as both a downstream and upstream link between all of these factors. Inflammation leads to increased endothelial dysfunction, repressed reverse cholesterol transport, adhesion and uptake of monocytes in vessel walls, and hypertension. Conversely, each of these factors has the potential to result in increased inflammation in a particularly vicious positive feedback loop (Tall and Yvan-Charvet 2015).

These low-level biological mechanisms sit downstream of a series of known risk factors. Age is the strongest single risk factor for most CVDs, while body mass index (BMI) and obesity, hypertension (as both a mechanistic and clinical risk factor), and lipid levels play important roles. Relevant lipids include low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), and triglycerides (TG). Genetics plays an important role in modifying these risk factors and clinical CVD risk, suggested by the moderate heritability of most CVDs and exemplified by successful risk stratification through polygenic scores (Khera et al. 2018). Modifiable lifestyle behaviors, including smoking, physical activity, and diet, account for a high proportion of cardiovascular risk. For example, analyses from the Global Burden of Disease Study suggest that 74% of stroke risk is attributable to these modifiable factors. Exposures during development, such as nutritional stress during gestation or emotional stress during early childhood, represent another source of risk that is being increasingly recognized (Murphy, Cohn, and Loria

2017). Accumulation of risk factors over time, owing to any of these causes, has also been demonstrated to associate with cardiovascular risk independently of current risk factor levels (Reinikainen et al. 2015).

1.2.2 Epigenomics for disease risk prediction

Epigenetics

Epigenetics refers to a group of cellular systems that confer heritable regulation of transcription. Typically, its definitions and interpretations revolve around the idea of cellular reprogramming and/or cellular memory (Lappalainen and Greally 2017). It encompasses a series of components and systems, including chemical modification of histones, methylation of DNA, non-coding RNA (such as microRNA), and chromatin conformation (Allis and Jenuwein 2016). It is important to distinguish these more stable and potentially heritable features from simple transcriptional regulation, in which genes are transiently up- or down-regulated in response to environmental signaling (Lappalainen and Greally 2017).

The “-omics” suffix in epigenomics captures the common notion that in genome-wide assays, “everything” is being measured in some sense (Schwartz et al. 2012). While not literally true for almost all current assays, many can measure epigenetic quantities near almost all genes and through much of the non-coding portion of the genome as well. Epigenomic assays include sequencing-based methods, such as ChIP-seq for histone marks, Hi-C for chromatin conformation, and whole-genome bisulfite sequencing for DNA methylation, as well as microarray-based methods, most notably those measuring DNA methylation.

DNA methylation biology

DNA methylation is a particularly attractive epigenetic biomarker of chronic disease risk due to its ease of measurement and semi-stable nature. These chemical marks are interindividually variable, associate with common chronic disease risk factors such as BMI, and can be stable over years to decades (Talens et al. 2010; Feinberg et al. 2010).

Biochemically, DNA methylation refers to the addition of a methyl group to DNA nucleotides, almost always at the cytosine of cytosine-phosphate-guanine (CpG) sites. CpG sites are generally depleted in the genome compared to statistical expectation. However, many CpGs are found at short (about 1kb long), CpG-rich regions enriched referred to as CpG islands (Jones 2012), which appear in approximately half of vertebrate genes. DNA methylation regulates gene expression through interactions with transcription factors and transcriptional machinery. The directional relationship between DNA methylation and gene expression is complex; the location of CpG marks with respect to genes appears to relate to their expected effect. For example, methylation at transcription factor binding sites in gene promoters typically downregulates transcription of the associated gene, while methylation in gene bodies tends to associate with increased gene expression (Jones 2012). Additionally, population-level studies in humans tend to highlight CpG sites with less direct effects on transcription. This is likely due to stronger evolutionary constraints against variability at loci with large effects (such as those in promoter CpG islands), leaving greater

interindividual methylation variability at less-constrained regions such as CpG shores and shelves (flanking CpG islands) and “open sea”/non-genic regions.

CpG sites are methylated by DNA methyltransferases (DNMTs). Specifically, DNMT1 methylates hemi-methylated DNA, most notably to ensure the propagation of methylation patterns during DNA replication. DNMT3A and DNMT3B perform de novo methylation based on recruitment by a complex set of transcription factors (Zhong, Agha, and Baccarelli 2016). Active demethylation is performed in a series of steps: ten-eleven-translocase (TET) proteins oxidize methyl groups to hydroxymethyl groups, then the hydroxymethylated base is excised (directly or after deamination) and replaced with unmethylated cytosine (Wu and Zhang 2010). Demethylation can also occur passively, due to a lack of faithful propagation of existing methyl marks during replication. Notably, DNA hydroxymethylation is being increasingly recognized as a possible independent regulatory mark, rather than simply an intermediate step in the process of active demethylation. However, this modification is much less common in the human genome and has not received as much attention to-date (Shukla, Sehgal, and Singh 2015).

DNA methylation reflects environmental exposures

While interindividual variation in DNA methylation may be due to genetic influences (e.g. methylation quantitative trait loci, or mQTLs), much of the scientific interest in its behavior relates to its reflection and possibly mediation of environmental influences. Environmental factors, i.e. all non-genetic influences, may account for up to 80% of interindividual variation in DNA methylation levels (Dongen et al. 2016). Early life is an important period for the establishment of DNA methylation patterns that may persist and modulate disease risk much later in life (Kessler et al. 2018; Zhong, Agha, and Baccarelli 2016). This is one of the mechanisms thought to contribute to the observed developmental origins of CVD.

Current prediction methods can explain only small amounts of the large variation attributable to environment. However, specific environmental factors have been shown to influence methylation patterns even into adulthood. These include stress (Matosin, Cruceanu, and Binder 2017) and exercise (Rönn et al. 2013). Diet can also impact methylation, both through direct mechanistic avenues (e.g. folate as a source of methyl groups for one-carbon metabolism (Jacob et al. 1998)) and more indirect or systemic effects (e.g. caloric restriction/weight loss (Milagro et al. 2011) and dietary fat composition (Perfilyev et al. 2017)).

Measurement of DNA methylation

Genome-wide assays for DNA methylation are typically based on bisulfite conversion (Kurdyukov and Bullock 2016). In this process, bisulfite treatment of DNA converts unmethylated cytosines to uridine and ultimately thymine after polymerase chain reaction (PCR) amplification. Comparison of bisulfite-converted DNA to a reference sequence then allows the determination of the methylation status of a given CpG site. Bisulfite-treated DNA can then be quantified using sequencing or microarray techniques. Whole-genome bisulfite sequencing is the most complete, while strategies such as reduced representation bisulfite sequencing (RRBS) and methylated DNA immunoprecipitation sequencing

(MeDIP-seq) present somewhat more cost-efficient alternatives that aim to capture specifically (methyl-) CpG-rich portions of the genome. Though sequencing-based approaches continue to drop in cost, more cost-efficient microarrays have been the de facto method of choice for genome-wide epigenomic interrogation in large numbers of subjects. Specifically, a series of Illumina microarrays have been the most prevalent over the last decade, providing fairly robust methylation estimates at 27 thousand to 1 million CpGs (~5-10% of the CpG sites in the genome). These include the HumanMethylation27K (Bibikova et al. 2009), HumanMethylation450 (Bibikova et al. 2011), and MethylationEPIC (Pidsley et al. 2016) chips. For smaller-scale measurement at specific loci, other bisulfite-based approaches including pyrosequencing, PCR-based techniques, and mass spectrometry-based techniques, can be used (Kurdyukov and Bullock 2016; Suchiman et al. 2015); these are typically relevant for hypothesis-based investigations or in confirmatory analyses.

Epigenome-wide association studies

The advent of cost-efficient microarrays measuring DNA methylation genome-wide has led to sample sizes large enough to support epigenome-wide associations studies (EWAS). EWAS are the epigenomic equivalent of genome-wide association studies (GWAS), in which univariate associations between a molecular quantity and a phenotype are assessed at a series of genetic loci (Rakyan et al. 2012). In a typical EWAS, a regression model is used to test the association between methylation at a CpG site and some phenotype. This is repeated over many loci, ultimately resulting in a genome-wide map of epigenetic associations.

While many variants exist, a set of standard preprocessing steps has emerged to facilitate more consistency in EWAS approaches and limit spurious results due to technical factors (Morris and Beck 2015). Basic quality control steps are similar to those for other 'omics technologies, with both samples and sites being removed if they are missing excessive amounts of data points and if they contain an excessive proportion of measurements whose microarray signal is statistically indistinguishable from baseline fluorescent noise. Some correction for technical effects can be accomplished through sample normalizations taking advantage of non-biological signal, such as that from out-of-band fluorescence (Triche et al. 2013) or microarray control probes (Fortin et al. 2014). A number of additional filters can then be applied to remove additional probes with high potential for cross-hybridization, nearby single nucleotide polymorphisms (SNPs), or other confounders. A further normalization step specific to recent Illumina methylation microarrays corrects for distributional differences between two different probe chemistries ("Type I" vs. "Type II"), and can be addressed using a number of algorithms (Teschendorff et al. 2013; Maksimovic, Gordon, and Oshlack 2012). The ultimate output of these preprocessing algorithms is a set of beta values (one per CpG site), representing an estimate of the fraction of chromosomes that are methylated within the input bulk sample; given heteroscedasticity in the beta values, these are sometimes converted through a logit transformation to "M-values".

Finally, standard procedures first developed for gene expression data can be applied to identify and adjust for batch effects and other technical and biological confounders. These include principal components analysis (PCA) to identify and possibly adjust for high-level effects, ComBat to estimate and remove variation associated with known batches (Johnson,

Li, and Rabinovic 2007), surrogate variable analysis (SVA) to create confounder-associated variables for statistical adjustment (Leek and Storey 2007), and many other similar strategies. Given that many EWAS are performed using samples taken from complex tissues, cell sub-types can be a major confounder of observed associations when the experimental hypothesis is not about cell proportion changes but rather epigenetic “reprogramming”. While cell counts are sometimes measured experimentally, highly accurate estimates of cell sub-type fractions in complex tissues (most notably blood) can be obtained from the epigenetic data themselves, to then be used as covariates during modeling. Various computational tools have been developed for this purpose, which can be based on known reference databases (Houseman et al. 2012) or inferred using unsupervised techniques by assuming that they will largely correspond to the principal axes of variation in the data (Houseman, Molitor, and Marsit 2014; Rahmani et al. 2016).).

DNA methylation as a disease biomarker

DNA methylation has been shown to act as an effective biomarker of disease and disease risk across multiple domains. Much of the existing research is related to cancer, which involves extensive epigenetic reprogramming and thus carries typically large effect sizes (Feinberg 2018). However, methylation-based predictors have also shown promise in the cardiometabolic realm, such as for type 2 diabetes risk (Wahl et al. 2017). These risk predictors can be developed using standard risk score methodology as with polygenic risk scores in which genome-wide significant loci are aggregated into a weighted sum based on their univariate effect sizes. For example, this approach was taken in the creation of a strikingly strong predictor of overall mortality (Zhang et al. 2017). DNA methylation datasets are also amenable to multivariate approaches that can perform reasonably in high-dimensional settings, such as the popular elastic net penalized regression method (Zou and Hastie 2005; Zhuang, Widschwendter, and Teschendorff 2012).

Most studies investigating DNA methylation and cardiovascular disease to-date have been cross-sectional. Some have been hypothesis-based, including links between lipid-related genes and coronary heart disease (CHD) (Peng et al. 2014) and between the “epigenetic clock” (a measure of biological age) and various CVD types (Horvath et al. 2016; Perna et al. 2016). Epigenome-wide studies have looked at myocardial infarction (MI) (Rask-Andersen et al. 2016), stroke (Gómez-Úriz et al. 2015), and various subcategories of CAD (Li et al. 2017; Sharma et al. 2014). Generally, most specific CpG sites and differentially-methylated genes related to CVD have not replicated well, but some themes have emerged at the pathway and disease process level, including RhoA signaling, endocrine function, cardiovascular development and function, inflammation, and immune cell trafficking (Fernández-Sanlés et al. 2017). The aforementioned studies focus primarily on blood DNA methylation due to its relative ease of collection compared to other cardiovascular tissues; important work has been done on other CVD-relevant tissues (e.g. aorta (Zaina et al. 2014)), though that will not be the primary focus here.

While cross-sectional studies are highly susceptible to “reverse causation”, in which the presence of a disease results in the observed methylation changes, one can investigate potential impact of methylation prior to disease onset through analysis of risk factors. Investigators have uncovered associations between methylation and lipids (Irvin et al. 2014; Hedman et al. 2017; Pfeiffer et al. 2015), blood pressure (Richard et al. 2017), BMI

(Wahl et al. 2017), postprandial lipid response to a meal (Lai et al. 2016), and circulating TNF- α (Aslibekyan et al. 2018). These studies have uncovered specific CpG sites with somewhat better replication across studies than those for CVD events, most notably cg06500161 in the *ABCG1* gene, which is robustly associated with multiple lipid traits and is predictive of incident CVD (Hedman et al. 2017).

DNA methylation and CVD

A few studies have evaluated direct prediction of incident cardiovascular events using DNA methylation. One observed strong associations of global methylation (using LINE-1 element methylation as a proxy) (Baccarelli et al. 2010), but relationships of global methylation in similar studies have been inconsistent (Kim et al. 2010). Another examined the relationship between genome-wide differentially-methylated regions and MI risk, demonstrating a modest improvement in predictive performance compared to standard risk factors. Multiple investigations have instead begun by linking CpGs to risk factors (as mentioned above), and only then examining whether top hits are also predictive of incident events (Hedman et al. 2017; Aslibekyan et al. 2018).

Mendelian Randomization (MR) is an alternative approach to suggesting causality of epigenetic associations. MR takes advantage of genetic variants as causal anchors (instrumental variables) to CpG-phenotype relationships. Here, genetic variants underlying strong mQTLs (genotype-CpG relationship) can be tested for association with the phenotype of interest to suggest causality of the epigenetic quantity, while the reverse hypothesis can be tested using a genetic instrumental variable for the phenotype of interest. One of the most prominent examples of MR-based systematic investigation into DNA methylation and CVD traits found potentially causal impacts at the *ADCY3* and *ADIPOQ* loci. Through bi-directional MR, another study provided evidence that blood lipids influence DNA methylation more commonly than vice versa (Dekkers et al. 2016). While avoiding some of the typical problems with cross-sectional analyses, MR analyses do have their own set of obstacles, including the potential for pleiotropy of the genetic variants as well as the major and untestable assumption that the instrumental variables have no association with confounders of the methylation-phenotype relationship [Bowden, Smith, and Burgess (2015)].

1.2.3 Genomics and epigenomics modulate response to interventions

Gene-by-environment interactions

In order to translate insights from genomics into improved treatments and behaviors by individuals, it is important to understand how molecular-level differences result in variable responses to environmental influences. Specifically, accurate prediction of responses to environmental exposures or interventions would allow the evolution of personalized medicine to encompass not only subject-specific risks, but also subject-specific treatments. This type of prediction is based on molecular and/or statistical associations in which genomic and epigenomic measurements (as well as other molecular quantities) modify environment-disease associations.

This type of association has a rich history in genetics, captured by the field of gene-by-environment (GxE) interactions. Univariate GxE interactions can be viewed through one

of two statistically equivalent lenses: as either a modification of environmental effects by genetic variants, or as a modification of genetic effects by environmental exposures. Furthermore, GxE investigations have been conducted in a variety of ways: as environment-focused (e.g. smoking (Do and Maes 2016)), phenotype-focused (e.g. asthma (Ober and Vercelli 2011)), or molecular phenotype-focused (e.g. based on gene expression as an outcome (Grishkevich and Yanai 2013)). In comparison to main-effect associations as in traditional GWAS, GxE interactions require substantially greater sample sizes (Dempfle et al. 2008). To achieve reasonable statistical power for case-control studies under typical parameter assumptions (such as interaction effect sizes), sample sizes in the thousands are required (Hunter 2005).

Variant prioritization techniques can be applied to reduce the multiple-testing burden in interaction studies by narrowing the set of possible loci of interest (Winkler et al. 2017). Hypothesis-based studies use a specific subset of variants based on prior biological knowledge of the gene(s) involved in the environmental factor and/or phenotype. Alternatively, resources such as CardioGxE provide repositories of known GxE interactions to guide future interaction studies (Parnell et al. 2014). Variants can also be filtered by some threshold of main effect; this strategy may improve power to detect quantitative or “pure” interactions, though it will be ineffective when the effects in each environmental stratum are in opposite directions (Winkler et al. 2017). The use of main-effect associations as a filtering strategy has shown mixed results in “real” large-scale datasets, such as in GxE meta-analyses from the CHARGE Consortium (Laville et al. 2019).

While GxE interactions in cross-sectional studies can provide important clues and allow much greater sample sizes, intervention trials enable much more direct querying of the questions of interest, such as the identification of “responders” to a given intervention or behavior. For example, genetic signatures of LDL-C response have been sought for pharmacological treatments including statins (Chasman et al. 2012) and fenofibrate (Aslibekyan et al. 2012).

Gene-diet interactions

Diet is a particularly important modifiable CVD risk factor, with estimates that up to 45% of cardiometabolic deaths in the US can be attributed to suboptimal dietary behaviors (Micha et al. 2017). The contribution of genetics to diet response has been convincingly demonstrated in mice, where separate strains show divergent cardiometabolic risk factor changes in response to a series of diets modeled after human dietary patterns (Barrington et al. 2018). In humans, GxE interactions related to dietary exposures, or gene-diet interactions (GDIs), have been a popular area of research, as their identification may inform genotype-based personalized nutrition. An abundance of studies have examined GDIs predicting cardiometabolic risk factors (Parnell et al. 2014).

Most of these studies investigate one or a small set of genetic variants based on plausible biology. For example, a Mediterranean diet pattern attenuates the detrimental effects of a *TCF7L2* variant on glycemic traits (Corella et al. 2013), and a variant in the *CYP1A2* gene modifies associations between caffeine intake and MI (Cornelis et al. 2006). Other studies employ a genetic risk score (GRS) for the cardiometabolic trait, including notable

explorations of the interaction between sugar-sweetened beverage intake and a GRS for obesity or adiposity influencing anthropometric traits (Qi et al. 2012; Olsen et al. 2016). Despite the difficulties with statistical power, some genome-wide gene-diet investigations (falling under the umbrella of genome-wide interaction studies, or GWIS) have also been conducted, including for various dietary associations with colorectal cancer (Figueiredo et al. 2014) and dairy associations with BMI (Smith et al. 2018).

Multiple conditions need to be met for these statistical GDIs to be practically useful for personalized nutrition. First, GDIs must be replicable across populations (which may or may not cross ancestral or cultural boundaries, depending on the intended scope of the application of a finding). Second, statistical interactions discovered in cross-sectional datasets must meaningfully predict the change in some outcome based on a change in a dietary behavior (the goal of the field of nutrigenetics). To this end, some studies have been dedicated to uncovering gene-diet interactions for changes in risk factors over the course of a dietary intervention trial. Two notable trials supporting many of these investigations include the PREDIMED study (Mediterranean diet intervention) (Corella et al. 2013) and the POUNDS LOST trial (weight-loss trial comparing macronutrient distributions) (Huang et al. 2016). Finally, a diet response score must account for the possible influence of genetic variants across the genome rather than those in just one or two genes of interest. This is common in main-effect polygenic risk scores (Khera et al. 2018), but has been historically difficult for responder scores owing to limited power in gene-environment interaction studies and limited sample size in dietary intervention trials. For example, Rudkowska and colleagues developed a genetic score to predict triglyceride response to omega-3 fatty acid supplementation in 141 subjects, but the resulting score could explain only 2% (non-statistically significant) of the variation in triglyceride response in an independent dataset (Rudkowska et al. 2014).

Epigenome-diet interactions

Despite the compelling potential biological mechanisms related to reflection of environmental or dietary exposures, comparatively little research effort has gone into epigenetic interactions and response prediction testing. This is likely due primarily to the lack of cohorts with sufficient sample sizes, given the aforementioned power problems with interaction studies. However, there are some examples in both the cancer field (*MGMT* methylation predicting response to dacarbazine in colorectal cancer (Amatu et al. 2013)) and the diet field (*TNF- α* methylation predicting weight loss in response to caloric restriction (Campión et al. 2009)). Epigenetics may also provide a mechanism underlying GxE associations, as has been uncovered in a series of studies related to genetics and body weight. Interactions between the rs5082 SNP in *APOA2* and saturated fat consumption influencing BMI have been discovered and replicated across cohorts, but without a clear mechanism (Corella et al. 2009, 2011). A recent epigenomic study provided additional clarity to this relationship, finding genotype-specific methylation at a particular CpG site only in high-saturated fat consumers along with potential downstream effects of this methylation difference on gene expression and plasma metabolite concentrations (Lai et al. 2018).

1.3 Central hypothesis and specific aims

Personalization of nutrition and behavior for cardiovascular disease (CVD) will require an improved understanding of who should be receiving treatments and what action can be taken by individuals based on their specific biology. Strong genetic predictors of CVD risk have recently been developed, but these do not take into account the large fraction of cardiovascular risk that has been consistently attributed to lifestyle factors including diet, exercise, and smoking. While multiple demographic, anthropometric, and blood-based risk factors have been strongly and causally linked to cardiovascular risk, they are not able to fully predict disease incidence nor the expected effect of specific lifestyle changes on specific individuals.

The long-term goal of this research is to move towards the ability to measure genomic quantities in individuals to assess cardiovascular risk and prescribe personalized nutritional changes. Our central hypothesis is two-fold: that epigenetic quantities can act as effective biomarkers of modifiable risk for CVD beyond current CVD risk prediction metrics, and that multi-omic markers can provide a basis for predicting the cardiometabolic response to dietary inputs. Together, these advances would enable the use of modern genomic techniques to provide earlier and more effective interventions for CVD prevention, and will be pursued in the following aims:

Aim 1: To conduct an epigenome-wide scan for incident cardiovascular disease events while using statistical techniques to enhance the interpretability of discoveries. The investigation will be conducted using DNA methylation microarray data from large-scale epidemiological cohorts including the Women's Health Initiative and Framingham Heart Study. Relationships with disease risk will be examined at the level of single sites, genetic regions, and correlation-based modules.

Aim 2: To assemble a multi-study DNA methylation-based risk score for incident cardiovascular disease. Statistical ensembling methods will be used to combine predictive models from a series of longitudinal studies. The resulting predictor will be compared to existing genomic and traditional risk factor-based risk metrics for redundancy and potential synergy.

Aim 3: To develop genomic response scores for dietary effects on cardiovascular risk factors. Gene-diet interactions will be assessed in cross-sectional data from the Women's Health Initiative and used to construct dietary fat response scores to be tested in the Women's Health Initiative Dietary Modification Trial. This process will explore the effectiveness of genome-wide aggregation of gene-environment interaction evidence and expand the investigation of diet response scores to chronic disease outcomes.

Taken together, these findings will provide important conceptual steps forward towards increasingly precise and personalized approaches to disease prevention. Aim 1 will develop new hypotheses for genetic exploration of CVD while deepening our understanding of existing mechanisms through epigenetic links to known risk factors. Aim 2 will provide an important benchmark for epigenomic prediction of CVD risk and establish the utility of epigenomic quantities in relation to other known risk predictors. Aim 3 will provide proof-of-concept for the utility of genome-wide predictors of response to diet for CVD risk reduction. Ultimately, these genomic and epigenomic biomarkers of cardiovascular risk

would represent key advances towards a more effective system for CVD prediction and prevention well before the onset of disease.

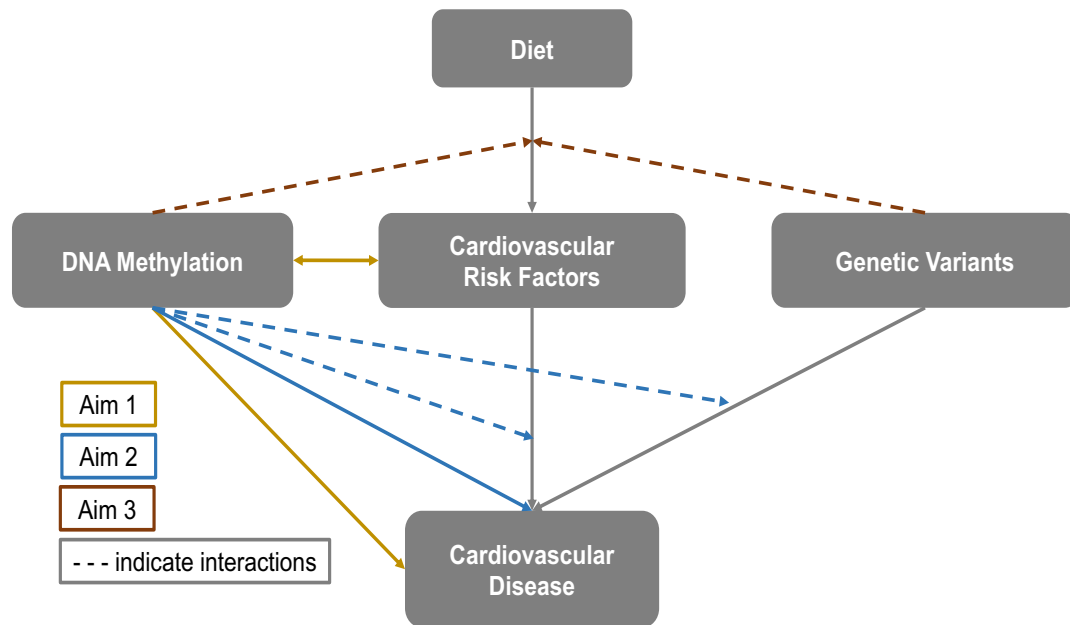


Diagram details the conceptual basis for the central hypothesis and specific aims of this research. Aim 1 will investigate the connection between DNA methylation and CVD at the level of regions and modules, while assessing potential epigenetic correlates of cardiovascular risk factors. Aim 2 will further explore the DNA methylation-CVD link in a more direct manner. Aim 3 will move towards response rather than risk biomarkers, seeking genetic variants that explain inter-individual variation in the known association between diet and CVD.

1.4 References

- Alexander, R. Wayne. 1995. "Hypertension and the Pathogenesis of Atherosclerosis." *Hypertension* 25 (2): 155–61. <https://doi.org/10.1161/01.HYP.25.2.155>.
- Allis, C. David, and Thomas Jenuwein. 2016. "The molecular hallmarks of epigenetic control." *Nature Reviews Genetics* 17 (8): 487–500. <https://doi.org/10.1038/nrg.2016.59>.
- Amatu, Alessio, Andrea Sartore-Bianchi, Catia Moutinho, Alessandro Belotti, Katia Bencardino, Giuseppe Chirico, Andrea Cassingena, et al. 2013. "Promoter CpG island hypermethylation of the DNA repair enzyme MGMT predicts clinical response to dacarbazine in a phase II study for metastatic colorectal cancer." *Clinical Cancer Research* 19 (8): 2265–72. <https://doi.org/10.1158/1078-0432.CCR-12-3518>.

Aslibekyan, Stella, Golareh Agha, Elena Colicino, Anh N. Do, Jari Lahti, Symen Ligthart, Riccardo E. Marioni, et al. 2018. "Association of Methylation Signals With Incident Coronary Heart Disease in an Epigenome-Wide Assessment of Circulating Tumor Necrosis Factor α ." *JAMA Cardiology* 3 (6): 463–72. <https://doi.org/10.1001/JAMACARDIO.2018.0510>.

Aslibekyan, Stella, Mark O. Goodarzi, Alexis C. Frazier-Wood, Xiaofei Yan, Marguerite R. Irvin, Eric Kim, Hemant K. Tiwari, et al. 2012. "Variants Identified in a GWAS Meta-Analysis for Blood Lipids Are Associated with the Lipid Response to Fenofibrate." *PLoS ONE*. <https://doi.org/10.1371/journal.pone.0048663>.

Baccarelli, A, R Wright, V Bollati, A Litonjua, A Zanobetti, L Tarantini, D Sparrow, P Vokonas, and J Schwartz. 2010. "Ischemic heart disease and stroke in relation to blood DNA methylation." *Epidemiology* 21 (6): 819–28. <https://doi.org/10.1097/EDE.0b013e3181f20457>.

Barrington, William T., Phillip Wulfridge, Ann E. Wells, Carolina Mantilla Rojas, Selene Y.F. Howe, Amie Perry, Kunjie Hua, et al. 2018. "Improving metabolic health through precision dietetics in mice." *Genetics*. <https://doi.org/10.1534/genetics.117.300536>.

Benjamin, Emelia J., Paul Muntner, Alvaro Alonso, Marcio S. Bittencourt, Clifton W. Callaway, April P. Carson, Alanna M. Chamberlain, et al. 2019. "Heart Disease and Stroke Statistics—2019 Update: A Report From the American Heart Association." *Circulation* 139 (10). <https://doi.org/10.1161/CIR.0000000000000659>.

Bibikova, Marina, Bret Barnes, Chan Tsan, Vincent Ho, Brandy Klotzle, Jennie M. Le, David Delano, et al. 2011. "High density DNA methylation array with single CpG site resolution." *Genomics* 98 (4): 288–95. <https://doi.org/10.1016/j.ygeno.2011.07.007>.

Bibikova, Marina, Jennie Le, Bret Barnes, Shadi Saedinia-Melnyk, Lixin Zhou, Richard Shen, and Kevin L. Gunderson. 2009. "Genome-wide DNA methylation profiling using Infinium assay." *Epigenomics* 1 (1): 177–200. <https://doi.org/10.2217/epi.09.14>.

Bowden, Jack, George Davey Smith, and Stephen Burgess. 2015. "Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression." *International Journal of Epidemiology* 44 (2). Oxford University Press: 512–25. <https://doi.org/10.1093/ije/dyv080>.

Cahill, Paul A., and Eileen M. Redmond. 2016. "Vascular endothelium - Gatekeeper of vessel health." <https://doi.org/10.1016/j.atherosclerosis.2016.03.007>.

Campión, Javier, Fermin I Milagro, Estibaliz Goyenechea, and J Alfredo Martínez. 2009. "TNF- α promoter methylation as a predictive biomarker for weight-loss response." *Obesity (Silver Spring, Md.)* 17 (6). Nature Publishing Group: 1293–7. <https://doi.org/10.1038/oby.2008.679>.

Chasman, Daniel I., Franco Giulianini, Jean MacFadyen, Bryan J. Barratt, Fredrik Nyberg, and Paul M. Ridker. 2012. "Genetic Determinants of Statin-Induced Low-Density Lipoprotein Cholesterol Reduction." *Circulation: Cardiovascular Genetics* 5 (2): 257–64. <https://doi.org/10.1161/CIRCGENETICS.111.961144>.

Corella, D., P. Carrasco, J. V. Sorli, R. Estruch, J. Rico-Sanz, M. A. Martinez-Gonzalez, J. Salas-Salvado, et al. 2013. "Mediterranean Diet Reduces the Adverse Effect of the TCF7L2-rs7903146 Polymorphism on Cardiovascular Risk Factors and Stroke Incidence: A randomized controlled trial in a high-cardiovascular-risk population." *Diabetes Care*, dc13–0955 –. <https://doi.org/10.2337/dc13-0955>.

Corella, Dolores, Gina Peloso, Donna K. Arnett, Serkalem Demissie, L. Adrienne Cupples, Katherine Tucker, Chao Qiang Lai, et al. 2009. "APOA2, dietary fat, and body mass index: Replication of a gene-diet interaction in 3 independent populations." *Archives of Internal Medicine*. <https://doi.org/10.1001/archinternmed.2009.343>.

Corella, D, E S Tai, J V Sorlí, S K Chew, O Coltell, M Sotos-Prieto, A García-Rios, R Estruch, and J M Ordovas. 2011. "Association between the APOA2 promoter polymorphism and body weight in Mediterranean and Asian populations: replication of a gene-saturated fat interaction." *International Journal of Obesity* 35 (5): 666–75. <https://doi.org/10.1038/ijo.2010.187>.

Cornelis, Marilyn C., Ahmed El-Sohemy, Edmond K. Kabagambe, and Hannia Campos. 2006. "Coffee, CYP1A2 genotype, and risk of myocardial infarction." *Journal of the American Medical Association*. <https://doi.org/10.1001/jama.295.10.1135>.

D'Agostino, Ralph B., Ramachandran S. Vasan, Michael J. Pencina, Philip A. Wolf, Mark Cobain, Joseph M. Massaro, and William B. Kannel. 2008. "General cardiovascular risk profile for use in primary care: The Framingham heart study." *Circulation* 117 (6): 743–53. <https://doi.org/10.1161/CIRCULATIONAHA.107.699579>.

Dekkers, Koen F, Maarten van Iterson, Roderick C Sliker, Matthijs H Moed, Marc Jan Bonder, Michiel van Galen, Hailiang Mei, et al. 2016. "Blood lipids influence DNA methylation in circulating cells." *Genome Biology* 17 (1): 138. <https://doi.org/10.1186/s13059-016-1000-6>.

Dempfle, Astrid, André Scherag, Rebecca Hein, Lars Beckmann, Jenny Chang-Claude, and Helmut Schäfer. 2008. "Gene-environment interactions for complex traits: Definitions, methodological requirements and challenges." <https://doi.org/10.1038/ejhg.2008.106>.

DeVos, T., Reimo Tetzner, Fabian Model, Gunter Weiss, Matthias Schuster, Jürgen Distler, Kathryn V. Steiger, et al. 2009. "Circulating Methylated SEPT9 DNA in Plasma Is a Biomarker for Colorectal Cancer." *Clinical Chemistry* 55 (7): 1337–46. <https://doi.org/10.1373/clinchem.2008.115808>.

Do, Elizabeth K., and Hermine H. Maes. 2016. "Genotype × Environment Interaction in Smoking Behaviors: A Systematic Review." *Nicotine & Tobacco Research*, June, ntw153. <https://doi.org/10.1093/ntr/ntw153>.

Dongen, Jenny van, Michel G Nivard, Gonneke Willemsen, Jouke-Jan Hottenga, Quinta Helmer, Conor V Dolan, Erik a Ehli, et al. 2016. "Genetic and environmental influences interact with age and sex in shaping the human methylome." *Nature Communications* 7 (1). Nature Publishing Group: 11115. <https://doi.org/10.1038/ncomms11115>.

Feinberg, Andrew P. 2018. "The Key Role of Epigenetics in Human Disease Prevention and Mitigation." *New England Journal of Medicine*. <https://doi.org/10.1056/NEJMr1402513>.

Feinberg, Andrew P, Rafael a Irizarry, Delphine Fradin, Martin J Aryee, Peter Murakami, Thor Aspelund, Gudny Eiriksdottir, et al. 2010. "Personalized Epigenomic Signatures That Are Stable Over Time and Covary with Body Mass Index." *Science Translational Medicine* 2 (49): 49ra67–49ra67. <https://doi.org/10.1126/scitranslmed.3001262>.

Fernández-Sanlés, Alba, Sergi Sayols-Baixeras, Isaac Subirana, Irene R Degano, and Roberto Elosua. 2017. "Association between DNA methylation and coronary heart disease or other atherosclerotic events: A systematic review." <https://doi.org/10.1016/j.atherosclerosis.2017.05.022>.

Figueiredo, Jane C., Li Hsu, Carolyn M. Hutter, Yi Lin, Peter T. Campbell, John A. Baron, Sonja I. Berndt, et al. 2014. "Genome-Wide Diet-Gene Interaction Analyses for Risk of Colorectal Cancer." *PLoS Genetics*. <https://doi.org/10.1371/journal.pgen.1004228>.

Fortin, Jean-Philippe, Aurélie Labbe, Mathieu Lemire, Brent W Zanke, Thomas J Hudson, Elana J Fertig, Celia Mt Greenwood, and Kasper D Hansen. 2014. "Functional normalization of 450k methylation array data improves replication in large cancer studies." *Genome Biology* 15 (12): 503. <https://doi.org/10.1186/s13059-014-0503-2>.

Goff, David C., Donald M. Lloyd-Jones, Glen Bennett, Sean Coady, Ralph B. D'Agostino, Raymond Gibbons, Philip Greenland, et al. 2014. "2013 ACC/AHA guideline on the assessment of cardiovascular risk: A report of the American college of cardiology/American heart association task force on practice guidelines." <https://doi.org/10.1016/j.jacc.2013.11.005>.

Gómez-Úriz, Ana M., Fermín I. Milagro, María L. Mansego, Paúl Cordero, Itziar Abete, Ana De Arce, Estíbaliz Goyenechea, et al. 2015. "Obesity and ischemic stroke modulate the methylation levels of KCNQ1 in white blood cells." *Human Molecular Genetics* 24 (5): 1432–40. <https://doi.org/10.1093/hmg/ddu559>.

Grishkevich, Vladislav, and Itai Yanai. 2013. "The genomic determinants of genotype × environment interactions in gene expression." <https://doi.org/10.1016/j.tig.2013.05.006>.

Hedman, Åsa K., Michael M. Mendelson, Riccardo E. Marioni, Stefan Gustafsson, Roby Joehanes, Marguerite R. Irvin, Degui Zhi, et al. 2017. "Epigenetic Patterns in Blood Associated With Lipid Traits Predict Incident Coronary Heart Disease Events and Are Enriched for Results From Genome-Wide Association Studies." *Circulation: Cardiovascular Genetics* 10 (1): e001487. <https://doi.org/10.1161/CIRCGENETICS.116.001487>.

Horvath, Steve, Michael Gurven, Morgan E Levine, Benjamin C Trumble, Hillard Kaplan, Hooman Allayee, Beate R Ritz, et al. 2016. "An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease." *Genome Biology* 17 (1): 171. <https://doi.org/10.1186/s13059-016-1030-0>.

Houseman, Eugene Andres, William P Accomando, Devin C Koestler, Brock C Christensen, Carmen J Marsit, Heather H Nelson, John K Wiencke, and Karl T Kelsey. 2012. "DNA methylation arrays as surrogate measures of cell mixture distribution." *BMC Bioinformatics* 13 (1): 86. <https://doi.org/10.1186/1471-2105-13-86>.

Houseman, Eugene Andres, John Molitor, and Carmen J. Marsit. 2014. "Reference-free cell mixture adjustments in analysis of DNA methylation data." *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btu029>.

Huang, Yen Tsung, Su Chu, Eric B Loucks, Chien Ling Lin, Charles B Eaton, Stephen L Buka, and Karl T Kelsey. 2016. "Epigenome-wide profiling of DNA methylation in paired samples of adipose tissue and blood." *Epigenetics* 11 (3): 227–36. <https://doi.org/10.1080/15592294.2016.1146853>.

Hunter, David J. 2005. "Gene–environment interactions in human diseases." *Nature Reviews Genetics* 6 (4): 287–98. <https://doi.org/10.1038/nrg1578>.

Irvin, Marguerite R., Degui Zhi, Roby Joehanes, Michael Mendelson, Stella Aslibekyan, Steven A. Claas, Krista S. Thibeault, et al. 2014. "Epigenome-wide association study of fasting blood lipids in the genetics of lipid-lowering drugs and diet network study." *Circulation* 130 (7): 565–72. <https://doi.org/10.1161/CIRCULATIONAHA.114.009158>.

- Jacob, R a, D M Gretz, P C Taylor, S J James, I P Pogribny, B J Miller, S M Henning, and M E Swendseid. 1998. "Moderate folate depletion increases plasma homocysteine and decreases lymphocyte DNA methylation in postmenopausal women." *The Journal of Nutrition* 128 (7): 1204–12.
- Johnson, W. Evan, Cheng Li, and Ariel Rabinovic. 2007. "Adjusting batch effects in microarray expression data using empirical Bayes methods." *Biostatistics* 8 (1). Oxford University Press: 118–27. <https://doi.org/10.1093/biostatistics/kxj037>.
- Jones, Peter A. 2012. "Functions of DNA methylation: islands, start sites, gene bodies and beyond." *Nature Reviews Genetics* 13 (7): 484–92. <https://doi.org/10.1038/nrg3230>.
- Kessler, Noah J., Robert A. Waterland, Andrew M. Prentice, and Matt J. Silver. 2018. "Establishment of environmentally sensitive DNA methylation states in the very early human embryo." *Science Advances* 4 (7). American Association for the Advancement of Science: eaat2624. <https://doi.org/10.1126/sciadv.aat2624>.
- Khera, Amit V., Mark Chaffin, Krishna G. Aragam, Mary E. Haas, Carolina Roselli, Seung Hoan Choi, Pradeep Natarajan, et al. 2018. "Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations." *Nature Genetics* 50 (9): 1219–24. <https://doi.org/10.1038/s41588-018-0183-z>.
- Kim, Myungjin, Tiffany I. Long, Kazuko Arakawa, Renwei Wang, Mimi C. Yu, and Peter W. Laird. 2010. "DNA Methylation as a Biomarker for Cardiovascular Disease Risk." *PLoS ONE* 5 (3): e9692. <https://doi.org/10.1371/journal.pone.0009692>.
- Kurdyukov, Sergey, and Martyn Bullock. 2016. "DNA Methylation Analysis: Choosing the Right Method." *Biology* 5 (1): 3. <https://doi.org/10.3390/biology5010003>.
- Lai, Chao-Qiang, Caren E. Smith, Laurence D. Parnell, Yu-Chi Lee, Dolores Corella, Paul Hopkins, Bertha A. Hidalgo, et al. 2018. "Epigenomics and metabolomics reveal the mechanism of the APOA2-saturated fat intake interaction affecting obesity." *The American Journal of Clinical Nutrition*, June. <https://doi.org/10.1093/ajcn/nqy081>.
- Lai, Chao-Qiang, Mary K. Wojczynski, Laurence D. Parnell, Bertha A. Hidalgo, Marguerite Ryan Irvin, Stella Aslibekyan, Michael A. Province, Devin M. Absher, Donna K. Arnett, and José M. Ordovás. 2016. "Epigenome-wide association study of triglyceride postprandial responses to a high-fat dietary challenge." *Journal of Lipid Research* 57 (12): 2200–2207. <https://doi.org/10.1194/jlr.M069948>.
- Lappalainen, Tuuli, and John M Greally. 2017. "Associating cellular epigenetic models with human phenotypes." *Nature Reviews Genetics*, May. <https://doi.org/10.1038/nrg.2017.32>.
- Laville, Vincent, Timothy Majarian, Yun J Sung, Mary F Feitosa, Daniel Chasman, Amy R Bentley, Charles N Rotimi, et al. 2019. "Genome-wide Interaction Studies by the CHARGE Gene-Lifestyle Interactions Working Group: what we have learned and what is coming next." *bioRxiv* (Preprint). <https://doi.org/10.1101/562157>.
- Leek, Jeffrey T., and John D. Storey. 2007. "Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis." *PLoS Genetics* 3 (9): e161. <https://doi.org/10.1371/journal.pgen.0030161>.

Li, Jun, Xiaoyan Zhu, Kuai Yu, Haijing Jiang, Yizhi Zhang, Siyun Deng, Longxian Cheng, et al. 2017. "Genome-Wide Analysis of DNA Methylation and Acute Coronary Syndrome." *Circulation Research* 120 (11): 1754–67. <https://doi.org/10.1161/CIRCRESAHA.116.310324>.

Maksimovic, Jovana, Lavinia Gordon, and Alicia Oshlack. 2012. "SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips." *Genome Biol* 13 (6): R44. <https://doi.org/10.1186/gb-2012-13-6-r44>.

Matosin, Natalie, Cristiana Cruceanu, and Elisabeth B Binder. 2017. "Preclinical and Clinical Evidence of DNA Methylation Changes in Response to Trauma and Chronic Stress." *Chronic Stress* 1 (January): 247054701771076. <https://doi.org/10.1177/2470547017710764>.

Micha, Renata, Jose L. Peñalvo, Frederick Cudhea, Fumiaki Imamura, Colin D. Rehm, and Dariush Mozaffarian. 2017. "Association Between Dietary Factors and Mortality From Heart Disease, Stroke, and Type 2 Diabetes in the United States." *JAMA* 317 (9): 912. <https://doi.org/10.1001/jama.2017.0947>.

Milagro, Fermin I, Javier Campión, Paúl Cordero, Estíbaliz Goyenechea, Ana M Gómez-Uriz, Itziar Abete, Maria a Zulet, and J Alfredo Martínez. 2011. "A dual epigenomic approach for the search of obesity biomarkers: DNA methylation in relation to diet-induced weight loss." *The FASEB Journal* 25 (4): 1378–89. <https://doi.org/10.1096/fj.10-170365>.

Morris, Tiffany J, and Stephan Beck. 2015. "Analysis pipelines and packages for Infinium HumanMethylation450 BeadChip (450k) data." *Methods* 72 (C): 3–8. <https://doi.org/10.1016/j.ymeth.2014.08.011>.

Murphy, M. O., D. M. Cohn, and A. S. Loria. 2017. "Developmental origins of cardiovascular disease: Impact of early life stress in humans and rodents." <https://doi.org/10.1016/j.neubiorev.2016.07.018>.

Ober, Carole, and Donata Vercelli. 2011. "Gene-environment interactions in human disease: Nuisance or opportunity?" <https://doi.org/10.1016/j.tig.2010.12.004>.

Olsen, Nanna J., Lars Ångquist, Sofus C. Larsen, Allan Linneberg, Tea Skaaby, Lise Lotte N. Husemoen, Ulla Toft, et al. 2016. "Interactions between genetic variants associated with adiposity traits and soft drinks in relation to longitudinal changes in body weight and waist circumference." *The American Journal of Clinical Nutrition* 104 (3): 816–26. <https://doi.org/10.3945/ajcn.115.122820>.

Parnell, Laurence D, Britt A Blokker, Hassan S Dashti, Paula-Dene Nesbeth, Brittany Elle Cooper, Yiyi Ma, Yu-Chi Lee, et al. 2014. "CardioGxE, a catalog of gene-environment interactions for cardiometabolic traits." *BioData Mining* 7 (1): 21. <https://doi.org/10.1186/1756-0381-7-21>.

Peng, Ping, Lu Wang, Xi Yang, Xiaoyan Huang, Yanna Ba, Xiaoliang Chen, Jian Guo, Jiangfang Lian, and Jianqing Zhou. 2014. "A Preliminary Study of the Relationship between Promoter Methylation of the ABCG1, GALNT2 and HMGCR Genes and Coronary Heart Disease." Edited by William B. Coleman. *PLoS ONE* 9 (8): e102265. <https://doi.org/10.1371/journal.pone.0102265>.

Perflyev, Alexander, Ingrid Dahlman, Linn Gillberg, Fredrik Rosqvist, David Iggman, Petr Volkov, Emma Nilsson, Ulf Risérus, and Charlotte Ling. 2017. "Impact of polyunsaturated and saturated fat overfeeding on the DNA-methylation pattern in human adipose tissue: a randomized controlled trial." *The American Journal of Clinical Nutrition* 105 (4). American Society for Nutrition: 991–1000. <https://doi.org/10.3945/ajcn.116.143164>.

- Perna, Laura, Yan Zhang, Ute Mons, Bernd Holleczech, Kai-Uwe Saum, and Hermann Brenner. 2016. "Epigenetic age acceleration predicts cancer, cardiovascular, and all-cause mortality in a German case cohort." *Clinical Epigenetics* 8 (1): 64. <https://doi.org/10.1186/s13148-016-0228-z>.
- Pfeiffer, Liliane, Simone Wahl, Luke C. Pilling, Eva Reischl, Johanna K. Sandling, Sonja Kunze, Lesca M. Holdt, et al. 2015. "DNA Methylation of Lipid-Related Genes Affects Blood Lipid Levels." *Circulation: Cardiovascular Genetics* 8 (2): 334–42. <https://doi.org/10.1161/CIRCGENETICS.114.000804>.
- Pidsley, Ruth, Elena Zotenko, Timothy J Peters, Mitchell G Lawrence, Gail P Risbridger, Peter Molloy, Susan Van Djik, et al. 2016. "Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling." *Genome Biology* 17 (1): 208. <https://doi.org/10.1186/s13059-016-1066-1>.
- Qi, Qibin, Audrey Y. Chu, Jae H. Kang, Majken K. Jensen, Gary C. Curhan, Louis R. Pasquale, Paul M. Ridker, et al. 2012. "Sugar-Sweetened Beverages and Genetic Risk of Obesity." *New England Journal of Medicine* 367 (15): 1387–96. <https://doi.org/10.1056/NEJMoa1203039>.
- Rahmani, Elior, Noah Zaitlen, Yael Baran, Celeste Eng, Donglei Hu, Joshua Galanter, Sam Oh, et al. 2016. "Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies." *Nature Methods*. <https://doi.org/10.1038/nmeth.3809>.
- Rakyan, Vardhman K, Thomas A Down, David J Balding, and Stephan Beck. 2012. "Epigenome-wide association studies for common human diseases." *Nature Reviews Genetics* 12 (8). Nature Publishing Group: 529–41. <https://doi.org/10.1038/nrg3000>.
- Rask-Andersen, Mathias, David Martinsson, Muhammad Ahsan, Stefan Enroth, Weronica E Ek, Ulf Gyllenstein, and Åsa Johansson. 2016. "Epigenome-wide association study reveals differential DNA methylation in individuals with a history of myocardial infarction." *Human Molecular Genetics* 25 (21): 4739–48. <https://doi.org/10.1093/hmg/ddw302>.
- Reinikainen, Jaakko, Tiina Laatikainen, Juha Karvanen, and Hanna Tolonen. 2015. "Lifetime cumulative risk factors predict cardiovascular disease mortality in a 50-year follow-up study in Finland." *International Journal of Epidemiology* 44 (1). Oxford University Press: 108–16. <https://doi.org/10.1093/ije/dyu235>.
- Richard, Melissa A., Tianxiao Huan, Symen Ligthart, Rahul Gondalia, Min A. Jhun, Jennifer A. Brody, Marguerite R. Irvin, et al. 2017. "DNA Methylation Analysis Identifies Loci for Blood Pressure Regulation." *American Journal of Human Genetics* 101 (6). Cell Press: 888–902. <https://doi.org/10.1016/j.ajhg.2017.09.028>.
- Rönn, Tina, Petr Volkov, Cajsa Davegårdh, Tasnim Dayeh, Elin Hall, Anders H. Olsson, Emma Nilsson, et al. 2013. "A Six Months Exercise Intervention Influences the Genome-wide DNA Methylation Pattern in Human Adipose Tissue." Edited by John M. Greally. *PLoS Genetics* 9 (6): e1003572. <https://doi.org/10.1371/journal.pgen.1003572>.
- Rudkowska, Iwona, Frédéric Guénard, Pierre Julien, Patrick Couture, Simone Lemieux, Olivier Barbier, Philip C. Calder, Anne Marie Miniñane, and Marie-Claude Vohl. 2014. "Genome-wide association study of the plasma triglyceride response to an n-3 polyunsaturated fatty acid supplementation." *Journal of Lipid Research*. <https://doi.org/10.1194/jlr.M045898>.

Schwartz, Stephen M., Hillel T. Schwartz, Steven Horvath, Eric Schadt, and Su In Lee. 2012. "A systematic approach to multifactorial cardiovascular disease: Causal analysis." *Arteriosclerosis, Thrombosis, and Vascular Biology*. <https://doi.org/10.1161/ATVBAHA.112.300123>.

Sharma, Priyanka, Gaurav Garg, Arun Kumar, Farhan Mohammad, Sudha Ramesh Kumar, Vinay Singh Tanwar, Satish Sati, et al. 2014. "Genome wide DNA methylation profiling for epigenetic alteration in coronary artery disease patients." *Gene* 541 (1): 31–40. <https://doi.org/10.1016/j.gene.2014.02.034>.

Shukla, Ankita, Manika Sehgal, and Tiratha Raj Singh. 2015. "Hydroxymethylation and its potential implication in DNA repair system: A review and future perspectives." <https://doi.org/10.1016/j.gene.2015.03.075>.

Smith, Caren E., Jack L. Follis, Hassan S. Dashti, Toshiko Tanaka, Mariaelisa Graff, Amanda M. Fretts, Tuomas O. Kilpeläinen, et al. 2018. "Genome-Wide Interactions with Dairy Intake for Body Mass Index in Adults of European Descent." *Molecular Nutrition and Food Research*. <https://doi.org/10.1002/mnfr.201700347>.

Suchiman, H Eka D, Roderick C Sliker, Dennis Kremer, P Eline Slagboom, Bastiaan T Heijmans, and Elmar W Tobi. 2015. "Design, measurement and processing of region-specific DNA methylation assays: the mass spectrometry-based method EpiTYPER." *Frontiers in Genetics* 6 (SEP). Frontiers Media SA: 287. <https://doi.org/10.3389/fgene.2015.00287>.

Talens, Rudolf P, Dorret I Boomsma, Elmar W Tobi, Dennis Kremer, J. Wouter Jukema, Gonneke Willemsen, Hein Putter, P. Eline Slagboom, and Bastiaan T Heijmans. 2010. "Variation, patterns, and temporal stability of DNA methylation: considerations for epigenetic epidemiology." *The FASEB Journal* 24 (9): 3135–44. <https://doi.org/10.1096/fj.09-150490>.

Tall, Alan R., and Laurent Yvan-Charvet. 2015. "Cholesterol, inflammation and innate immunity." <https://doi.org/10.1038/nri3793>.

Teschendorff, Andrew E, Francesco Marabita, Matthias Lechner, Thomas Bartlett, Jesper Tegner, David Gomez-Cabrero, and Stephan Beck. 2013. "A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data." *Bioinformatics* 29 (2): 189–96. <https://doi.org/10.1093/bioinformatics/bts680>.

Triche, Timothy J., Daniel J. Weisenberger, David Van Den Berg, Peter W. Laird, and Kimberly D. Siegmund. 2013. "Low-level processing of Illumina Infinium DNA Methylation BeadArrays." *Nucleic Acids Research* 41 (7): e90–e90. <https://doi.org/10.1093/nar/gkt090>.

Wahl, Simone, Alexander Drong, Benjamin Lehne, Marie Loh, William R. Scott, Sonja Kunze, Pei-Chien Tsai, et al. 2017. "Epigenome-wide association study of body mass index and the adverse outcomes of adiposity." *Nature* 541 (7635): 81–86. <https://doi.org/10.1038/nature20784>.

Winkler, Thomas W., Anne E. Justice, L. Adrienne Cupples, Florian Kronenberg, Zoltán Kutalik, and Iris M. Heid. 2017. "Approaches to detect genetic effects that differ between two strata in genome-wide meta-analyses: Recommendations based on a systematic evaluation." Edited by David Meyre. *PLOS ONE* 12 (7): e0181038. <https://doi.org/10.1371/journal.pone.0181038>.

Wu, Susan C., and Yi Zhang. 2010. "Active DNA demethylation: many roads lead to Rome." *Nature Reviews Molecular Cell Biology* 11 (9). Nature Publishing Group: 607–20. <https://doi.org/10.1038/nrm2950>.

Zaina, Silvio, Holger Heyn, F. Javier Carmona, Nuray Varol, Sergi Sayols, Enric Condom, José Ramírez-Ruz, et al. 2014. "DNA methylation map of human atherosclerosis." *Circulation: Cardiovascular Genetics*. <https://doi.org/10.1161/CIRCGENETICS.113.000441>.

Zhang, Yan, Rory Wilson, Jonathan Heiss, Lutz P. Breitling, Kai-Uwe Saum, Ben Schöttker, Bernd Holleczek, Melanie Waldenberger, Annette Peters, and Hermann Brenner. 2017. "DNA methylation signatures in peripheral blood strongly predict all-cause mortality." *Nature Communications* 8 (March): 14617. <https://doi.org/10.1038/ncomms14617>.

Zhong, Jia, Golareh Agha, and Andrea A. Baccarelli. 2016. "The Role of DNA Methylation in Cardiovascular Risk and Disease: Methodological Aspects, Study Design, and Data Analysis for Epidemiological Studies." <https://doi.org/10.1161/CIRCRESAHA.115.305206>.

Zhuang, Joanna J, Martin Widschwendter, and Andrew E. Teschendorff. 2012. "A comparison of feature selection and classification methods in DNA methylation studies using the Illumina Infinium platform." *BMC Bioinformatics* 13 (1): 59. <https://doi.org/10.1186/1471-2105-13-59>.

Zou, Hui, and Trevor Hastie. 2005. "Regularization and variable selection via the elastic net." *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 67 (2): 301–20. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>.

Chapter 2

Epigenomics of CVD

2.1 DNA methylation modules associate with incident cardiovascular disease and cumulative risk factor exposure

Accepted for publication in Clinical Epigenetics

2.1.1 Abstract

Background: Epigenome-wide association studies using DNA methylation have the potential to uncover novel biomarkers and mechanisms of cardiovascular disease (CVD) risk. However, the direction of causation for these associations is not always clear, and investigations to-date have often failed to replicate at the level of individual loci.

Methods: Here, we undertook module- and region-based DNA methylation analyses of incident CVD in the Women's Health Initiative (WHI) and Framingham Heart Study Offspring Cohort (FHS) in order to find more robust epigenetic biomarkers for cardiovascular risk. We applied weighted gene correlation network analysis (WGCNA) and the Comb-p algorithm to find methylation modules and regions associated with incident CVD in the WHI dataset.

Results: We discovered two modules whose activation correlated with CVD risk and replicated across cohorts. One of these modules was enriched for development-related processes and overlaps strongly with epigenetic aging sites. For the other, we showed preliminary evidence for monocyte-specific effects and statistical links to cumulative exposure to traditional cardiovascular risk factors. Additionally, we found three regions (associated with the genes *SLC9A1*, *SLC1A5*, and *TNRC6C*) whose methylation associates with CVD risk.

Conclusions: In sum, we present several epigenetic associations with incident CVD that reveal disease mechanisms related to development and monocyte biology. Furthermore, we show that epigenetic modules may act as a molecular readout of cumulative cardiovascular risk factor exposure, with implications for the improvement of clinical risk prediction.

2.1.2 Background

Genetic approaches to cardiovascular disease (CVD) research have led to important breakthroughs in mechanistic understanding and therapeutic strategies. However, the mechanisms for gene variant-disease relationships are often difficult to determine, and their effects may often be mediated by epigenetic regulation (Bonder et al. 2016). DNA methylation is one such mechanism that can reflect both genetic variation and environmental exposures and potentially drive their effects on CVD outcomes (Ordovás and Smith 2010).

A series of recent epigenome-wide association studies (EWAS) have examined relationships between DNA methylation at cytosine-phosphate-guanine (CpG) sites and various subtypes of CVD, including prior myocardial infarction (MI) (Rask-Andersen et al. 2016), acute coronary syndrome (Li et al. 2017), and atherosclerosis (Nakatochi et al.

2017). These cross-sectional studies may reveal important mechanistic insights, but are susceptible to reverse causation, i.e. methylation being influenced by the presence of CVD. Indeed, Mendelian randomization approaches across multiple phenotypes have suggested that reverse causation is more common (Dekkers et al. 2016; Wahl et al. 2017) than the causal methylation effect that is often implicitly assumed. One approach to this problem is to examine epigenetic associations with cardiovascular risk factors. Multiple investigations have explored these relationships genome-wide (Pfeiffer et al. 2015; Irvin et al. 2014), and have even uncovered prognostic CpG sites for incident coronary heart disease (CHD) in the process (Hedman et al. 2017; Aslibekyan et al. 2018). A few studies looking directly at incident CVD as a binary variable have found relationships with global DNA methylation (as approximated by LINE-1 methylation levels) and with a specific cluster of CpG sites in the ZBTB12 gene (Baccarelli et al. 2010; Guarrera et al. 2015).

Studies linking CVD and methylation have additionally shown a notable lack of replication, especially at the level of single CpG sites (Fernández-Sanlés et al. 2017). One approach to this problem is to aggregate CpGs and test their phenotype associations at the group level. Differentially methylated region (DMR) searches may improve detection by combining sites based on physical proximity on the genome (Jaffe et al. 2012; Pedersen et al. 2012). An alternative grouping strategy is to search for correlation-based clusters, which may boost biological signal and improve the interpretability of results (Zhang and Horvath 2005). This approach was originally developed for use with gene expression data, but has been successfully applied to higher-dimensional DNA methylation microarray datasets (Horvath et al. 2012; Busch et al. 2016).

To address the problem of reverse causation by CVD while achieving more robust results, we set out to analyze relationships between group-level CpG methylation and incident CVD using time-to-event models in two cohorts. We used module- and region-based techniques to improve detection and provide more interpretable results. We sought context for two specific modules of interest using gene- and chromatin-based annotations, and compared module activations to past and current cardiovascular risk factor levels to better understand their potential biological mechanisms.

2.1.3 Results

Weighted correlation network approach finds CVD-related modules

Population characteristics are described in Table 1. The discovery set, Women’s Health Initiative (n=2023), had a median age of 65 at blood draw and is entirely female, while being selected for an approximately equal ratio of subjects who did and did not experience an incident CVD event following the methylation measurement timepoint. The replication set, Framingham Heart Study Offspring Cohort (n=2587), had a median age of 66 at blood draw (Exam 8) and is approximately half female, with 305 subjects experiencing incident CVD events. Cardiovascular events were defined here as encompassing CHD, stroke, and death from CVD (see Methods section for further details).

Table 1: Population description

	FHS	WHI
Sample Size	2587	2023
% female	55 %	100 %
Age at blood draw	66 (60-73)	65 (59-70)
Mixed ancestry	No	Yes
Body mass index (BMI)	27.7 (24.5-31)	29.1 (25.5-33.3)
% smoke currently	9%	10%
Smoking pack-years	0 (0-0)	0 (0-12.5)
# prior CVD events	331	0
# incident CVD events	305	1009
Total cholesterol (Chol; mg/dL)	185 (161-211)	230 (206-259)
LDL-cholesterol (LDL; mg/dL)	104 (84-126)	150 (126-175)
HDL-cholesterol (HDL; mg/dL)	55 (44-68)	51 (43-60)
Triglycerides (TG; mg/dL)	102 (74-142)	127 (92-177)
Fasting glucose (Glu; mg/dL)	101 (94-110)	96 (88.6-108)
High-sensitivity C-reactive protein (hsCRP; mg/dL)	1.5 (0.8-3.2)	3.1 (1.4-6.5)
Systolic blood pressure (SBP; mmHg)	127 (116-139)	131 (120-143)

* Continuous values shown as: median (interquartile range).

* 92 subjects experienced both prior and incident CVD.

We first set out to find biologically relevant modules in an unsupervised manner (agnostic to incident CVD information) using the WGCNA algorithm for 422952 CpGs in WHI passing quality control filters (study overview in Supp. Fig. S1). After weighted correlation network construction, topological overlap calculation, and subsequent clustering, 110 modules were uncovered, ranging in size from 28 to 35361 CpGs. These modules were assigned unique color labels as identifiers.

Principal component eigenvectors for each module were calculated in order to examine the characteristics of these modules as a whole. The first principal component of each module tended to explain approximately half of the total variance, while the rest contributed only small fractions (see Supp. Fig. S2 for selected Scree plots). Thus, these first eigenvectors, or “eigenCpGs”, were subsequently used to describe module behavior. Cox proportional hazards models were used to assess the relationships between these module eigenCpGs and incident CVD. In partially-adjusted models (adjusted for technical factors and estimated white blood cell proportions), three modules were found to be associated at multiple test-corrected false discovery rate (FDR) < 0.2 (Table 2; correction based on 110 modules). Adjustment for biological covariates (age, BMI, sex/race, and smoking behavior) attenuated these relationships to marginal statistical significance (all $0.01 < p < 0.1$; direct risk factor associations shown in Fig. 3). These modules showed strong (FDR < 10) enrichment for different sets of GO terms, ranging from immune activation (myeloid or T cell) to developmental processes.

Table 2: Modules associated with incident CVD at $FDR < 0.2$

Module	Size	EigenCpG		Enrichment Analysis		
		Var. expl. (%)	p	GO terms	CpG Islands	Gene-based
Blue	29441	44.61	0.00027	Development	N_shore	1stExon/TSS/ 5' UTR
Brown	953	53.08	0.00455	Immune activation	Open sea	
Purple	568	44.88	0.00500	T cell activation	Open sea	Body

All three modules showed very strong preservation in FHS (all $Z_{summary}$ statistics > 50 , where 10 is a typical threshold for strong preservation), when evaluated using established density and connectivity preservation techniques (Langfelder et al. 2011). Of these, two associations with incident CVD (blue and brown) replicated strongly in FHS, while purple showed nominal replication ($p=0.0203$) in partially-adjusted models (Supp. Table 1). Fully-adjusted models including age as a covariate attenuated (brown) or abolished (blue and purple) these associations in FHS.

Though the existence of past CVD events (experienced prior to sample collection for DNA methylation measurement) could represent a confounder in the FHS dataset, sensitivity analyses adjusting for past events did not appreciably reduce the strength of these module-trait relationships. Also of potential relevance to this replication is the demographic heterogeneity between the two cohorts. To address this possibility, we performed additional analyses including interaction terms between eigenCpGs for each module and either sex (in FHS) or race (in WHI). None of these analyses produced significant interaction terms at $p < 0.05$.

Genome-wide associations between DNA methylation and incident CVD events

To investigate more specific DNA methylation signals, we performed an epigenome-wide association study (EWAS) for incident CVD. Of single sites from the EWAS, 3 reached a genome-wide Bonferroni threshold, but none replicated strongly in FHS (Supp. Table 2). In order to improve statistical power, we focused on differentially methylated regions (DMRs) with respect to incident CVD status. Single-site EWAS p-values were used as input to the Comb-p algorithm, which seeks regions enriched for low p-values while accounting for autocorrelation based on genomic distance. Comb-p was applied separately to EWAS results from WHI and FHS.

206 DMRs were found in WHI after Sidak multiple testing correction for each DMR based on its length. Of these, 3 were both found in FHS and replicated at a Bonferroni level (Table 3; Fig. 1). These regions were annotated to two cellular transport genes (*SLC9A1* and *SLC1A5*) and *TNRC6C*, which codes for a scaffolding protein involved in miRNA-mediated translational repression. Of the three WGCNA modules identified above, brown CpG sites constituted part of 2 DMRs (at *SLC9A1* & *SLC1A5*), while a single CpG from the blue module was also a member of the *SLC9A1* DMR.

Table 3: Comb-*p* regions with multiple test-corrected $p < 0.05$ in WHI and Bonferroni $p < 0.05$ in FHS

Location	# CpGs	Annotated gene	Genomic region	Discovery		Replication
				P	Adj. P	P
chr1:27440462-721	3	SLC9A1	Body	1.75e-08	2.85e-05	1.03e-04
chr19:47287777-8263	6	SLC1A5	CpG shelf near TSS	5.91e-07	5.14e-04	3.80e-10
chr17:76037034-562	6	TNRC6C	CpG island in 5' UTR	1.67e-05	1.33e-02	1.89e-04

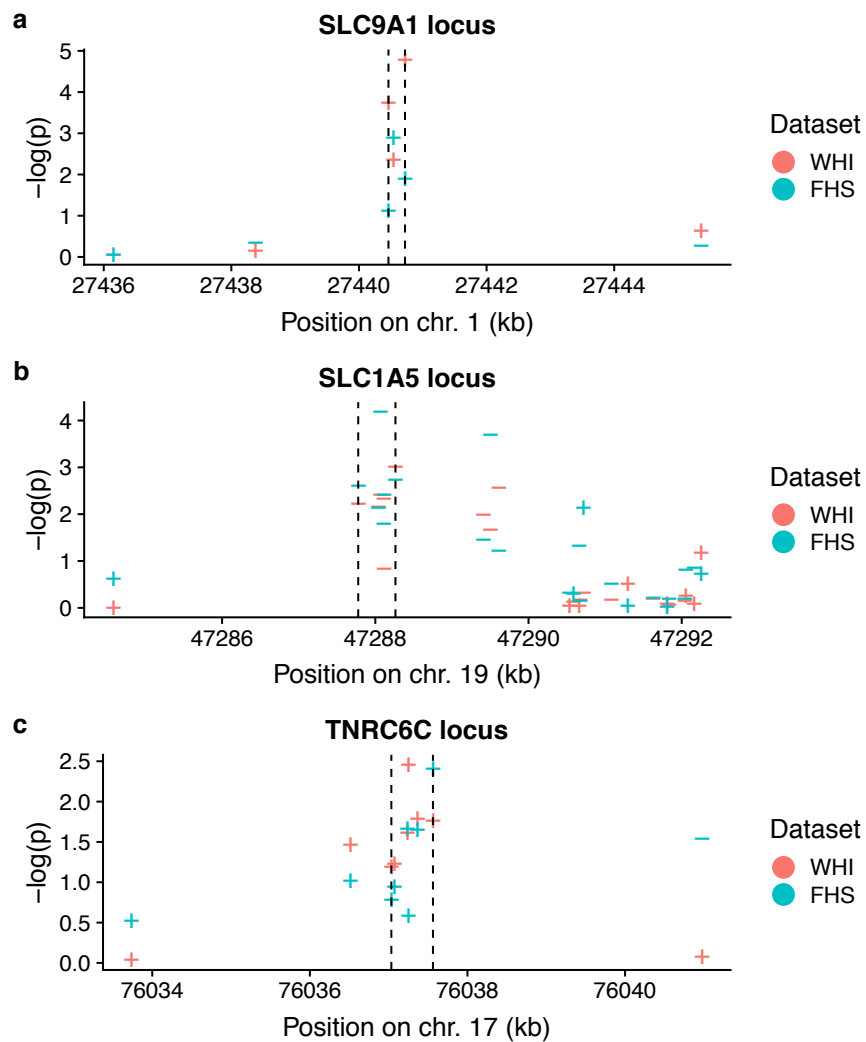


Figure 1: DMRs identified by Comb-*p* in WHI and validated in FHS at the (a) *SLC9A1*, (b) *SLC1A5*, and (c) *TNRC6C* loci. Negative logarithms of EWAS *p*-values are shown as a function of the genomic coordinate. EWAS *p*-values from WHI are in red and FHS in green. Dotted lines demarcate the DMR boundaries. Shapes for each point indicate the direction of association of the CpG site methylation with incident CVD.

Of CpGs in these 3 identified DMRs, 4 (2 in *SLC1A5* and 1 each in *SLC9A1* and *TNRC6C*) were involved in methylation-quantitative trait loci (mQTL) based on the mQTLdb (Gaunt et al. 2016). For these 4 CpGs, a Mendelian randomization analysis was undertaken using the MR-Base platform (Hemani et al. 2018) to assess the potential causality of their methylation for myocardial infarction (MI) and coronary heart disease CHD). While no associations were found for 3 of the 4 CpGs, a modest single-SNP relationship linked a cis-mQTL for cg22304262 with MI ($p=0.013$) and CHD ($p=0.072$). This SNP, rs8105903, was also found to modify *SLC1A5* expression levels in whole blood based on the GTEx database (Aguet et al. 2017). Taken together, these data are consistent with a small effect of DNA methylation at *SLC1A5* on CVD, possibly acting through regulation of *SLC1A5* gene expression.

Exploration of the brown and blue modules

Based on the results from the module- and region-centric analyses, we investigated the brown and blue modules further for biological significance. The brown module was associated with immune-related genes as noted above, and was enriched strongly for “open sea” sites ($p=1.1e-42$) and annotated enhancers ($p=1.7e-33$). In contrast, the blue module was associated with development-related genes, and was enriched moderately for sites near genic transcription start sites and strongly for CpG islands ($p < 2.2e-16$) (Fig. 2a,b).

Given these observations, we examined relative enrichments of enhancer- and promoter-associated histone marks across different blood cell subtypes to better understand the cell type specificity of this signal. Epigenetic peaks were annotated using data from the Roadmap Epigenomics Project (Kundaje et al. 2015) and relative enrichments were calculated as the fraction of module CpGs found in peaks divided by the fraction of all CpGs found in peaks (see Methods section for details).

We observed the greatest enrichment of brown CpGs in 2 enhancer-associated chromatin annotations, DNase hypersensitivity sites (DHS) and H3K4me1 histone peaks, from monocytes compared to other blood cell subtypes (Fig. 2c). This could point towards monocyte-related biology and inflammatory processes as an important shared mechanism for cardiovascular risk between the two cohorts examined here. To validate this observation, an additional analysis based on the CellDMC method (Zheng et al. 2018) was undertaken, based on the idea that cell type-specific epigenetic changes will result in statistical interactions between cell types and epigenetic quantities. Adapting this method to the epigenetic module level and applying it using partially-adjusted Cox models in FHS, no meaningful statistical interaction between monocyte fractions and brown module activation were observed (positive interaction; $p=0.83$). In contrast to the monocyte-related enrichments for the brown module, blue CpGs were enriched for DHS and promoter-associated H3K4me3 histone peaks from hematopoietic stem cells (HSCs), providing a link to the observed enrichment of development-related genes in this set.

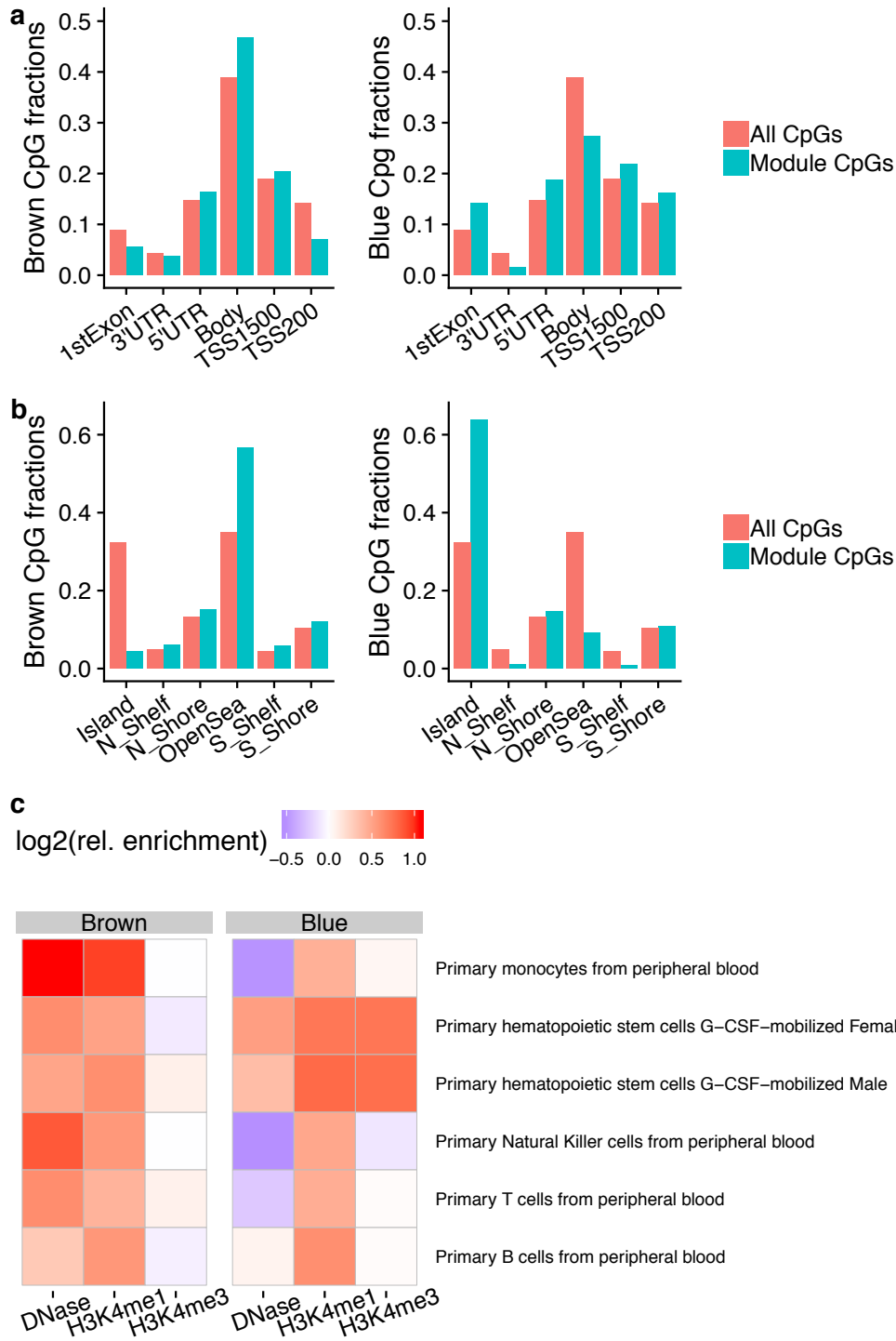


Figure 2: Genomic and epigenomic annotations of the brown and blue modules. *a,b*) Relative proportions of module CpGs compared to the full set of CpGs tested, with respect to gene-based (*a*) or CpG island-based (*b*) annotations (UTR = untranslated region; TSS_X = sites within X base pairs upstream of the gene transcription start site). *c*) Cell type-specific enrichments based on Roadmap Epigenomics datasets. Shown are relative enrichments of peaks (ratio of in-module fraction to all-CpG fraction) for a given epigenetic mark across many blood cell types, for each of the modules of interest.

The module CpG sets were also compared to two existing methylation-based age predictors from Horvath and Hannum et al., as well as the recent morbidity-directed phenoAge (Horvath 2013; Hannum et al. 2013; Levine et al. 2018). While enrichments for brown CpGs were moderate to nonexistent, blue CpGs were strongly enriched for all three of these sets, most highly for the original DNAm age developed by Horvath (46/353; $p=3.4e-5$; hypergeometric test), despite the fact that this model was developed based on only ~21,000 CpGs shared between multiple versions of the Illumina methylation microarray platform. Furthermore, 28 of these 46 CpGs had associated positive coefficients in the DNAm age predictor. This subset has been previously observed to contain a disproportionate amount of Polycomb-group target genes, which are known to associate with developmental processes and to be generally hypermethylated with age (Dozmorov 2015). Using SUZ12 binding regions (Lee et al. 2006) as a proxy for Polycomb-group targets, we confirmed their enrichment in the blue module ($p = 1.37e-7$). Surprisingly, the blue eigenCpG showed only a modest correlation with age itself ($r=0.09$).

Module-risk factor relationships

Next, we examined correlations between these module eigenCpGs and traditional cardiovascular risk factors. Though no extremely strong module-risk factor correlations were observed (all $|r|<0.25$), they tended to be stronger for the brown module, especially in FHS (Fig. 3a). Age showed the greatest association, while lipid and glycemic parameters also showed moderate associations. To further probe relationships between the brown module and risk factors in FHS, we retrieved historical risk factors measured in previous Offspring Cohort exams. Visual inspection revealed a notably stronger correlation between the module eigenCpG and cumulative (mean of all previous exams) compared to current risk factor exposure. This pattern applied for systolic blood pressure (strongly), triglycerides, glucose, BMI, and LDL (which correlated in the “expected” direction cumulatively, but non-intuitively at Exam 8) (Fig. 3b).

To better investigate this phenomenon, we tested associations between the brown module and each of the cumulative risk factors after adjustment for potential confounders. Specifically, for each risk factor, linear models were used to predict the brown eigenCpG value from either the current or cumulative risk factor level while adjusting for the full set of EWAS covariates other than BMI (age/sex/smoking/cell counts/study center/7 ctrl-probe PCs). Only for the brown module did cumulative risk factor exposure show strong associations, which were generally equal to or stronger than those of the current risk factors, most notably for BMI, hsCRP, and triglycerides (Table 4). Though more recent medication use could possibly explain discrepancies between biological relationships with current and past risk factors, adjustment for hypertension and lipid medication use did not notably affect the results of these models.

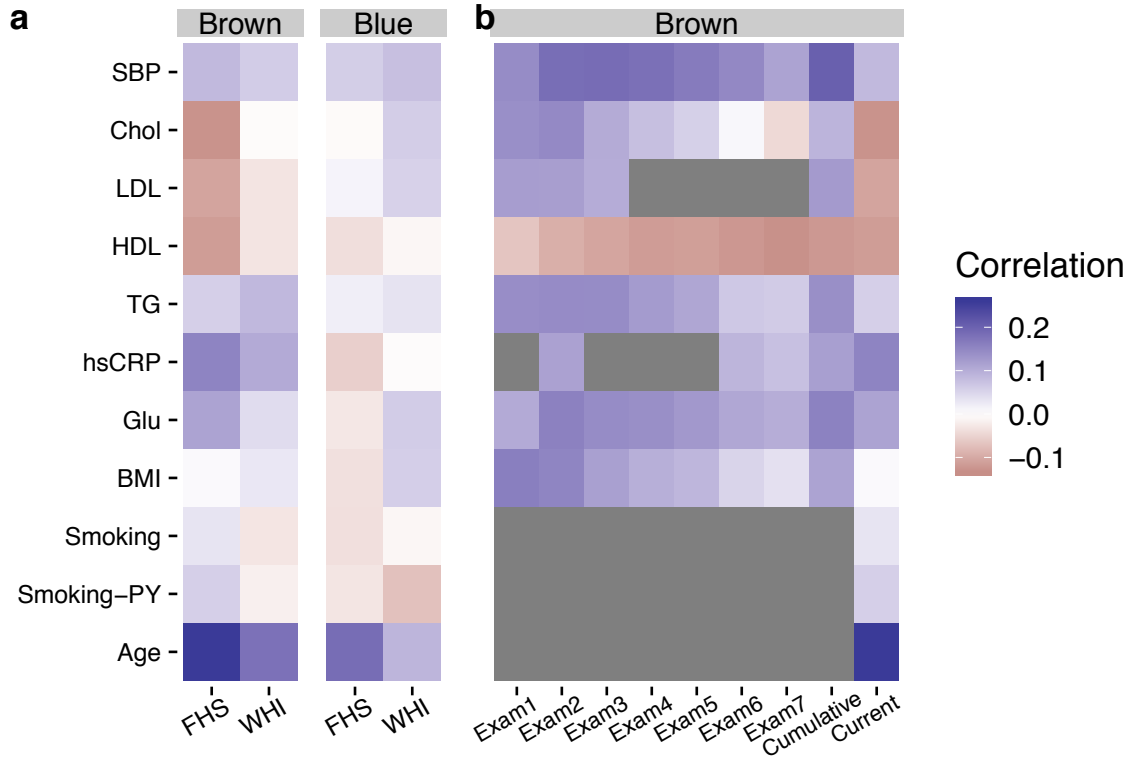


Figure 3: Risk factor-module relationships. (a) Pearson correlations between a series of traditional cardiovascular risk factors and module eigenCpGs (blue and brown) are shown in each study population. (b) Pearson correlations between historical risk factor levels in FHS (across previous exams, x-axis) and current brown module activation are shown. Grey panels indicate that the risk factor in question was not available for the corresponding exam (LDL and hsCRP) or was not analyzed with respect to past exams (smoking and age).

Table 4: Module-risk factor relationships (current and cumulative) after adjustment for covariates

Risk factor	Brown		Blue	
	Current	Cumulative	Current	Cumulative
BMI	0.036 (4.3e-06)	0.051 (2.3e-10)	0.026 (0.0061)	0.019 (0.051)
Glu	0.021 (0.011)	0.027 (0.0011)	-0.00065 (0.95)	-0.0045 (0.66)
hsCRP	0.021 (0.0077)	0.039 (1e-06)	0.025 (0.011)	0.014 (0.15)
TG	0.018 (0.02)	0.042 (5e-07)	0.021 (0.03)	0.015 (0.14)
HDL	-0.021 (0.015)	-0.017 (0.056)	-0.038 (0.00031)	-0.03 (0.0067)
LDL	-0.012 (0.13)	0.0089 (0.29)	-0.00072 (0.94)	0.027 (0.0088)
Chol	-0.014 (0.11)	0.019 (0.021)	-0.013 (0.2)	0.012 (0.24)
SBP	0.012 (0.15)	0.024 (0.0084)	-0.011 (0.27)	-0.0015 (0.89)

* Regression results are presented as: beta (p-value).

* Models are adjusted for age, sex, smoking status and pack-years, estimated cell counts, study center, and 7 control probe principal components.

Finally, we used the basic mediation approach of Baron and Kenny (Baron and Kenny 1986) to test whether brown module activation may mediate a portion of the effects of cumulative risk factor exposure on cardiovascular risk. A series of Cox models were created in FHS for these three most strongly-associated risk factors (BMI, hsCRP, and triglycerides). Covariates in all models included current values for the risk factor in question, as well as technical factors, estimated cell counts, age, and sex. Current risk factors did not show notable relationships with incident CVD in any of the models. Having established the exposure-mediator relationships (Table 4), we tested the association with CVD risk of 1) cumulative risk factors, 2) module eigenCpGs, and 3) both quantities together (Table 5; example causal diagram using hsCRP in Supp. Fig. S4). In general, the significance of the module relationships with CVD tended to decrease in the presence of cumulative risk factor values. This fits with a model in which, rather than mediating cardiovascular risk, module activation acts as a biomarker for the actions of cumulative risk factor exposures by some other mechanism. As only subjects with current risk factor values were included in each model, sample sizes were largely identical across models.

Table 5: *CVD risk models using cumulative risk factor exposure and brown module activation*

Risk factor	Risk factors only	Brown only	Full Model	
	Cumulative	Module	Cumulative	Module
BMI	0.061 (0.009)	0.012 (0.1)	0.057 (0.015)	0.01 (0.16)
hsCRP	0.64 (<0.001)	0.014 (0.054)	0.62 (<0.001)	0.012 (0.12)
TG	1.7 (<0.001)	0.016 (0.039)	1.7 (<0.001)	0.012 (0.1)

* Regression results are presented as: beta (p-value).

* Models are adjusted for age, sex, smoking status and pack-years, estimated cell counts, study center, and 7 control probe principal components.

2.1.4 Discussion

Here, we performed a primarily module-based epigenetic analysis of incident cardiovascular events in order to find robust, prospective biomarkers and uncover novel mechanisms contributing to disease risk. We began by constructing correlation-based clusters in the methylation data from WHI using the WGCNA algorithm. This network-based feature clustering approach can potentially improve the signal-to-noise ratio of high-dimensional DNA methylation data while facilitating more clear biological interpretation of results (Langfelder and Horvath 2008). As WGCNA does not consider class labels (i.e. incident CVD status), the 110 modules uncovered were not *a priori* expected to be related to CVD and rather reflected unbiased patterns in the data. After correction for multiple testing, the first principal components (eigenCpGs) of three of these modules were found to be related to incident cardiovascular events. A gene ontology-based enrichment analysis of the genes annotated to these modules found strong enrichment for either immune-related or development-related processes. The finding of immune-related processes is intuitive given that DNA from blood measures primarily immune cells, while the development-related enrichment could possibly reflect influences during early life (Hao et al. 2018). Notably, these two module “types” (immune and development) have been

uncovered in a prior network-based DNA methylation analysis related to asthma (Busch et al. 2016), suggesting that similar module types are a potentially general feature of blood-based methylation patterns and that these patterns may not be fully cardiovascular-specific, reflecting instead a predisposition toward general inflammatory disease processes. Both in WHI and in replication in FHS, two modules (blue and brown) showed strong relationships with incident CVD that were attenuated after adjustment for age (direct correlations of these modules with age are presented in Fig. 3).

We examined the set of module eigenvector loadings as a proxy for the relative importance of their component CpGs, in a similar approach to the standard calculation of gene-module correlations (or “kME” statistics) in WGCNA analyses. As we did not observe any obvious peaks distinguishing particularly important groups of CpGs, we undertook an epigenome-wide association study (EWAS) in order to identify potentially stronger locus-specific signals. Though we did not find any single sites replicating in FHS after stringent correction for multiple tests, a subsequent region-based analysis using the Comb-p algorithm revealed three regions replicating strongly across the two cohorts examined here. One was found on chromosome 1 in the body of the *SLC9A1* (also known as *NHE-1*) gene, which codes for an integral membrane ion transporter involved in intracellular pH maintenance. *SLC9A1* has been shown to be required for the increased adhesion, migration, and phagocytosis of oxidized LDL seen in monocytes in response to stimuli including leptin, adrenaline, and hyperglycemia (Sarigianni 2010). Another region discovered was on chromosome 19 near the transcription start site (TSS) of *SLC1A5*, which codes for a neutral amino acid transporter. Though strong evidence does not yet exist linking *SLC1A5* to cardiovascular mechanisms, its CpGs have shown associations with diabetes, blood pressure, and mortality (Walaszczyk et al. 2018; Richard et al. 2017; Zhang et al. 2017), and we note that its companion amino acid transporter, *SLC7A5*, is known to regulate metabolic and inflammatory reprogramming of monocytes in response to stimulation by lipopolysaccharide (LPS). Notably, CpG sites in both *SLC9A1* and *SLC1A5* were discovered and replicated in a recent EWAS for BMI (including the FHS cohort) (Mendelson et al. 2017), though the specific *SLC9A1* site from that study was not one of the three constituent CpGs in the region found here. These two SLC transporter DMRs contained CpGs belonging to blue (1 in *SLC9A1*) and brown (1 in *SLC9A1*, 5 in *SLC1A5*) modules. The third region was found near the TSS of *TNRC6C* on chromosome 17. This gene codes for a component of the miRNA-mediated translational repression cascade, has shown up in a genome-wide association study (GWAS) for heart failure (not one of the phenotypes included in our CVD definition here) (Villard et al. 2011), and was identified as a potential target gene in the monocyte-to-macrophage transition upon exposure to CSF-1 (Wallner et al. 2016). Common to these three DMRs is a potential involvement in monocyte biology specific to a stimulus response. This concept of “priming” for subsequent response to stimulus has been observed with respect to both monocyte activity in CVD (Short et al. 2017) and DNA methylation in general (Lam et al. 2012). While a two-step Mendelian randomization analysis here found modest potential causal evidence for CVD for only one DMR constituent CpG (in *SLC1A5*), a lack of available mQTL SNPs for the full set of relevant CpGs prevented a full exploration.

Based on the module- and region-level replication in FHS, we further explored the characteristics of the brown and blue modules. Enrichment analyses of gene-based and

locus-based annotations demonstrated that these two modules occupy distinct biological niches. Broadly, the brown module (consisting of about 1000 CpG sites) is enriched for enhancers and other non-proximal regions near immune-related genes, while the blue module (a notably large module of almost 30,000 CpG sites) is enriched for CpG islands near the TSS of development-related genes. One could speculate that these modules also represent different mechanisms of cardiovascular risk: one related to inflammatory burden and the other to long-term effects of early-life exposures, both of which are well-established as contributing to cardiovascular risk (Nahrendorf 2018; Hao et al. 2018). Analyses based on cross-tissue epigenome annotations added an additional dimension to these insights by suggesting differential importance of blood cell sub-types for these modules. A cell type specificity analysis, adapted from the eFORGE algorithm (Breeze et al. 2016), revealed the enrichment of monocyte-specific regions of open chromatin (DNase hypersensitivity sites and H3K4me1 peaks) in the brown module. This observation reinforces the idea of monocyte-specific activity suggested by the replicated DMRs as well as that of “monocyte priming” (Short et al. 2017). Based on the tendency of blue module CpGs to be proximal to gene TSS, we focused on enrichment for a promoter-associated marker, H3K4me3, and found a distinct signal related to hematopoietic stem cells. This finding supports a potential mechanism linking early-life exposure to consequences in adult life (Farlik et al. 2016; Laiosa and Tate 2015). We also observed that the blue module was strongly enriched for components of a popular epigenetic age marker (Horvath 2013) as well as for binding regions of the Polycomb-group member SUZ12. As Polycomb-group targets are known to be related to developmental processes (Dozmorov 2015), this finding contributes additional support to the module’s role as a bridge between development, aging, and disease risk.

It is not clear whether these methylation modules associate with cardiovascular risk upstream, downstream, or independently of traditional cardiovascular risk factors (including age, blood pressure, BMI, smoking, and lipid levels). To explore these relationships, we began by calculating correlations between risk factor levels and blue and brown module activations. Blue correlations were largely weak, while brown correlations were somewhat stronger, following the hypothesis that the blue module is more relevant to early-life, rather than adult, exposures as compared to brown. However, as a semi-stable biological quantity, methylation may have the ability to act as a “molecular recorder” of past exposures, ranging from heavy metals to stress (Wright et al. 2010; Matosin, Cruceanu, and Binder 2017). We thus retrieved risk factor measurements from seven prior exams in FHS to compare “cumulative” (calculated as the mean of past exam values) versus current correlations with brown activation. Surprisingly, we observed stronger correlations with cumulative values across almost all risk factors. To address the possibility of confounding in these relationships, we tested linear models predicting brown eigenCpG values from current or cumulative risk factors adjusting for the full set of EWAS covariates. Here, we again observed multiple instances of stronger cumulative relationships, especially for BMI, hsCRP, and triglycerides. Though such a finding could be partially explained by the greater stability in a mean over seven values compared to one, we note that we did not observe this same pattern with respect to the blue module, where associations with current risk factors tended to be stronger. Our observation agrees with a conceptual model in which known risk factors, such as the three noted here, act partially through their cumulative

impact over time on immune cell DNA methylation and thus inflammatory processes known to be related to CVD pathogenesis.

To more directly test this proposal, we used a basic mediation approach in which we sequentially tested the relationships between cumulative risk factor levels, brown eigenCpG values, and both factors together in predicting incident CVD. Though neither factor exerted a strong effect on the relationship of the other, module activation associations were more weakened after adjustment for cumulative risk factors than the converse. Thus, our models replicate previous findings that cumulative risk factor exposure correlates with CVD risk (Reinikainen et al. 2015) while suggesting that brown methylation module activation may be sensing, rather than mediating, this effect. One concrete example supporting this observation is the DMR near *SLC1A5* containing primarily brown CpGs, one of which (cg02711608) was suggested in Mendelian randomization analysis to be causally downstream of blood pressure (Richard et al. 2017).

A few limitations should be acknowledged in interpreting the results of this study. First, its observational nature made it impossible to clearly determine causality of the relationships between methylation and cardiovascular risk. While the examination of incident CVD reduced concerns about reverse causation, the discovered associations may only be markers of other disease-causing processes (such as cumulative risk factor exposure, as discussed above). Second, assessment of methylation in blood samples prevented the understanding of potentially causal epigenetic effects in other CVD-relevant tissues. Although some studies report promising findings with respect to blood as a proxy tissue (Bacos et al. 2016; Huang et al. 2016), and although development-related epialleles may persist across tissues, there is a gap in our ability to discover non-blood-related epigenetic patterns in this analysis. Finally, experimental follow-up will be necessary to confirm these findings and establish their potential for supporting therapeutic interventions.

2.1.5 Conclusions

The modules and regions discovered in this investigation provide insights into the complex relationships between DNA methylation and cardiovascular disease risk. We show that epigenetic modules track with diverse biological sources of CVD risk, ranging from development- to immune-related processes, and may provide a molecular readout of past exposure to cardiovascular risk factors. We further discover specific differentially methylated regions that show limited evidence for a causal impact on CVD but may be related to monocyte activation in response to biological stimuli. This work opens the door to further investigation of the epigenetic basis of CVD risk as well as the ability of DNA methylation to act as a biomarker of prior exposures that may be important for disease-relevant prognosis and interventions.

2.1.6 Methods

Study participants and phenotype collection

Data for the discovery set came from a combined case-control and pseudo case-cohort sampling of 2129 women from the Women's Health Initiative study, a larger prospective

cohort beginning in 1993 that included over 160,000 postmenopausal women from across the United States (Anderson et al. 1998). Included subjects had no self-reported CVD at baseline, and cases were chosen based on incident centrally adjudicated angina, revascularization, or CHD event during follow-up. Inclusion criteria for methylation measurement resulted in an oversampling of African American and Hispanic participants. Blood samples used for measurement of DNA methylation and clinical biochemistry were taken at Exam 1. Data are available in the dbGaP public repository (accession: phs000200.v11.p3; downloaded on September 27, 2017).

Data for the validation set came from a substudy of the Framingham Heart Study that measured DNA methylation in 2726 subjects from the Offspring Cohort. The Framingham Offspring Cohort was originally established in 1971 to follow 5209 descendants of the original Framingham Heart Study participants and their spouses (Kannel et al. 1979). Fasting blood samples for both methylation and clinical biochemistry were collected from participants at Exam 8, which took place from 2005-8. Blood samples were also provided for clinical biochemistry measurements in previous exams, constituting the “past exposures” examined here. Data are available in the dbGaP public repository (accession: phs000007.v29.p10; downloaded on September 27, 2017). Adjudicated cardiovascular event data was collected through 2015, and events were defined here as any of: MI, angina pectoris, stroke (approximately 90% being ischemic), or death from CHD (Framingham event codes 1-29).

Blood-based biochemical markers (total cholesterol, low-density lipoprotein cholesterol (LDL), high-density lipoprotein cholesterol (HDL), triglycerides, glucose, hsCRP, and systolic blood pressure) were log10-transformed for all analyses. In addition, median imputation was used to fill missing values for BMI (20 individuals in total), medication use, and smoking status (thus assuming no medication use and no smoking where these values were missing). Diabetes was defined as either use of diabetes medication or a measured fasting blood glucose level of >125 mg/dL. While directly available in WHI, pack-years of smoking was approximated in FHS by multiplying the number of years since starting smoking by the current number of packs per day.

DNA methylation data processing

In both cohorts, DNA methylation data were collected using the Illumina HumanMethylation450 microarray platform (Bibikova et al. 2011) and downloaded as raw intensity files. Preprocessing was performed using the *minfi* and *wateRmelon* packages for R (Aryee et al. 2014; Pidsley et al. 2013). As a quality control step, samples were removed if they showed weak overall signal based on visual inspection of an intensity plot, if they had more than 10% of probes undetected at a detection threshold of $p < 1e-16$, or if the reported sex did not match the predicted sex based on methylation patterns. Probes were removed if they met any of the following criteria: more than 10% of samples undetected at a detection threshold of $p < 1e-16$, location in the X or Y chromosomes, non-CpG probes, cross-hybridizing probes, probes measuring SNPs, and probes with an annotated SNP at the CpG site or in the single-base extension region. Samples were normalized using the Noob method for background correction and dye-bias normalization, followed by the BMIQ method for probe type correction (Fortin, Triche, and Hansen 2016; Teschendorff et al. 2013). For each dataset, principal components analysis was performed on the set of

control probes using code adapted from the CPACOR method of Lehne et al. to account for technical variation (Lehne et al. 2015). Blood cell counts for 6 blood cell types (CD4+ T-cells, CD8+ T-cells, B-cells, natural killer cells, monocytes, and granulocytes) were estimated using a common reference-based method (Houseman et al. 2012). After quality control and filtering steps, 422952 (WHI) and 425326 (FHS) CpG sites remained for downstream analysis, formatted as beta values (ratio of methylated signal to total microarray signal). The vast majority of these sites (422688) were available in both datasets.

Weighted gene correlation network analysis

Weighted gene correlation network analysis (WGCNA) was used to find highly correlated modules of CpG sites (Zhang and Horvath 2005). The full set of 422952 CpGs passing quality control from WHI were used as input. For computational tractability, blockwise module detection was performed, which treats blocks of features separately for network creation and module detection, followed by eventual merging of highly similar modules. To allow for reasonable computation time, the initial pre-clustering analysis (used to inform the choice of blocks) was performed in a random subset of 100 subjects. A block size of 20,000 was used, and a soft-thresholding power of 8 was chosen to balance approximately scale-free network properties with network connectivity. Unsigned networks were used, based on the fact that the biological consequences of an increase vs. decrease in DNA methylation are much less clear than those of gene transcripts. Whole-module behavior was assessed using the first component from a principal components analysis, performed separately for each module. Scree plots were used to inform the variance explained by each module as well as to justify the use of a single eigenvector as a proxy for module behavior. Module preservation assessment was completed in FHS to confirm cross-dataset robustness of modules. The *modulePreservation* function calculates permutation-based $Z_{summary}$ statistics reflecting the preservation of density (of within-module adjacencies) and connectivity (maintenance of within-module node correlations) when modules are evaluated in a test set (Langfelder et al. 2011). EigenCpGs were then calculated (according to the principal component weights from WHI), followed by assessment of associations with incident CVD.

Module associations with cardiovascular disease were assessed using Cox proportional hazards regressions, with eigenCpGs as the independent variable and time-to-event measures for incident CVD as the dependent variable. Minimal models adjusted for estimated blood cell counts as well as technical covariates (DNA pull batch in WHI; analysis center + 7 control-probe principal components in FHS – see EWAS section for details). Fully-adjusted models adjusted additionally for biological covariates (age, BMI, smoking status, and pack-years of smoking; sex in FHS; race in WHI). Proportional hazards checks were implemented (cox.zph function in R), and no violations of the Cox regression assumptions were detected at $p < 0.05$ for any of the modules in WHI or FHS. Mixed models to account for family structure in FHS were also explored, but were found to generate highly similar results (Supp. Table S1).

Epigenome-wide associations of DNA methylation with incident CVD events

For the EWAS analysis, each CpG site was assessed using the same regression framework as in the module-based models, separately in both WHI and FHS. Methylation beta-values replaced eigenCpGs as the independent variable, and the full set of technical and biological covariates, including cell type composition estimates, was used. To remove the influence of beta-value outliers, samples were excluded for each CpG if their beta value was outside of the interval $[25^{\text{th}}\text{ile} - 3 * IQR, > 75^{\text{th}}\text{ile} + 3 * IQR]$. QQ plots and calculation of the genomic inflation factor λ revealed that genomic inflation was not initially adequately controlled in FHS, but after additional adjustment for 7 CPACOR principal components (chosen based on a Scree plot assessment of CPACOR results), a reasonable inflation of $\lambda = 1.09$ was achieved. CPACOR uses principal components analysis on the set of control probes from the methylation array in order to estimate and control for potential batch effects without disturbing biological signal (Lehne et al. 2015). Proportional hazards checks were implemented as in the module-based analysis for the top EWAS hits in WHI, and no systematic departure from the Cox regression assumptions were detected.

Comb-p, implemented as a Python module, was used to call differentially methylated regions (DMRs). The algorithm takes as input p-values from the EWAS, removing the requirement for additional covariate adjustment. Comb-p first calculates an autocorrelation function (ACF), for which a maximum distance of 1kb and a step size of 50 bases were used. Next, it uses the ACF to adjust each p-value using a Stouffer-Liptak-Kechris correction (Kechris, Biehs, and Kornberg 2010), followed by identification of contiguous regions of sites with adjusted p-values below some threshold (here, $p < 0.1$ with no more than 500 bases between neighboring sites in a region). Finally, the ACF is recalculated out to the maximum region size (a step size of 50 was used here as well) and regional p-values are calculated using the Stouffer-Liptak test. For Sidak multiple testing correction of DMRs (Šidák 1967), Comb-p calculates the number of effective tests separately for each DMR as the number of loci tested divided by the number of loci in the region, thus approximating a correction for the total number of regions while accounting for region size.

DMRs were examined to evaluate whether their constituent CpGs contained any residual SNPs-under-probe that escaped filtering based on the Illumina HumanMethylation450 annotation. These checks were performed manually using the UCSC Genome Browser (Kent et al. 2002) and a dbSNP-based annotation track displaying common ($\geq 1\%$ minor allele frequency) variants.

Mendelian randomization (MR) analyses were undertaken for the 4 DMR CpGs that participated in mQTLs in whole blood in the mQTLdb (Gaunt et al. 2016). Relevant mQTL SNPs were retrieved from mQTLdb and used as input to the MR-Base platform (Hemani et al. 2018). MR analysis was run using MR-Egger, weighted median, inverse variance weighted, and weighted mode methods where possible, and the Wald ratio method where only one SNP was available. Outcome summary statistics for MI and CHD were taken from GWAS in the CARDIoGRAMplusC4D consortium (Nikpay et al. 2015). LD-based clumping was used to prune correlated SNPs prior to analysis, with default values used for all other parameters. For follow-up of one SNP of interest, tissue-specific eQTL data were explored in the GTEx Portal (release v7).

Module enrichment analyses

Gene ontology-based enrichment analysis of modules was performed using the *gometh* function from the *missMethyl* package for R (Phipson, Maksimovic, and Oshlack 2015). In this procedure, CpG sites are annotated to genes using the HumanMethylation450 microarray annotation from Illumina, resulting in a binary vector indicating whether each gene is associated with any of the CpG sites of interest (for example, CpGs constituting a module). Prior probabilities for each gene being selected are estimated based on the total number of associated CpG sites on the array. Enrichment analysis is then performed for each gene ontology category using Wallenius' noncentral hypergeometric distribution, which generalizes the basic hypergeometric distribution to account for biased sampling.

Locus-based enrichment analyses were performed using basic two-tailed hypergeometric tests for overlap between module membership and annotation category membership. CpG annotations with respect to both CpG islands (Island, North shore, Open sea, etc.) and genes (TSS1500, 3' UTR, Body, etc.) were retrieved from the standard Illumina HumanMethylation450 microarray annotation. CpG sites were annotated for Polycomb-group target status using embryonic stem cell SUZ12 binding regions retrieved from Lee et al. (Lee et al. 2006). A similar approach was taken to calculate enrichment of epigenetic age CpGs in module sets, replacing annotation category membership with membership in the corresponding epigenetic age biomarker.

Inference of cell type specificity

Epigenomic annotations were used to test for relative enrichment of module CpGs in cell type-specific regulatory regions. Annotations for broad peaks in DNase sensitivity as well as ChIP-seq signal for H3K4me1 and H3K4me3 were obtained for 6 blood cell types (monocytes, natural killer cells, T-cells, B-cells, and hematopoietic stems cells from males and females) from the NIH Roadmap Epigenomics Project database (Kundaje et al. 2015). For each combination of epigenomic feature and cell type, CpGs from the HumanMethylation450 array were classified as to their membership in a peak region. Relative enrichments of in-peak CpGs for modules were then calculated as the ratio of $\frac{\#CpG_{in-peak}}{\#CpG_{total \ module}}$ to $\frac{\#CpG_{in-peak}}{\#CpG_{total \ all}}$ and presented as $\log_2(\text{relative enrichment})$ for ease of visualization. Cell type specificity of different modules can then be compared by examining relative enrichments across cell types, especially with respect to highly-represented regulatory annotation types (e.g. DNase hypersensitivity sites for a module enriched in enhancers). We note that this method borrows from the permutation-based eFORGE tool methodology (Breeze et al. 2016), which could not be used here due to the size of the blue module. However, we confirmed similarity of our results to those from the eFORGE method for the brown module (Supp. Fig. S3).

Cell type-module interaction analyses were undertaken using a similar approach to that of the CellDMC algorithm for cell type-specific differentially-methylated cytosine discovery (Zheng et al. 2018). Current implementations of CellDMC do not support time-to-event models. Partially-adjusted models were built in FHS as was done for the initial module-CVD replication, while including interaction terms between the brown module and each of 5 estimated cell type fractions (leaving out estimated granulocytes). The reported

interaction term of interest was that between the brown module and monocyte fractions as influences time-to-event for CVD.

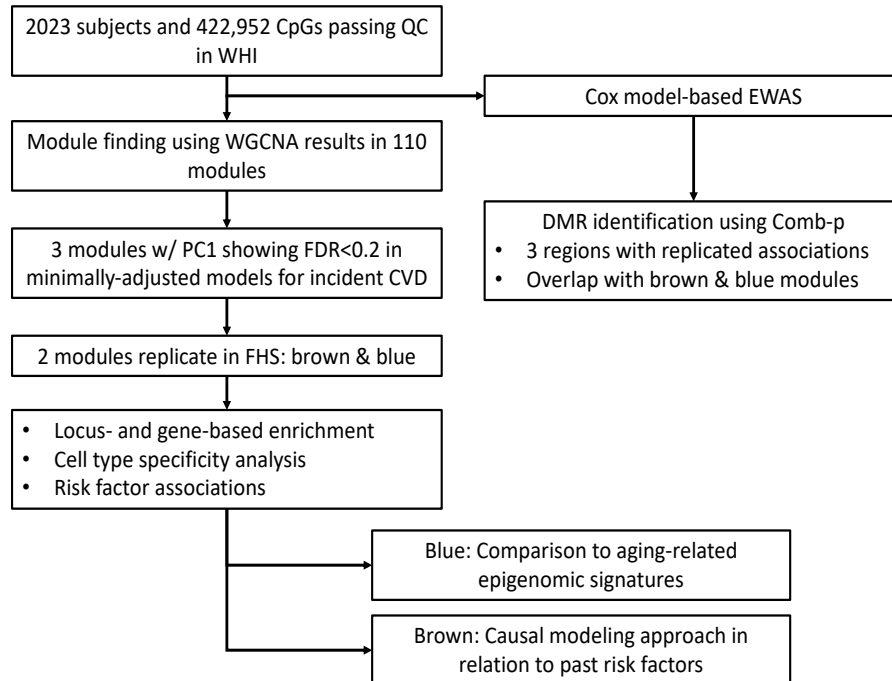
Risk factor integration

Risk factors were incorporated into the module-based analysis in a series of steps. First, Pearson correlations between risk factor levels and module eigenCpGs were calculated to provide a high-level understanding of the strength of their relationship. Risk factors in WHI were all measured at Exam 1 (concurrently with the methylation measurement), while risk factors in FHS were collected for all exams prior to and including Exam 8 (the time of the methylation measurement). In FHS, correlations with past risk factor levels as well as a “cumulative” exposure level (equal to the mean of each set of risk factor levels from Exam 1-7) were also calculated.

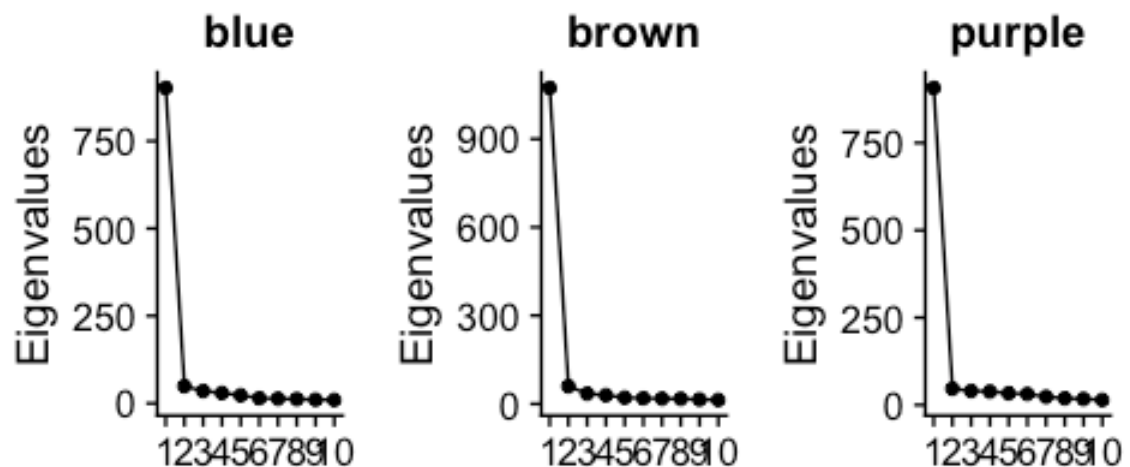
Next, linear models were used to assess these same module-risk factor correlations in FHS while adjusting for potential confounding variables. These models predicted module eigenCpGs using either cumulative (Exams 1-7) or current (Exam 8) risk factors, while adjusting for the same set of technical and biological covariates as in the EWAS (described above). In this step, both eigenCpGs and risk factors were standardized before modeling in order to facilitate effect size comparisons across risk factors and across modules.

Finally, the relationship between cumulative risk factors, the brown module, and incident CVD was examined, using the same Cox regression setup as in the EWAS to perform a basic mediation analysis for BMI, hsCRP, and triglycerides. Here, cumulative risk factor exposure (as defined above) acted as the exposure, brown methylation module activation (represented by the brown eigenCpG) acted as the mediator, and incident CVD acted as the outcome. Having established the strong exposure-mediator links, three subsequent Cox models were examined: cumulative risk factors only, brown eigenCpG only, and both simultaneously. All models adjusted for the full set of technical and biological covariates as well as the “current” level (i.e. at Exam 8) of the risk factor in question.

2.1.7 Supplementary Data



Supp. Fig. S1: Study overview, including module- and region-based analyses as well as follow-up.



Supp. Fig. S2: Scree plots for PCA on the set of CpGs corresponding to each of the top modules.

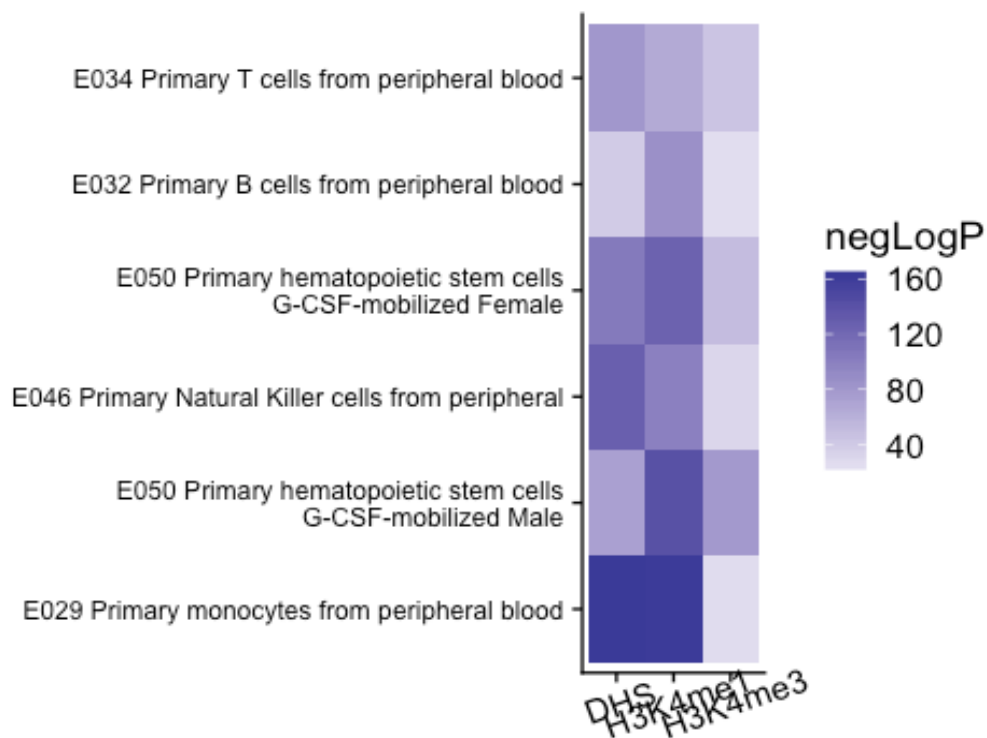
Supp. Table S1: *P-values for module associations with incident CVD in discovery and replication.*

Module	WHI (discovery)		FHS (replication)		
	Partially adjusted	Fully adjusted	Partially adj.	Partially adj. (mixed)	Fully adjusted
Blue	0.0002736	0.0500018	0.0000085	0.0000085	0.8189348
Brown	0.0045462	0.0872688	0.0000962	0.0000963	0.0997390
Purple	0.0050028	0.0210976	0.0202819	0.0202804	0.1580861

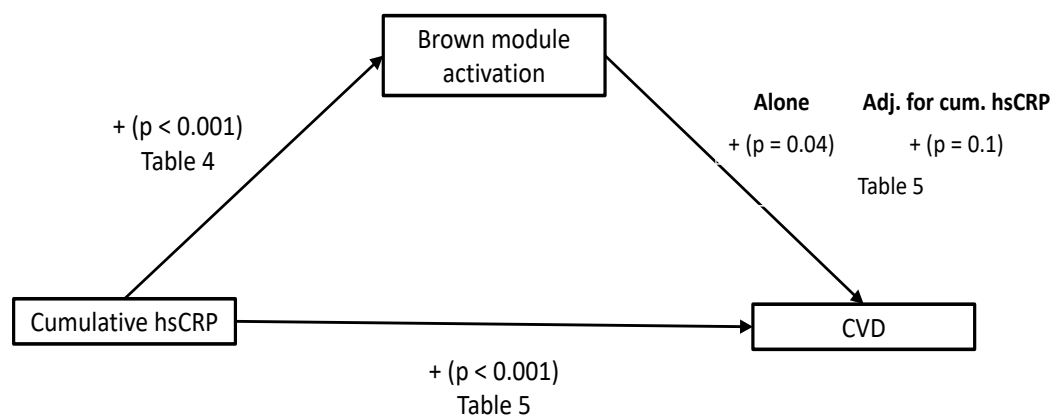
Partially-adjusted models are adjusted for technical covariates (DNA pull batch in WHI and study center + 7 control probe PCs in FHS) and estimated cell counts. Fully-adjusted models are additionally adjusted for age, sex, smoking status and smoking pack-years. Mixed model contains a random intercept for each family.

Supp. Table S2: *CpGs with FDR < 0.05 in the discovery set (Bonferroni threshold = 1.18e-7)*

CpG	Chromosome	Dir. of Assoc.	P-value	Location	Annotated Gene	Replication P-value
cg09155044	chr16	+	6.63e-09	TSS1500	VKORC1	0.107
cg24434800	chr1	+	5.04e-08			0.629
cg11691298	chr2	+	1.1e-07	Body	FAM59B	0.525
cg02379107	chr20	+	4.72e-07	TSS1500	KIAA1755	0.930



Supp. Fig. S3: eFORGE cell type-specificity plot for the brown module.



Supp. Fig. S4: Example diagram of cumulative risk factor mediation by brown methylation module activation. Results from 4 regressions are shown: cumulative risk factor exposure to brown activation, cumulative risk factor exposure to incident CVD, and brown activation to incident CVD with and without adjustment for cumulative risk factor exposure. Regression terms represented as: sign of coefficient (p -value).

2.1.8 References

- Aguet, François, Andrew A. Brown, Stephane E. Castel, Joe R. Davis, Yuan He, Brian Jo, Pejman Mohammadi, et al. 2017. "Genetic effects on gene expression across human tissues." *Nature*. <https://doi.org/10.1038/nature24277>.
- Anderson, Garnet L, Steven R Cummings, Laurence S Freedman, Curt Furberg, Maureen M Henderson, Susan R Johnson, Lewis H Kuller, et al. 1998. "Design of the Women's Health Initiative Clinical Trial and Observational Study." *Controlled Clinical Trials* 19 (1): 61–109. [https://doi.org/10.1016/S0197-2456\(97\)00078-0](https://doi.org/10.1016/S0197-2456(97)00078-0).
- Aryee, Martin J., Andrew E. Jaffe, Hector Corrada-Bravo, Christine Ladd-Acosta, Andrew P. Feinberg, Kasper D. Hansen, and Rafael A. Irizarry. 2014. "Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays." *Bioinformatics* 30 (10). Oxford University Press: 1363–9. <https://doi.org/10.1093/bioinformatics/btu049>.
- Aslibekyan, Stella, Golareh Agha, Elena Colicino, Anh N. Do, Jari Lahti, Symen Ligthart, Riccardo E. Marioni, et al. 2018. "Association of Methylation Signals With Incident Coronary Heart Disease in an Epigenome-Wide Assessment of Circulating Tumor Necrosis Factor α ." *JAMA Cardiology* 3 (6): 463–72. <https://doi.org/10.1001/JAMACARDIO.2018.0510>.
- Baccarelli, A, R Wright, V Bollati, A Litonjua, A Zanobetti, L Tarantini, D Sparrow, P Vokonas, and J Schwartz. 2010. "Ischemic heart disease and stroke in relation to blood DNA methylation." *Epidemiology* 21 (6): 819–28. <https://doi.org/10.1097/EDE.0b013e3181f20457>.
- Bacos, Karl, Linn Gillberg, Petr Volkov, Anders H Olsson, Torben Hansen, Oluf Pedersen, Anette Prior Gjesing, et al. 2016. "Blood-based biomarkers of age-associated epigenetic changes in human islets associate with insulin secretion and diabetes." *Nature Communications* 7. Nature Publishing Group: 11089. <https://doi.org/10.1038/ncomms11089>.
- Baron, Reuben M., and David A. Kenny. 1986. "The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations." *Journal of Personality and Social Psychology* 51 (6): 1173–82. <https://doi.org/10.1037/0022-3514.51.6.1173>.
- Bibikova, Marina, Bret Barnes, Chan Tsan, Vincent Ho, Brandy Klotzle, Jennie M. Le, David Delano, et al. 2011. "High density DNA methylation array with single CpG site resolution." *Genomics* 98 (4): 288–95. <https://doi.org/10.1016/j.ygeno.2011.07.007>.
- Bonder, Marc Jan, René Luijk, Daria V Zhernakova, Matthijs Moed, Patrick Deelen, Martijn Vermaat, Maarten van Iterson, et al. 2016. "Disease variants alter transcription factor levels and methylation of their binding sites." *Nature Genetics* 49 (1): 131–38. <https://doi.org/10.1038/ng.3721>.
- Breeze, Charles E., Dirk S. Paul, Jenny van Dongen, Lee M. Butcher, John C. Ambrose, James E. Barrett, Robert Lowe, et al. 2016. "eFORGE: A Tool for Identifying Cell Type-Specific Signal in Epigenomic Data." *Cell Reports* 17 (8). Cell Press: 2137–50. <https://doi.org/10.1016/j.celrep.2016.10.059>.
- Busch, Robert, Weiliang Qiu, Jessica Lasky-Su, Jarrett Morrow, Gerard Criner, and Dawn DeMeo. 2016. "Differential DNA methylation marks and gene comethylation of COPD in African-Americans with COPD exacerbations." *Respiratory Research* 17 (1): 143. <https://doi.org/10.1186/s12931-016-0459-8>.

Dekkers, Koen F, Maarten van Iterson, Roderick C Sliker, Matthijs H Moed, Marc Jan Bonder, Michiel van Galen, Hailiang Mei, et al. 2016. "Blood lipids influence DNA methylation in circulating cells." *Genome Biology* 17 (1): 138. <https://doi.org/10.1186/s13059-016-1000-6>.

Dozmorov, Mikhail G. 2015. "Polycomb repressive complex 2 epigenomic signature defines age-associated hypermethylation and gene expression changes." *Epigenetics* 10 (6). Taylor & Francis: 484–95. <https://doi.org/10.1080/15592294.2015.1040619>.

Farlik, Matthias, Florian Halbritter, Fabian Müller, Fizzah A Choudry, Peter Ebert, Johanna Klughammer, Samantha Farrow, et al. 2016. "DNA Methylation Dynamics of Human Hematopoietic Stem Cell Differentiation." *Cell Stem Cell* 19 (6). Elsevier: 808–22. <https://doi.org/10.1016/j.stem.2016.10.019>.

Fernández-Sanlés, Alba, Sergi Sayols-Baixeras, Isaac Subirana, Irene R Degano, and Roberto Elosua. 2017. "Association between DNA methylation and coronary heart disease or other atherosclerotic events: A systematic review." <https://doi.org/10.1016/j.atherosclerosis.2017.05.022>.

Fortin, Jean-Philippe, Timothy J Triche, and Kasper D Hansen. 2016. "Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi." *Bioinformatics* 33 (4). Oxford University Press: btw691. <https://doi.org/10.1093/bioinformatics/btw691>.

Gaunt, Tom R, Hashem A Shihab, Gibran Hemani, Josine L Min, Geoff Woodward, Oliver Lyttleton, Jie Zheng, et al. 2016. "Systematic identification of genetic influences on methylation across the human life course." *Genome Biology* 17 (1): 61. <https://doi.org/10.1186/s13059-016-0926-z>.

Guarrera, Simonetta, Giovanni Fiorito, N Charlotte Onland-Moret, Alessia Russo, Claudia Agnoli, Alessandra Allione, Cornelia Di Gaetano, et al. 2015. "Gene-specific DNA methylation profiles and LINE-1 hypomethylation are associated with myocardial infarction risk." *Clinical Epigenetics* 7 (1): 133. <https://doi.org/10.1186/s13148-015-0164-3>.

Hannum, Gregory, Justin Guiney, Ling Zhao, Li Zhang, Guy Hughes, Srinivas Sadda, Brandy Klotzle, et al. 2013. "Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates." *Molecular Cell* 49 (2): 359–67. <https://doi.org/10.1016/j.molcel.2012.10.016>.

Hao, Guang, Nagy A. Youssef, Catherine L. Davis, and Shaoyong Su. 2018. "The role of DNA methylation in the association between childhood adversity and cardiometabolic disease." *International Journal of Cardiology* 255 (March). Elsevier: 168–74. <https://doi.org/10.1016/j.ijcard.2017.12.063>.

Hedman, Åsa K., Michael M. Mendelson, Riccardo E. Marioni, Stefan Gustafsson, Roby Joehanes, Marguerite R. Irvin, Degui Zhi, et al. 2017. "Epigenetic Patterns in Blood Associated With Lipid Traits Predict Incident Coronary Heart Disease Events and Are Enriched for Results From Genome-Wide Association Studies." *Circulation: Cardiovascular Genetics* 10 (1): e001487. <https://doi.org/10.1161/CIRCGENETICS.116.001487>.

Hemani, Gibran, Jie Zheng, Benjamin Elsworth, Kaitlin H Wade, Valeriia Haberland, Denis Baird, Charles Laurin, et al. 2018. "The MR-Base platform supports systematic causal inference across the human phenome." *eLife* 7 (May). <https://doi.org/10.7554/eLife.34408>.

Horvath, Steve. 2013. "DNA methylation age of human tissues and cell types." *Genome Biology* 14 (10): R115. <https://doi.org/10.1186/gb-2013-14-10-r115>.

- Horvath, Steve, Yafeng Zhang, Peter Langfelder, René S Kahn, Marco P.M. Boks, Kristel van Eijk, Leonard H van den Berg, and Roel A Ophoff. 2012. "Aging effects on DNA methylation modules in human brain and blood tissue." *Genome Biology* 13 (10). BioMed Central: R97. <https://doi.org/10.1186/gb-2012-13-10-r97>.
- Houseman, Eugene Andres, William P Accomando, Devin C Koestler, Brock C Christensen, Carmen J Marsit, Heather H Nelson, John K Wiencke, and Karl T Kelsey. 2012. "DNA methylation arrays as surrogate measures of cell mixture distribution." *BMC Bioinformatics* 13 (1): 86. <https://doi.org/10.1186/1471-2105-13-86>.
- Huang, Yen Tsung, Su Chu, Eric B Loucks, Chien Ling Lin, Charles B Eaton, Stephen L Buka, and Karl T Kelsey. 2016. "Epigenome-wide profiling of DNA methylation in paired samples of adipose tissue and blood." *Epigenetics* 11 (3): 227–36. <https://doi.org/10.1080/15592294.2016.1146853>.
- Irvin, Marguerite R., Degui Zhi, Roby Joehanes, Michael Mendelson, Stella Aslibekyan, Steven A. Claas, Krista S. Thibeault, et al. 2014. "Epigenome-wide association study of fasting blood lipids in the genetics of lipid-lowering drugs and diet network study." *Circulation* 130 (7): 565–72. <https://doi.org/10.1161/CIRCULATIONAHA.114.009158>.
- Jaffe, Andrew E, Peter Murakami, Hwajin Lee, Jeffrey T Leek, M Daniele Fallin, Andrew P Feinberg, and Rafael A Irizarry. 2012. "Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies." *International Journal of Epidemiology* 41 (1): 200–209. <https://doi.org/10.1093/ije/dyr238>.
- Kannel, William B, Manning Feinleib, Patricia M. Mcnamara, Robert J Garrison, and William P Castelli. 1979. "An investigation of coronary heart disease in families: The framingham offspring study." *American Journal of Epidemiology* 110 (3): 281–90. <https://doi.org/10.1093/oxfordjournals.aje.a112813>.
- Kechris, Katerina J., Brian Biehs, and Thomas B. Kornberg. 2010. "Generalizing Moving Averages for Tiling Arrays Using Combined P-Value Statistics." *Statistical Applications in Genetics and Molecular Biology* 9 (1). <https://doi.org/10.2202/1544-6115.1434>.
- Kent, W James, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and a. D. Haussler. 2002. "The Human Genome Browser at UCSC." *Genome Research* 12 (6). Cold Spring Harbor Laboratory Press: 996–1006. <https://doi.org/10.1101/gr.229102>.
- Kundaje, Anshul, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, et al. 2015. "Integrative analysis of 111 reference human epigenomes." *Nature* 518 (7539): 317–30. <https://doi.org/10.1038/nature14248>.
- Laios, Michael D., and Everett R. Tate. 2015. "Fetal Hematopoietic Stem Cells Are the Canaries in the Coal Mine That Portend Later Life Immune Deficiency." *Endocrinology* 156 (10). Oxford University Press: 3458–65. <https://doi.org/10.1210/en.2015-1347>.
- Lam, Lucia L, Eldon Emberly, Hunter B Fraser, Sarah M Neumann, Edith Chen, Gregory E Miller, and Michael S Kobor. 2012. "Factors underlying variable DNA methylation in a human community cohort." *Proceedings of the National Academy of Sciences* 109 (October): 17253–60. <https://doi.org/10.1073/pnas.1121249109>.
- Langfelder, Peter, and Steve Horvath. 2008. "WGCNA: An R package for weighted correlation network analysis." *BMC Bioinformatics* 9. <https://doi.org/10.1186/1471-2105-9-559>.

- Langfelder, Peter, Rui Luo, Michael C Oldham, and Steve Horvath. 2011. "Is My Network Module Preserved and Reproducible?" Edited by Philip E. Bourne. *PLoS Computational Biology* 7 (1). Public Library of Science: e1001057. <https://doi.org/10.1371/journal.pcbi.1001057>.
- Lee, Tong Ihn, Richard G. Jenner, Laurie A. Boyer, Matthew G. Guenther, Stuart S. Levine, Roshan M. Kumar, Brett Chevalier, et al. 2006. "Control of Developmental Regulators by Polycomb in Human Embryonic Stem Cells." *Cell* 125 (2). Cell Press: 301–13. <https://doi.org/10.1016/j.cell.2006.02.043>.
- Lehne, Benjamin, Alexander W. Drong, Marie Loh, Weihua Zhang, William R. Scott, Sian Tsung Tan, Uzma Afzal, et al. 2015. "A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies." *Genome Biology* 16 (1). <https://doi.org/10.1186/s13059-015-0600-x>.
- Levine, Morgan E., Ake T. Lu, Austin Quach, Brian H. Chen, Themistocles L. Assimes, Stefania Bandinelli, Lifang Hou, et al. 2018. "An epigenetic biomarker of aging for lifespan and healthspan." *Aging* 10 (4): 573–91. <https://doi.org/10.18632/aging.101414>.
- Li, Jun, Xiaoyan Zhu, Kuai Yu, Haijing Jiang, Yizhi Zhang, Siyun Deng, Longxian Cheng, et al. 2017. "Genome-Wide Analysis of DNA Methylation and Acute Coronary Syndrome." *Circulation Research* 120 (11): 1754–67. <https://doi.org/10.1161/CIRCRESAHA.116.310324>.
- Matosin, Natalie, Cristiana Cruceanu, and Elisabeth B Binder. 2017. "Preclinical and Clinical Evidence of DNA Methylation Changes in Response to Trauma and Chronic Stress." *Chronic Stress* 1 (January): 247054701771076. <https://doi.org/10.1177/2470547017710764>.
- Mendelson, Michael M, Riccardo E Marioni, Roby Joehanes, Chunyu Liu, Åsa K Hedman, Stella Aslibekyan, Ellen W Demerath, et al. 2017. "Association of Body Mass Index with DNA Methylation and Gene Expression in Blood Cells and Relations to Cardiometabolic Disease: A Mendelian Randomization Approach." Edited by Cathryn Lewis. *PLOS Medicine* 14 (1): e1002215. <https://doi.org/10.1371/journal.pmed.1002215>.
- Nahrendorf, Matthias. 2018. "Myeloid cell contributions to cardiovascular health and disease." <https://doi.org/10.1038/s41591-018-0064-0>.
- Nakatochi, Masahiro, Sahoko Ichihara, Ken Yamamoto, Keiko Naruse, Shigeki Yokota, Hiroyuki Asano, Tatsuaki Matsubara, and Mitsuhiro Yokota. 2017. "Epigenome-wide association of myocardial infarction with DNA methylation sites at loci related to cardiovascular disease." *Clinical Epigenetics* 9 (1). <https://doi.org/10.1186/s13148-017-0353-3>.
- Nikpay, Majid, Anuj Goel, Hong Hee Won, Leanne M. Hall, Christina Willenborg, Stavroula Kanoni, Danish Saleheen, et al. 2015. "A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease." *Nature Genetics* 47 (10): 1121–30. <https://doi.org/10.1038/ng.3396>.
- Ordovás, José M., and Caren E. Smith. 2010. "Epigenetics and cardiovascular disease." *Nature Reviews Cardiology* 7 (9): 510–19. <https://doi.org/10.1038/nrcardio.2010.104>.
- Pedersen, Brent S., David A. Schwartz, Ivana V. Yang, and Katerina J. Kechris. 2012. "Comb-p: Software for combining, analyzing, grouping and correcting spatially correlated P-values." *Bioinformatics* 28 (22): 2986–8. <https://doi.org/10.1093/bioinformatics/bts545>.
- Pfeiffer, Liliane, Simone Wahl, Luke C. Pilling, Eva Reischl, Johanna K. Sandling, Sonja Kunze, Lesca M. Holdt, et al. 2015. "DNA Methylation of Lipid-Related Genes Affects Blood Lipid

Levels.” *Circulation: Cardiovascular Genetics* 8 (2): 334–42. <https://doi.org/10.1161/CIRCGENETICS.114.000804>.

Phipson, Belinda, Jovana Maksimovic, and Alicia Oshlack. 2015. “MissMethyl: An R package for analyzing data from Illumina’s HumanMethylation450 platform.” *Bioinformatics* 32 (2). Oxford University Press: 286–88. <https://doi.org/10.1093/bioinformatics/btv560>.

Pidsley, Ruth, Chloe C Y Wong, Manuela Volta, Katie Lunnon, Jonathan Mill, and Leonard C Schalkwyk. 2013. “A data-driven approach to preprocessing Illumina 450K methylation array data.” *BMC Genomics* 14 (1). BioMed Central: 293. <https://doi.org/10.1186/1471-2164-14-293>.

Rask-Andersen, Mathias, David Martinsson, Muhammad Ahsan, Stefan Enroth, Weronica E Ek, Ulf Gyllensten, and Åsa Johansson. 2016. “Epigenome-wide association study reveals differential DNA methylation in individuals with a history of myocardial infarction.” *Human Molecular Genetics* 25 (21): 4739–48. <https://doi.org/10.1093/hmg/ddw302>.

Reinikainen, Jaakko, Tiina Laatikainen, Juha Karvanen, and Hanna Tolonen. 2015. “Lifetime cumulative risk factors predict cardiovascular disease mortality in a 50-year follow-up study in Finland.” *International Journal of Epidemiology* 44 (1). Oxford University Press: 108–16. <https://doi.org/10.1093/ije/dyu235>.

Richard, Melissa A., Tianxiao Huan, Symen Ligthart, Rahul Gondalia, Min A. Jhun, Jennifer A. Brody, Marguerite R. Irvin, et al. 2017. “DNA Methylation Analysis Identifies Loci for Blood Pressure Regulation.” *American Journal of Human Genetics* 101 (6). Cell Press: 888–902. <https://doi.org/10.1016/j.ajhg.2017.09.028>.

Sarigianni, Maria. 2010. “Involvement of Signaling Molecules on Na⁺/H⁺ Exchanger-1 Activity in Human Monocytes.” *The Open Cardiovascular Medicine Journal* 4 (1): 181–88. <https://doi.org/10.2174/1874192401004010181>.

Short, John D., Sina Tavakoli, Huynh Nga Nguyen, Ana Carrera, Chelbee Farnen, Laura A. Cox, and Reto Asmis. 2017. “Dyslipidemic Diet-Induced Monocyte ‘Priming’ and Dysfunction in Non-Human Primates Is Triggered by Elevated Plasma Cholesterol and Accompanied by Altered Histone Acetylation.” *Frontiers in Immunology* 8 (August). Frontiers: 958. <https://doi.org/10.3389/fimmu.2017.00958>.

Šidák, Zbyněk. 1967. “Rectangular Confidence Regions for the Means of Multivariate Normal Distributions.” *Journal of the American Statistical Association* 62 (318): 626–33. <https://doi.org/10.1080/01621459.1967.10482935>.

Teschendorff, Andrew E, Francesco Marabita, Matthias Lechner, Thomas Bartlett, Jesper Tegner, David Gomez-Cabrero, and Stephan Beck. 2013. “A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data.” *Bioinformatics* 29 (2): 189–96. <https://doi.org/10.1093/bioinformatics/bts680>.

Villard, Eric, Claire Perret, Françoise Gary, Carole Proust, Gilles Dilanian, Christian Hengstenberg, Volker Ruppert, et al. 2011. “A genome-wide association study identifies two loci associated with heart failure due to dilated cardiomyopathy.” *European Heart Journal* 32 (9). Oxford University Press: 1065–76. <https://doi.org/10.1093/eurheartj/ehr105>.

Wahl, Simone, Alexander Drong, Benjamin Lehne, Marie Loh, William R. Scott, Sonja Kunze, Pei-Chien Tsai, et al. 2017. “Epigenome-wide association study of body mass index and the adverse outcomes of adiposity.” *Nature* 541 (7635): 81–86. <https://doi.org/10.1038/nature20784>.

Walaszczyk, Eliza, Mirjam Luijten, Annemieke M. W. Spijkerman, Marc J. Bonder, Helen L. Lutgers, Harold Snieder, Bruce H. R. Wolffenbuttel, and Jana V. van Vliet-Ostaptchouk. 2018. “DNA methylation markers associated with type 2 diabetes, fasting glucose and HbA1c levels: a systematic review and replication in a case–control sample of the Lifelines study.” *Diabetologia* 61 (2). Springer Berlin Heidelberg: 354–68. <https://doi.org/10.1007/s00125-017-4497-7>.

Wallner, Stefan, Christopher Schröder, Elsa Leitão, Tea Berulava, Claudia Haak, Daniela Beißer, Sven Rahmann, et al. 2016. “Epigenetic dynamics of monocyte-to-macrophage differentiation.” *Epigenetics & Chromatin* 9 (1): 33. <https://doi.org/10.1186/s13072-016-0079-z>.

Wright, Robert O., Joel Schwartz, Rosalind J. Wright, Valentina Bollati, Letizia Tarantini, Sung Kyun Park, Howard Hu, David Sparrow, Pantel Vokonas, and Andrea Baccarelli. 2010. “Biomarkers of lead exposure and DNA methylation within retrotransposons.” *Environmental Health Perspectives* 118 (6): 790–95. <https://doi.org/10.1289/ehp.0901429>.

Zhang, Bin, and Steve Horvath. 2005. “A General Framework for Weighted Gene Co-Expression Network Analysis.” *Statistical Applications in Genetics and Molecular Biology* 4 (1): 2194. <https://doi.org/10.2202/1544-6115.1128>.

Zhang, Yan, Rory Wilson, Jonathan Heiss, Lutz P. Breitling, Kai-Uwe Saum, Ben Schöttker, Bernd Holleczeck, Melanie Waldenberger, Annette Peters, and Hermann Brenner. 2017. “DNA methylation signatures in peripheral blood strongly predict all-cause mortality.” *Nature Communications* 8 (March): 14617. <https://doi.org/10.1038/ncomms14617>.

Zheng, Shijie C., Charles E. Breeze, Stephan Beck, and Andrew E. Teschendorff. 2018. “Identification of differentially methylated cell types in epigenome-wide association studies.” *Nature Methods* 15 (12): 1059–66. <https://doi.org/10.1038/s41592-018-0213-x>.

Chapter 3

Epigenomic Prediction of CVD Risk

3.1 Epigenomic prediction of cardiovascular disease risk and interactions with traditional risk metrics

In preparation for submission to Nature Communications

3.1.1 Abstract

Epigenome-wide association studies for cardiometabolic risk factors have discovered multiple loci associated with incident cardiovascular disease (CVD). However, few studies have sought to directly optimize a predictor of CVD risk. Furthermore, it is traditionally challenging to train multivariate models across multiple studies in the presence of study- or batch effects. Here, we analyzed existing DNA methylation data collected using the Illumina HumanMethylation450 microarray to create a predictor of CVD risk across three cohorts: Women's Health Initiative, Framingham Heart Study Offspring Cohort, and Lothian Birth Cohorts. We trained Cox proportional hazards-based elastic net regressions for incident CVD separately in each cohort, and used a recently-introduced cross-study learning approach to integrate these individual predictions into an ensemble predictor. The methylation-based risk score (MRS) predicted CVD time-to-event in a held-out fraction of the Framingham dataset (HR per SD = 1.32, $p = 2e-4$) and predicted myocardial infarction status in the independent REGICOR dataset (OR per SD = 2.2, $p = 1e-6$). These associations remained after adjustment for traditional cardiovascular risk factors, and were consistently stronger than those from elastic net models trained on a directly merged dataset. Additionally, we investigated interactions between the MRS and both genetic and biochemical CVD risk, showing preliminary evidence of an enhanced predictive power in those with less traditional risk factor elevation and suggesting that the MRS may enable the discovery of high-risk individuals that would be missed by alternative risk metrics.

3.1.2 Introduction

DNA methylation is an important epigenetic pathway through which genetic variants and environmental exposures impact disease risk (Bonder et al. 2016; Tobi et al. 2018). Methylation at specific cytosine-phosphate-guanine (CpG) sites has been associated with disease in epigenome-wide association studies, even showing associations in blood as a convenient but non-target tissue such as for type 2 diabetes (Bacos et al. 2016). Methylation-based risk scores allow genome-wide aggregation of epigenetic information, similarly to the more established genetic risk scores, and allow for the use of models with arbitrary complexity. These risk scores are often developed initially by using methylation as a proxy for disease risk factors, such as BMI (Wahl et al. 2017) and general aging-related morbidity (Levine et al. 2018). Alternatively, given sufficient sample size, epigenetic associations with disease risk can be modeled directly (Hao et al. 2017).

Associations between DNA methylation and cardiovascular disease (CVD) have been explored in many different cohorts and using diverse approaches. Cross-sectional associations have been found across multiple relevant tissues, namely blood, aorta, and other vascular tissues (Fernández-Sanlés et al. 2017). Some investigations aimed at cardiovascular risk factors have discovered CpGs predictive of CVD development

(Hedman et al. 2017; Aslibekyan et al. 2018), while Mendelian randomization approaches have suggested causality of at least some of these CpG-risk factor associations (Richardson et al. 2017). The few studies directly modeling incident CVD as a primary outcome have either been conducted using only global (not locus-specific) methylation levels (Baccarelli et al. 2010), or have found limited additional predictive power in the presence of known risk factors (Guarrera et al. 2015). We have previously investigated methylation regions and modules associating with incident CVD, generating mechanistic insights but without aggregating these results into a direct predictor of risk (Westerman et al. 2018). Additionally, it is unclear how the CVD risk tracked by DNA methylation is redundant with or complementary to existing risk metrics, including genetic scores (Khera et al. 2018) and those based on traditional cardiovascular risk factors (e.g. the Framingham Risk Score for generalized CVD) (D’Agostino et al. 2008).

Due to the many sources of heterogeneity across population-scale cohorts, it is beneficial to combine signal across cohorts when performing association studies or training predictive genomic models. Such a strategy can increase sample size while attenuating the effect of study-specific biases and confounding factors, but can be prone to emergent sources of confounding from “batch” effects or other systematic biases in methylation data across cohorts. This is especially problematic when there is notable class imbalance (different outcome frequencies or distributions) across cohorts (Goh, Wang, and Wong 2017). The most common and straightforward method for dealing with this heterogeneity is to use meta-analysis, but standard meta-analysis approaches are restricted to univariate (i.e. one CpG site at a time) models. Other approaches include batch effect correction on the input dataset (e.g. ComBat (Johnson, Li, and Rabinovic 2007)), direct adjustment for batch/study in linear models, or adjustment for derived variables intended to capture technical biases (e.g. surrogate variable analysis (Leek and Storey 2007)), but these approaches can often lead to over- or under-estimates of true biological effects (Goh, Wang, and Wong 2017), and may be less amenable to complex or nonlinear models. Another approach described recently, cross-study learning, uses an alternative strategy to tackle the problem by creating an ensemble predictor consisting of one or multiple models per cohort, all trained to predict the same outcome (Patil and Parmigiani 2018). This strategy allows the use of arbitrarily complex models and the contribution of multiple training cohorts while avoiding technical confounding from direct combination of the datasets.

In order to develop an improved DNA methylation-based cardiovascular risk predictor using multiple training cohorts, we used a cross-study learning method to develop an ensemble of penalized time-to-event regression risk models. The resulting composite risk score performed well in a held-out data subset, predicting survival even in the presence of traditional risk factors, and surpassing the performance of models trained on naively merged datasets. External validation was achieved in a case-control for prevalent myocardial infarction (MI). Further, interactions were assessed between the composite methylation-based risk score and other risk predictors, finding no consistent relationship to a genome-wide genetic risk score, but a potentially enhanced prediction in those with low Framingham Risk Scores.

3.1.3 Results

Cross-study learner model development

Epigenomic model development was performed in three cohorts, including the Women's Health Initiative (WHI), Framingham Heart Study Offspring Cohort (FHS), and Lothian Birth Cohorts of 1936 (LBC). The FHS dataset was divided into two functionally separate groups (FHS-JHU and FHS-UM) based on differences in subject selection and geographic location of laboratory methylation analysis (see Methods). Further details of the populations can be found in Table 1.

Table 1: Baseline parameters of the populations used for model development

	WHI	FHS-JHU	LBC	FHS-UM
Sample size	2023	484	818	2103
Age	65 (59-70)	71 (64-77)	69 (68-70)	64 (59-71)
Sex (% female)	100	30	50	60
% European	47	100	100	100
% African American	32	0	0	0
% Hispanic	20	0	0	0
Body mass index	29.1 (25.5-33.3)	28.2 (25.5-31.3)	27.5 (24.9-30.3)	27.4 (24.3-31)
LDL cholesterol	150 (126-175)	88 (73-107)	118 (89.5-150.3)	107 (87-128)
HDL cholesterol	51 (43-60)	49 (40-60)	56.1 (47.2-68.3)	56 (45.8-69)
Triglycerides	127 (92-177)	101.5 (75-141.2)	128.4 (97.4-171.2)	102 (73-142)
Fasting glucose	96 (88.6-108)	106 (97-116)	Unavailable	100 (94-109)
Systolic blood pressure	131 (120-143)	130 (117-143)	148.7 (137-161.3)	126 (116-138)
Prior only	0	127	70	112
Incident only	1009	67	133	146
Prior and incident	0	58	164	34

* Continuous values shown as: median (interquartile range).

* WHI = Women's Health Initiative, FHS-JHU = Framingham Heart Study Offspring Cohort (Johns Hopkins University subset), LBC = Lothian Birth Cohorts 1936, FHS-UM = Framingham Heart Study Offspring Cohort (University of Minnesota subset).

Fig. 1 outlines the computational workflow. Briefly, a cross-study learning (CSL) model was developed by training time-to-event elastic net regressions on three of the datasets, while holding out the FHS-UM subset for evaluation. Next, a model trained on all four datasets was subject to external replication in the REGICOR study. CSL model CpGs were characterized as to their potential biological function, and model performance was assessed across strata of alternative cardiovascular risk metrics.

The initial predictor was developed by training individual penalized Cox proportional hazards regression models (single-study learners, or SSLs) in each of the three training cohorts (WHI, FHS-JHU, and LBC). Predictions from these models were aggregated through the "stacking" method, in which the outcomes and model predictions from each of the individual datasets are combined, and a regression is used to assign weights to each of the model predictions (see Methods). This procedure led to FHS-JHU dropping out of the ensemble model, with weights for this initial predictor as follows: 0.6 (WHI), 0.0 (FHS-JHU), and 0.4 (LBC).

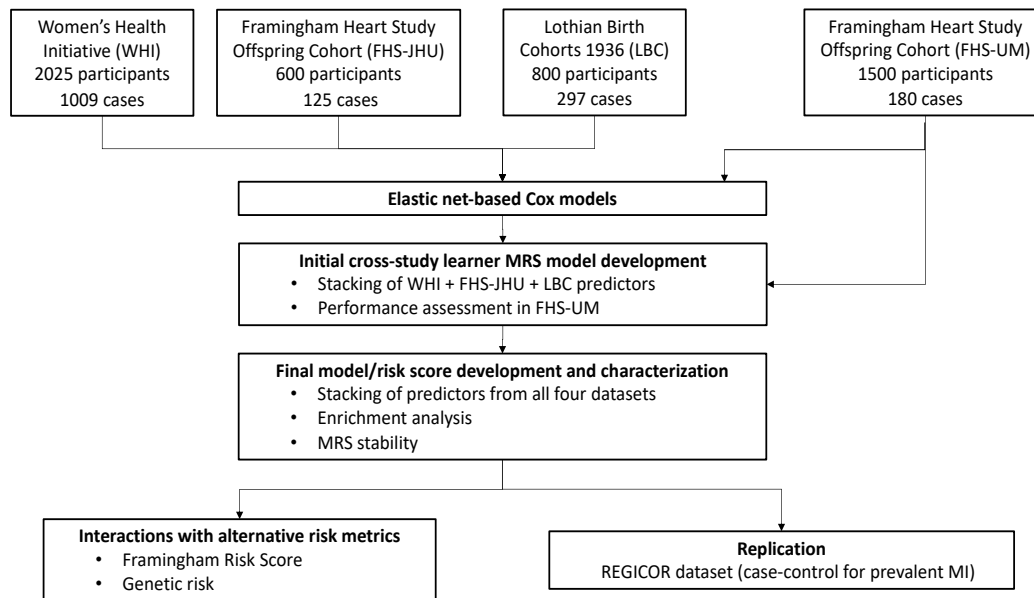


Figure 1: Computational workflow for MRS development and evaluation. The initial MRS was trained in three cohorts with FHS-UM held out to evaluate performance. The final MRS was then trained using all four datasets and examined for biological significance, before testing for prevalent MI discrimination in an independent cohort and assessment of interactions with genetic and traditional risk scores.

Assessment in held-out FHS subset

Table 2: MRS performance in held-out FHS subset

Model	HR per s.d. MRS	P-value
Unadjusted	1.62	7.5e-12
Basic	1.32	2.2e-04
Plus risk factors	1.29	2.2e-03
FRS only	1.40	5.9e-06

1 No covariates

2 Adjusted for age, sex, and estimated cell type fractions

3 Additionally adjusted for BMI, LDKL, HDL, SBP, diabetes status, and current smoking

4 Adjusted for Framingham Risk Score

Stacking of the three initial predictors resulted in model weights of 0.6, 0, and 0.4 for WHI, FHS-JHU, and LBC, respectively (i.e. FHS-JHU did not ultimately contribute to the initial model). The resulting ensemble predictor was evaluated using robust Cox proportional hazards models in FHS-UM, showing strong associations with incident CVD in an unadjusted model (HR=1.62, $p=6.03\text{e-}14$), which was attenuated partially through adjustment for standard covariates (age, sex, and estimated cell type fractions) as well as CVD risk factors (Table 2). These results were robust to sensitivity analyses excluding all individuals who experienced prior CVD events (Supp. Table S1).

Table 3: Comparison of Cox regression coefficient estimates between the CSL and combined model options.

Model	Hazard ratio per standard deviation MRS		
	Combined	ComBat	CSL
Unadjusted	1.51 [1.3-1.76]	1.75 [1.54-1.99]	1.62 [1.41-1.86]
Basic	1.21 [1.02-1.42]	1.39 [1.18-1.63]	1.32 [1.14-1.54]
Plus risk factors	1.12 [0.94-1.33]	1.28 [1.07-1.53]	1.29 [1.1-1.52]
FRS only	1.39 [1.2-1.62]	1.56 [1.36-1.78]	1.4 [1.21-1.62]

* Results are presented as: OR per s.d. [95% CI]

* Combined model naively combines all datasets while adjusting for study, whereas ComBat adjusts the data themselves to remove variation across studies while preserving class differences.

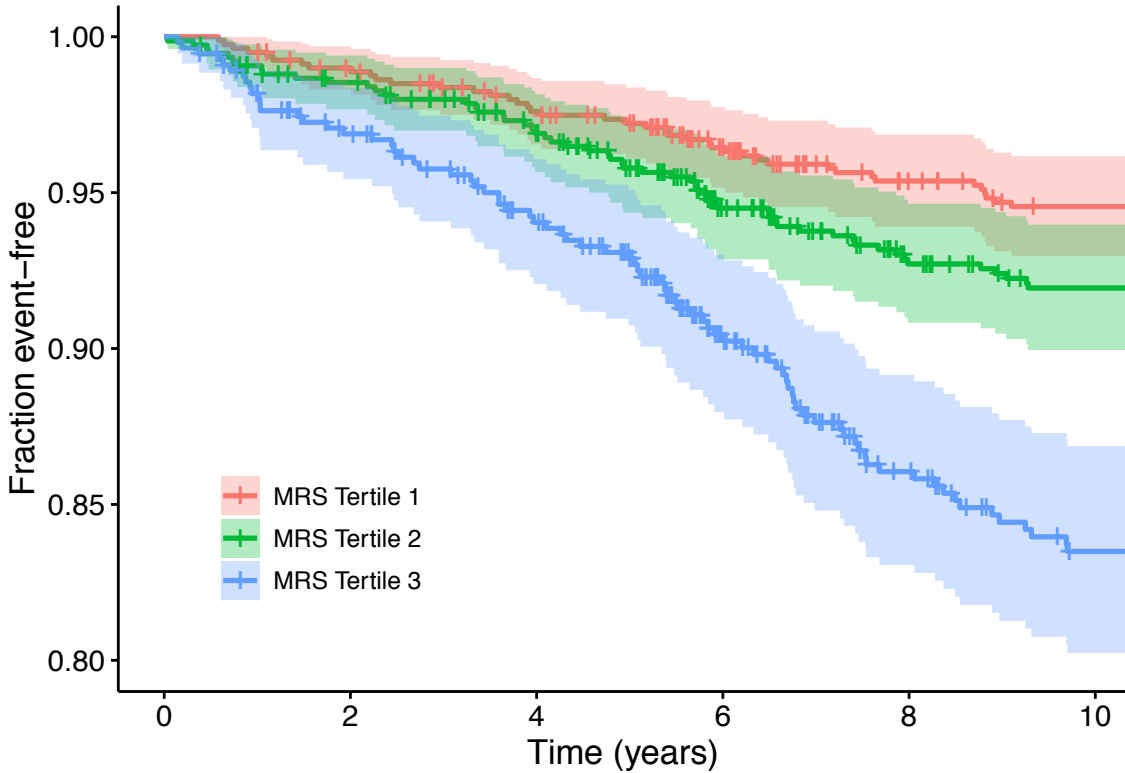


Figure 2: Kaplan-Meier survival curves in the held-out FHS-UM dataset. Individual curves correspond to tertiles of the initial (3-dataset) MRS. Vertical ticks correspond to censored observations, and colored bands represent 95% confidence intervals for tertile-specific survival curves.

When compared to the models based on naive combination of the datasets (“Combined”), the CSL model performed consistently better across all sets of covariates (Table 3). However, its performance was equivalent or worse than the single model trained on a ComBat-preprocessed dataset (note that FHS-UM was not included in the ComBat adjustment dataset so as not to allow the procedure to inject bias).

Final CSL model characterization

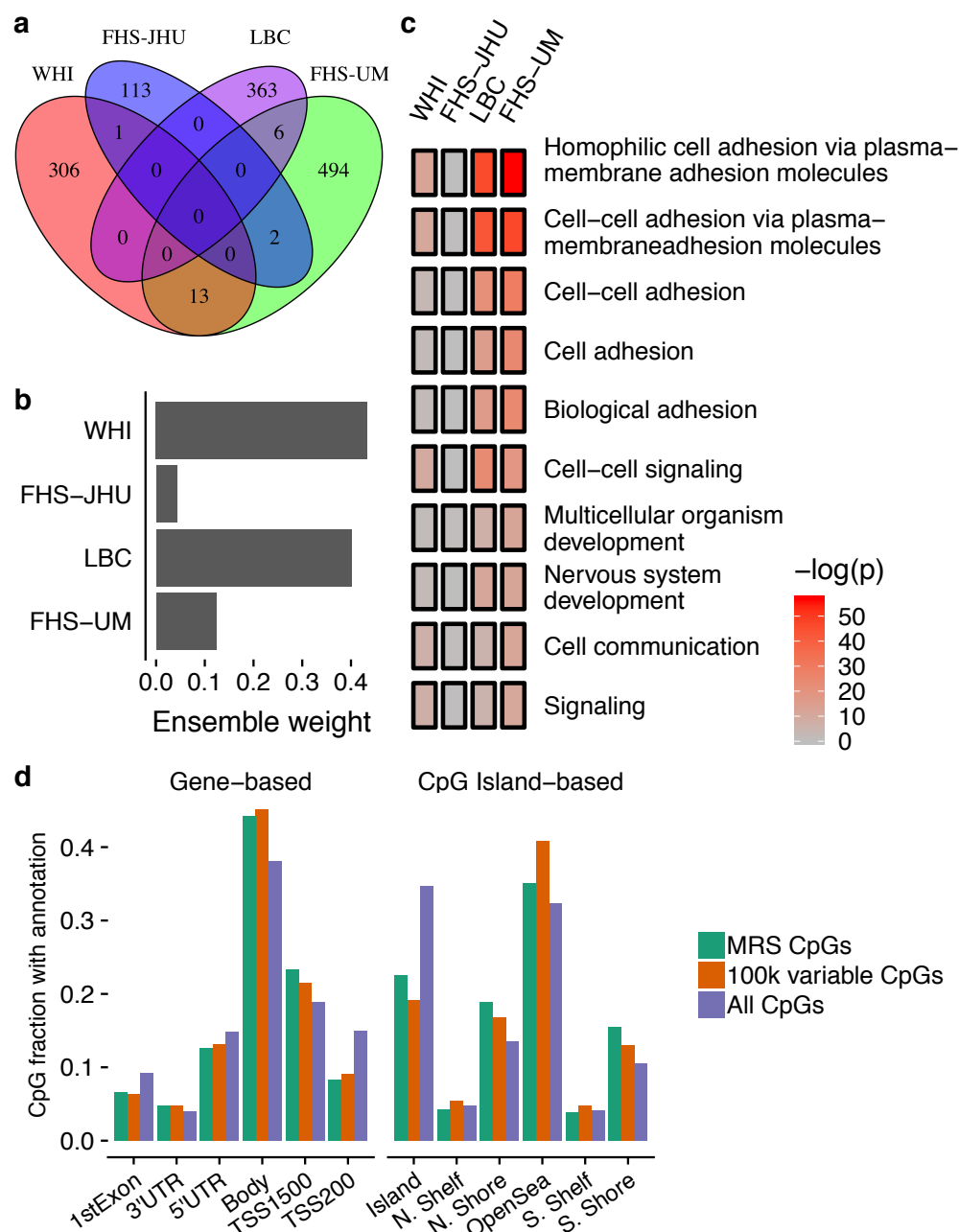


Figure 3: Characterization of the final CSL model. *a)* Overlap of CpG sites in the four individual predictors constituting the final model. *b)* Study-specific weights for constructing the ensemble model (derived from the “stacking” regression). *c)* Results from Gene Ontology-based enrichment analysis using genes annotated to SSL component CpGs. All GO terms with false discovery rate < 0.2 in any cohort are shown, and colored according to $-\log(p\text{-value})$ for enrichment in each SSL. *d)* Proportion of CpGs in the full set of CSL CpGs (union of CpG sets in each component SSL) compared to the 100,000 most variable CpGs (as used in SSL model development) and the full set of available CpGs. Groupings according to both gene-based and CpG island-based CpG annotations are shown.

The stacking regression in the final CSL model gave the most weight to WHI (0.43) and LBC (0.40), while retaining nonzero weights for FHS-JHU (0.04) and FHS-UM (0.12). There was very little overlap of specific CpG sites across cohort-specific models, with a maximum of 13 CpGs shared between two models (WHI and FHS-UM) and no CpGs shared between three or more models (Fig. 2a). Despite this lack of site-specific overlap, there was broad agreement for three of the four component SSL models at the level of enriched biological processes, with all except FHS-JHU enriched most strongly for proximity to genes involved in homophilic cell adhesion (Fig. 2b). MRS component CpGs tended to be found in similar genomic loci to the overall set of variable CpGs, which tended to be enriched in gene bodies and depleted in CpG islands compared to the full microarray CpG set. However, MRS CpGs did show a modest enrichment in and around CpG islands compared to the set of variable CpGs (Fig. 3d).

To gain more clarity as to potential biological mechanisms represented by the MRS, the HOMER tool was used to calculate enrichment of transcription factor (TF) binding motifs in the MRS component CpG sites (union of all individual SSL sites). Though no strong enrichments were found (all q-values >0.05), we note that two of the top ten hits relate to motifs found for circadian regulatory TFs (CLOCK and NPAS) in liver (full list of HOMER results with $q < 0.2$ in Supp. Table S2).

To better understand the stability of the risk score over time, intraclass correlation coefficients (ICCs) were calculated for two sets of grouped samples: 26 technical replicates from FHS and approximately 1000 longitudinal samples (across 3 visits, or about 6 years total) from LBC (Supp. Table S3). The technical replicates showed an ICC of 0.81, while the longitudinal samples showed an ICC of 0.68. As would be expected, the ICC for samples closer in time (Waves 1 & 2) were higher than that for samples more distant in time (Waves 1 & 3). Based on the observation of imperfect stability of the MRS over time as well as the partial attenuation its predictive power after adjustment for age, its component CpGs (the 1298-element union of all CpGs in any of the four individual SSL models) were examined for overlap with established epigenetic age metrics. While no enrichment was seen for the original cross-tissue DNAm age from Horvath (Horvath 2013), strong enrichment was seen for the morbidity-directed PhenoAge (Levine et al. 2018) (9 of 513 CpGs; $p=2.3e-5$) and especially the blood-specific aging marker from Hannum et al. (Hannum et al. 2013) (14 of 71 CpGs; $p=9.4e-23$). We note that these overlaps do not constitute a major fraction of either CpG set, but are nonetheless highly statistically significant.

Discrimination in myocardial infarction case-control

Table 4: Results from replication in REGICOR MI case-control

Model	ComBat	Combined	CSL
Unadjusted	1.71 [1.31-2.23]	1.66 [1.27-2.18]	1.85 [1.4-2.43]
Basic	2.11 [1.52-2.92]	1.75 [1.24-2.47]	2.23 [1.62-3.07]
Plus risk factors	1.8 [1.21-2.67]	1.36 [0.88-2.11]	1.65 [1.12-2.44]

* Results are presented as: OR per s.d. MRS [95% CI]

* Model covariates as in Table 2

* All models above are adjusted for two surrogate variable analysis (SVA) components

As one form of replication, the MRS was investigated for its discriminative performance in a nested case-control for prior myocardial infarction in the REGICOR cohort (cohort description in Supp. Table S4). Though this dataset did not contain incident events, its matching of sex and age allowed an evaluation free of potential confounding by these factors. The MRS was able to discriminate cases and controls in both unadjusted (odds ratio = 1.85, $p = 1.29\text{e-}5$) and, to a lesser degree, risk factor-adjusted models (odds ratio = 1.65, $p = 0.012$). The improved performance of the CSL model over the “Combined” MRS and its equivalent performance to the ComBat-based MRS was also confirmed in this dataset.

Interactions with alternate risk metrics

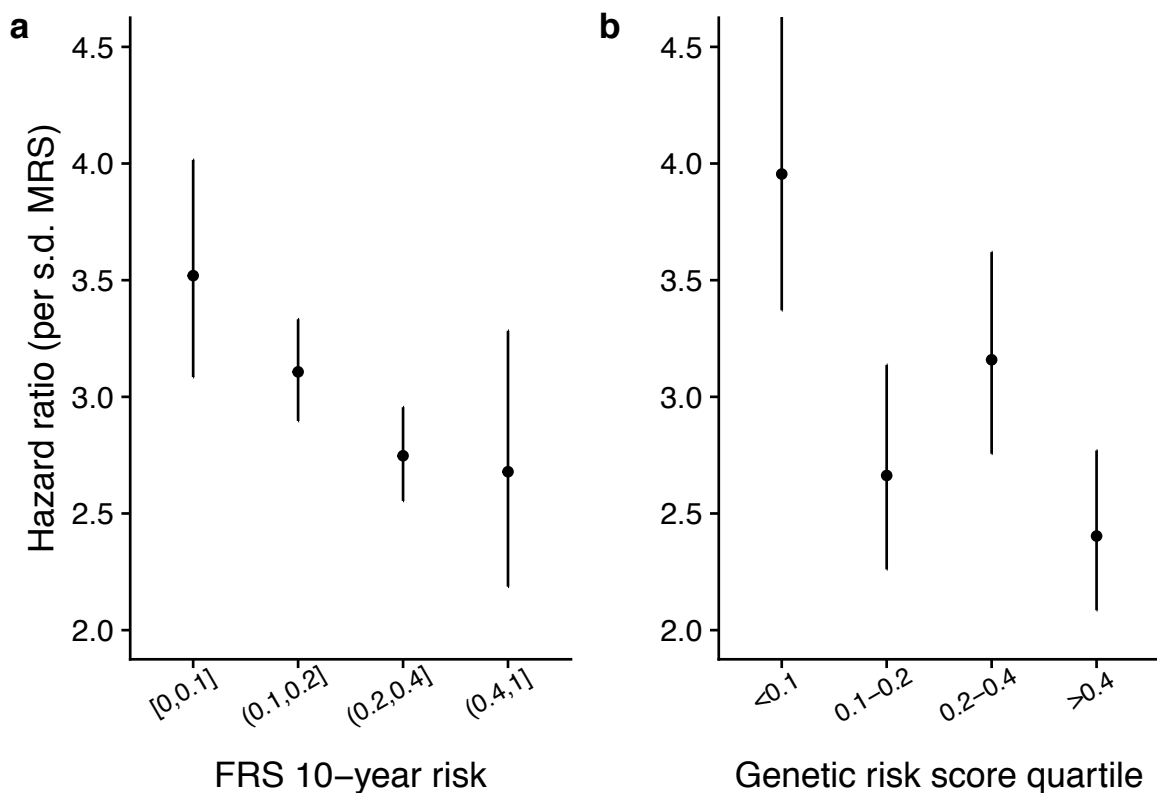


Figure 4: Interactions of MRS with other biomarkers of CVD risk. a) Hazard ratios for the MRS within subsets of 10-year generalized CVD risk according to the Framingham Risk Score. b) Hazard ratios for the MRS within quartiles of a genetic cardiovascular risk score (in white participants only). Hazard ratios are estimated using the final MRS, which was trained using each of these datasets. Stratum-specific Cox regressions were adjusted for age, sex, and estimated cell subtype fractions. Error bars represent standard errors for the hazard ratio estimates.

To understand how the present risk score interacts with other established CVD risk metrics, the performance of the MRS was re-evaluated after stratifying individuals by risk scores reflecting either demographic and biochemical features (Framingham Risk Score), or

genetic variants (based on Khera et al. 2018). First, the marginal effects of these risk scores were confirmed in each population. The Framingham Risk Score (FRS) was strongly predictive in WHI and FHS, while surprisingly showing no association with disease incidence in LBC (Supp. Table S5). As the genetic score does not change over time, it was evaluated with respect to any past or future CVD event, showing a moderate association in WHI but no association in FHS (Supp. Table S6).

In Cox models using baseline hazards stratified by study, it appeared that the MRS was more effective in those in lower “traditional” risk strata (according to the FRS; Fig. 4). As a sensitivity analysis, the cohorts were fully stratified into separate models, in which this pattern was visually clear in WHI, FHS-JHU, and to some extent FHS-UM (Supp. Fig. S1). The pattern did not appear in LBC, although we note that the Framingham Risk Score also did not show a “main effect” for predicting incident CVD in this cohort. A similar pattern appeared with respect to genetic risk in WHI (white participants only based on the formulation of the relevant risk score) and FHS, in which maximum MRS performance was achieved in the lowest alternative risk stratum. Supplementing these visual comparisons, combined Cox regressions across all cohorts (allowing for different baseline hazards across studies) showed a strong MRS-FRS interaction effect (12% reduction in HR for the MRS per 10% increase in FRS ($p = 0.000404$)), while that for the MRS-GRS interaction did not reach nominal statistical significance (0% reduction in HR for the MRS per standard deviation increase in GRS ($p = 0.853$)).

To explore the clinical potential of these interactions further, we returned to the initial MRS (trained in 3 datasets with FHS-UM held out). The FHS-UM dataset was filtered to include only participants with lower CVD risk based on either FRS ($<10\%$ estimated 10-year risk) or GRS (lowest quartile of the GRS). Within this lower-risk subset, participants in the upper MRS quartile had more than double the risk of the remainder of the participants: 10% (29/289) versus 4% (34/866). While a similar risk increase was observed when taking the intersection of these low-risk groups (low FRS and low GRS), the sample size was not enough to draw reasonable statistical conclusions (only 6 events in total).

FRS could not be calculated in the REGICOR dataset, as not all risk factors were available as continuous values. However, stratified models replicated the observation of greater MRS discrimination in lower alternative risk strata. An empirical risk function was generated through logistic regression of MI status on cardiovascular risk factors (age, sex, BMI, diabetes, smoking status, hyperlipidemia (binary), and hypertension). Predicted MI risk using this model was used to stratify subjects into four risk groups, with MRS odds ratios of 4.97 in the lowest-risk group versus 1.36 in the highest-risk group. More detailed results from these analyses are shown in Supp. Table S7.

3.1.4 Discussion

Epigenetic signatures of cardiometabolic diseases and aging in general are being actively explored as biomarkers of disease risk that are potentially modifiable and reveal underlying biological mechanisms. Here, in a novel application of a cross-study ensembling method, we introduce a DNA methylation-based score specific to cardiovascular disease risk. The model performs better than one trained on a naive combination of the entire dataset (though

similarly to use of a batch preprocessing algorithm), and may be most strongly predictive in individuals predicted to be at lower risk based on traditional risk factors.

We opted to use cross-study learning to train our risk model because we expect that differences across cohorts (e.g. demographic, behavioral) may contribute to heterogeneity in both the marginal distribution of the CpG features and the conditional distribution of the CVD outcome. Under these conditions, the generalizability of a single-study predictor is often obscured or overstated (Chang et al. 2015; Zhang et al. 2018). Using the CSL, we observed improved performance in test sets as compared to the naïve strategy of combining cohorts and training a single-study model. This suggests that up- and down-weighting single-study models offers an advantage compared to the increased combined sample size of the cohort-merging strategy.

Notably, the performance of the CSL model was similar to that of the model trained on the merged cohorts after batch adjustment via ComBat. This suggests that the assumptions regarding heterogeneity in the marginal distributions of the model covariates made by the ComBat hierarchical model are met. Any heterogeneity structure that can be captured by variation in the marginal effects can be accounted for and removed by ComBat (Johnson, Li, and Rabinovic 2007). In practice, this underlying structure is unknown, and we highlight that the CSL makes no specific assumptions and was able to produce similar gains in predictive accuracy in both the held-out FHS-UM and independent REGICOR datasets.

In assessing the stability of the MRS, we observed reasonable reproducibility between technical replicates (ICC=0.81). ICCs for LBC subjects over time were somewhat lower (ICC=0.68), which is to be expected due to not only changes in environment, but also the known epigenetic evolution with age that we observed to be enriched in the components of our score. These ICC values suggest an imperfect but usable reproducibility of the MRS, and an aggregate marker that is fairly robust considering the low replicability that has been observed for individual sites in technical replicates (general median ICC of 0.3 and mode of 0.75 in a “high reliability” cluster) (Bose et al. 2014).

The enrichment of the MRS component CpGs for proximity to genes related to cell-cell adhesion (in all subsets except FHS-JHU) suggests one of the biological patterns it tracks. As we have previously observed in a subset of these datasets (Westerman et al. 2018), it appears that immune activation is central to the prognostic information contained in leukocyte DNA methylation. For example, epigenetic processes have been shown to be involved in the activation and increased adhesion of monocytes in response to environmental insults and metabolic stress, though these have been explored primarily in relation to histone modifications (Short et al. 2017). Our results provide preliminary support for an attractive model in which a methylation-based score could act as a monitor of cumulative stress in leukocytes and their corresponding activation towards a more atherogenic state.

Existing epigenetic scores have shown varying strength in predicting incident cardiovascular disease. An early investigation examined blood-based methylation in LINE-1 elements, finding strong associations of global hypomethylation with prevalent and incident ischemic heart disease global (LINE-1), though additional reports showed opposite associations of methylation at repetitive elements with CVD (Kim et al. 2010).

Guarrerra et al. developed a biomarker for MI based on global LINE-1 and ZBTB12 gene methylation that provided a modest net reclassification index improvement (0.23-0.47) compared to traditional risk factors only. Multiple epigenetic aging metrics, though not developed specifically for CVD, have been shown to predict incident CHD, including PhenoAge (odds ratios from 1.02 to 1.08) and GrimAge (hazard ratio = 1.07, adjusted for age and technical factors) (Levine et al. 2018; Lu et al. 2019). While these associations are statistically significant, they do not represent clinically meaningful improvements in discrimination. Our observed hazard ratio of 1.32 (in the held-out FHS-UM dataset) suggests a biomarker that may be closer to clinical relevance. We note that our component CpG sites overlap strongly with those of these established epigenetic metrics including PhenoAge, suggesting that it captures some of the same biological patterns. However, the biological significance of the specific methylation changes observed in these aging- and morbidity-related metrics, whether as markers of failure in epigenetic regulation breakdown versus the work down by an “epigenetic maintenance system” is still unclear (Horvath 2013; Lund et al. 2019).

In examining the potential clinical utility of an epigenetic risk score for CVD, it is important to understand to what extent it is redundant or complementary to existing risk metrics. We saw a pattern of improved risk prediction in individuals whose cardiovascular risk based on traditional metrics (here, the Framingham Risk Score) was low. This pattern replicated in the REGICOR dataset (though FRS could not be directly calculated), with improved MRS discrimination in lower-risk subjects based on an empirical risk function. While these associations are preliminary, they suggest that an epigenetic risk score could help identify higher-risk individuals who otherwise would not have been detected by other metrics. We did not identify any robust patterns of differential MRS performance in strata based on a recent genetic cardiovascular risk score, meaning that individuals cannot be prioritized for measurement of this score purely based on germline genetic variants. There may have been lower power to detect any such patterns from the outset, since the GRS performed only modestly in WHI and had no discriminative power in FHS.

Multiple limitations should be acknowledged. While lymphocytes are known to be important in CVD pathogenesis, there is likely additional biological signal in other CVD-relevant tissues not examined here. Additionally, the present definition of CVD was chosen to balance specificity of CVD subtypes with sample size, but this balance could be altered to focus on more specific disease subtypes (e.g. myocardial infarction) or a broader definition of CVD (e.g. including heart failure).

In sum, we have developed an epigenetic risk score for cardiovascular disease that provides additional predictive power beyond existing risk measures, and may show improved performance in populations otherwise designated as low-risk. Furthermore, we have shown a novel application of a cross-cohort ensembling method that may provide significant value to future investigations in genomic epidemiology.

3.1.5 Methods

Study participants and phenotype collection

WHI methylation data came from a combined case-control and pseudo case-cohort sampling of 2129 women from the Women's Health Initiative study, a larger prospective cohort beginning in 1993 that included over 160,000 postmenopausal women from across the United States (Anderson et al. 1998). Included subjects had no self-reported CVD at baseline, and cases were chosen based on incident centrally adjudicated angina, revascularization, or CHD event during follow-up. Inclusion criteria for methylation measurement resulted in an oversampling of African American and Hispanic participants. Blood samples used for measurement of DNA methylation and clinical biochemistry were taken at Exam 1. Data are available in the dbGaP public repository (accession: phs000200.v11.p3; downloaded on September 27, 2017).

FHS methylation data came from a substudy of the Framingham Heart Study that measured DNA methylation in 2726 subjects from the Offspring Cohort. The Framingham Offspring Cohort was originally established in 1971 to follow 5209 children of the original Framingham Heart Study participants and their spouses (Kannel et al. 1979). Fasting blood samples for both methylation and clinical biochemistry were collected from participants at Exam 8, which took place from 2005-8. Blood samples were also provided for clinical biochemistry measurements in previous exams, constituting the "past exposures" examined here. Data are available in the dbGaP public repository (accession: phs000007.v29.p10; downloaded on September 27, 2017). Adjudicated cardiovascular event data was collected through 2015, and events were defined here as any of: myocardial infarction, angina pectoris, stroke (approximately 90% being ischemic), or death from CHD (Framingham event codes 1-29). FHS methylation data were collected in two primary batches in two centers – one in subjects from a nested case-control for CVD measured at Johns Hopkins University (FHS-JHU) (JoeHanes et al. 2013), and the other in a larger set of remaining Framingham Offspring participants measured at the University of Minnesota (FHS-UM).

Blood-based biochemical markers (total cholesterol, LDL, HDL, triglycerides, glucose, hsCRP, and systolic blood pressure) were log10-transformed for all analyses. In addition, median imputation was used to fill missing values for BMI (20 individuals in total), medication use, and smoking status (thus assuming no medication use and no smoking where these values were missing). Diabetes was defined as either use of diabetes medication or a measured fasting blood glucose level of >125 mg/dL. While directly available in WHI, pack-years of smoking was approximated in FHS by multiplying either the number of years since starting smoking by the current number of packs per day, or else by multiplying the number of years smoked by the median number of packs per day (15) for those who had quit. Framingham Risk Scores were calculated as described previously (D'Agostino et al. 2008).

The Lothian Birth Cohorts consist of two birth cohorts (born in 1921 and 1936) established in the Lothian region of Scotland (Deary et al. 2012). Only the 1936 cohort was analyzed here. Blood samples were collected in three waves starting in 2004, with our primary analyses here focusing on only those samples from Wave 1 (2004-2007). LBC data are

accessible through the European Genome-phenome Archive (accession: EGAS00001000910).

The REGICOR dataset analyzed here consisted of a nested case-control for myocardial infarction within the larger REGICOR (REgistre Gironí del COR) cohort from the Girona Province in Catalonia (Spain). Whole blood samples were collected from 391 total participants, with those from cases generally collected within 24 hours of the event. Characteristics for this population are available in Supp. Table S4.

DNA methylation data processing

DNA methylation data for all initial cohorts (WHI, FHS, and LBC) were collected using the Illumina HumanMethylation450 microarray platform (Bibikova et al. 2011) and downloaded as raw intensity files. Preprocessing was performed using the *minfi* and *wateRmelon* packages for R (Aryee et al. 2014; Pidsley et al. 2013). Sample-wise filters were as follows: robust overall signal in the main cluster based on visual inspection of an intensity plot, less than 10% of probes undetected at a detection threshold of $p < 1e-16$, and a reported sex matching methylation-based sex prediction. Probes were removed using the following criteria: more than 10% of samples undetected at a detection threshold of $p < 1e-16$, location in the X or Y chromosomes, non-CpG probes, cross-hybridizing probes, probes measuring SNPs, and probes with an annotated SNP at the CpG site or in the single-base extension region. Samples were normalized using the Noob method for background correction and dye-bias normalization, followed by the BMIQ method for probe type correction (Fortin, Triche, and Hansen 2016; Teschendorff et al. 2013). Blood cell fractions for 6 blood cell types (CD4+ T-cells, CD8+ T-cells, B-cells, natural killer cells, monocytes, and granulocytes) were estimated using a common reference-based method (Houseman et al. 2012), and 5 of these (excluding granulocytes) were included in cell count-adjusted statistical models. After quality control and filtering steps, 390597 CpG sites were shared between the 3 datasets, formatted as beta values (ratio of methylated signal to total microarray signal).

DNA methylation data for the REGICOR cohort were collected using the Illumina MethylationEPIC microarray platform (Pidsley et al. 2016) and analyzed using the *wateRmelon* (Pidsley et al. 2013) and *methyumi* (Davis et al. 2019) R packages. Samples were excluded based on detection p -value > 0.05 in at least 1% of probes or failure to cluster in the appropriate sex based on X chromosome methylation. Probes were excluded based on detection p -value > 0.05 in at least 1% of samples, a bead count < 3 in at least 5% of samples, discarding by Illumina based on underperformance ($n=1,031$) or changes in the manufacturing process ($n=977$), non-CpG targets, and cross-hybridization ($n=43,979$). A batch normalization was performed by standardizing beta values to mean zero and unit variance within each bisulfite conversion batch prior to analysis. After quality control and preprocessing, 811,610 CpG sites across 391 individuals were available for analysis. Participants were further excluded from analysis due to unknown smoking habits ($n=10$) and unavailable information regarding diabetes, hypertension, or hyperlipidemia ($n=53$). Surrogate variable analysis (Leek and Storey 2007) was used to calculate two surrogate variables, representing potential technical and biological confounders, for adjustment in MRS replication models.

CVD risk prediction modeling

Study-specific CVD risk prediction models were trained using penalized Cox proportional hazards regressions with the elastic net penalty. CVD events were defined as above, and times were right-censored based on the most recent exam available in each cohort. The elastic alpha parameter was set at 0.05, based on prior observations of good performance on Illumina methylation microarray datasets (Zhuang, Widschwendter, and Teschendorff 2012), and the penalty parameter λ was optimized through 5-fold cross-validation. For each model, only the most variable 100,000 CpGs according to median absolute deviation (~25% of all available sites shared across platforms) were included in order to decrease the computational burden and ensure that the selected CpGs would have meaningful interindividual variation.

The cross-study learner (CSL) was constructed as an ensemble of study-specific regression models. Predictions from each single-study learner (SSL) were combined using the “stacking” approach (Patil and Parmigiani 2018), implemented as follows. First, predictions from each SSL to both itself and the other training datasets were combined into a design matrix (with dimensions $N_{total} \times \# \text{ SSLs}$). This formed the input to an additional penalized Cox regression (ridge regression with λ optimized through 5-fold CV and coefficients restricted to be non-negative) of all training studies at once. Coefficients from this regression, corresponding to input study-specific SSLs, were normalized to sum to one to produce the CSL weights. For prediction in new datasets, SSL predictions were each standardized to mean zero and unit variance before calculating their weighted sum (using the “stacking” weights) as the final CSL score.

A series of approaches for combining information across cohorts were tested as alternatives to the CSL. The naive “combined” approach consisted of simply aggregating observations from all training sets into a single dataset and training an elastic net regression as described above while adjusting for study as a fixed effect. The ComBat method trained across all studies as with the “combined” approach, but included an empirical Bayes-based preprocessing step to directly adjust the dataset for study differences while preserving variation along the “axis” of incident CVD events (Johnson, Li, and Rabinovic 2007).

MRS evaluation in FHS was performed using Cox proportional hazards models, with a series of models adjusting for covariates including demographics, anthropometrics, biochemical values, and cell subtype estimates. Robust standard errors (also known as the Huber-White sandwich estimator) were used to account for family structure as has been suggested for clustered data (Rogers 1993) and used for epigenetic risk models in FHS (Lu et al. 2019). MRS evaluation in the REGICOR case-control used logistic regression models, adjusting for the same sets of covariates where possible, though traditional biochemical risk factors were only available in discrete low vs. high categories.

Genomic risk score calculation

Imputed genotype data were retrieved from dbGaP (accession: phs000746.v2.p3 (WHI) and phs000342.v18.p11 (FHS)). Variants were filtered for imputation R-squared > 0.3 , and annotated with rsIDs, loci, and allelic information using the 1000 Genomes Phase 3 download from dbSNP (download date: April 13, 2018). Weights for the genetic risk score

calculation were based on the genome-wide CVD score developed by Khera et al (Khera et al. 2018). We note that these scores were developed only for populations of European descent, and thus are not optimized for the mixed-ancestry WHI population. GRS were then calculated as the weighted sum of allelic dosages, normalized by the number of relevant SNPs available. Genotype data processing and GRS calculation were performed using PLINK 2.0.

Risk score interaction analysis

Interaction analysis was performed using similar Cox regression models to those above, adjusting for the “basic” set of covariates (age, sex, and estimated blood cell type fractions) and using robust standard error estimates. To facilitate visual comparisons, main-effect regressions for the MRS were fitted within risk strata defined by the FRS or GRS separately in each dataset. To obtain overall interaction effect estimates, an interaction between MRS and either FRS or GRS was introduced into a combined regression including all datasets, while allowing stratified baseline hazards (strata() argument to the coxph function). We note that all interaction analyses were performed using the final MRS in the same group of datasets used to train the MRS model, meaning that the main-effect estimates were biased upwards. All regressions assessing the GRS excluded non-white participants, based on the fact that the polygenic CVD score was developed in individuals of European ancestry (Khera et al. 2018).

For quasi-replication of these associations in the REGICOR dataset, stratified logistic regressions were used to discriminate MI cases from controls using the MRS, while adjusting for estimated cell count fractions as well as two SVA components (as in the main REGICOR models). In the absence of continuous values for blood pressure and lipids, an empirical risk function was generated by first performing a logistic regression on the following cardiovascular risk factors: age, sex, estimated cell count fractions, BMI, diabetes, smoking status, hyperlipidemia (binary), and hypertension (binary), along with two SVA components. Predicted risks based on this model were then used to stratify subjects into four risk groups by evenly splitting the range of predicted risks into four segments (thus resulting in strata based on raw risk, rather than percentiles).

3.1.6 Supplementary Data

Supp. Table S1: MRS performance in held-out FHS subset without past CVD events

Model	HR per s.d. MRS	P-value
Unadjusted ¹	1.60	4.3e-10
Basic ²	1.32	8.2e-04
Plus risk factors ³	1.31	5.0e-03
FRS only ⁴	1.41	1.8e-05

¹No covariates

²Adjusted for age, sex, and estimated cell type fractions

³Additionally adjusted for BMI, HDL, SBP, diabetes status, and current smoking

⁴Adjust for Framingham Risk Score only

Supp. Table S2: MRS enrichment for transcription factor binding motifs: HOMER results

Motif	Consensus	BH q-value	MRS Coverage	Background Coverage
Foxf1(Forkhead)/Lung-Foxf1-ChIP-Seq(GSE77951)/Homer	WWATRTAAACAN	0.169	6.41%	4.27%
Pit1(Homeobox)/GCrat-Pit1-ChIP-Seq(GSE58009)/Homer	ATGMATATDC	0.169	6.16%	4.11%
HEB(bHLH)/mES-Heb-ChIP-Seq(GSE53233)/Homer	VCAGCTGBNN	0.169	28.69%	24.53%
Smad4(MAD)/ESC-SMAD4-ChIP-Seq(GSE29422)/Homer	VBSYGTCTGG	0.169	20.34%	16.75%
c-Myc(bHLH)/LNCAP-cMyc-ChIP-Seq(Unpublished)/Homer	VCCACGTG	0.169	8.44%	6.13%
Ascl1(bHLH)/NeuralTubes-Ascl1-ChIP-Seq(GSE55840)/Homer	NNVVCAGCTGBN	0.169	22.87%	19.36%
CLOCK(bHLH)/Liver-Clock-ChIP-Seq(GSE39860)/Homer	GHCACGTG	0.169	7.76%	5.70%
E2A(bHLH)/proBcell-E2A-ChIP-Seq(GSE21978)/Homer	DNRCAGCTGY	0.169	22.53%	19.17%
NPAS(bHLH)/Liver-NPAS-ChIP-Seq(GSE39860)/Homer	NVCACGTG	0.169	17.22%	14.28%
bHLHE40(bHLH)/HepG2-BHLHE40-ChIP-Seq(GSE31477)/Homer	KCACGTGMCN	0.169	5.06%	3.49%
Max(bHLH)/K562-Max-ChIP-Seq(GSE31477)/Homer	RCCACGTGGYYN	0.169	8.52%	6.48%
Ascl2(bHLH)/ESC-Ascl2-ChIP-Seq(GSE97712)/Homer	SSRGCAGCTGCH	0.169	17.72%	14.87%
Lhx2(Homeobox)/HFSC-Lhx2-ChIP-Seq(GSE48068)/Homer	TAATTAGN	0.169	7.00%	5.20%
FOXK1(Forkhead)/HEK293-FOXK1-ChIP-Seq(GSE51673)/Homer	NVWTGTTTAC	0.175	7.43%	5.63%
bHLHE41(bHLH)/proB-Bhlhe41-ChIP-Seq(GSE93764)/Homer	KCACGTGMCN	0.197	15.61%	13.11%

Supp. Table S3: MRS stability as evaluated by using multiple within-subject measurements. Generic ICC heuristics for reference: 0-0.5 = poor, 0.5-0.75 = moderate, 0.75 - 0.9 = good, 0.9-1 = excellent.

Cohort	Group_type	# of pairs/groups	ICC
FHS	Duplicates	26	0.81
LBC36	Samples over multiple visits	758	0.68
LBC36	Samples over subsequent visits (Wave 1 & 2)	758	0.68
LBC36	Samples over longer time frame (Wave 1 & 3)	758	0.64

Supp. Table S4: Description of REGICOR myocardial infarction nested case-control population

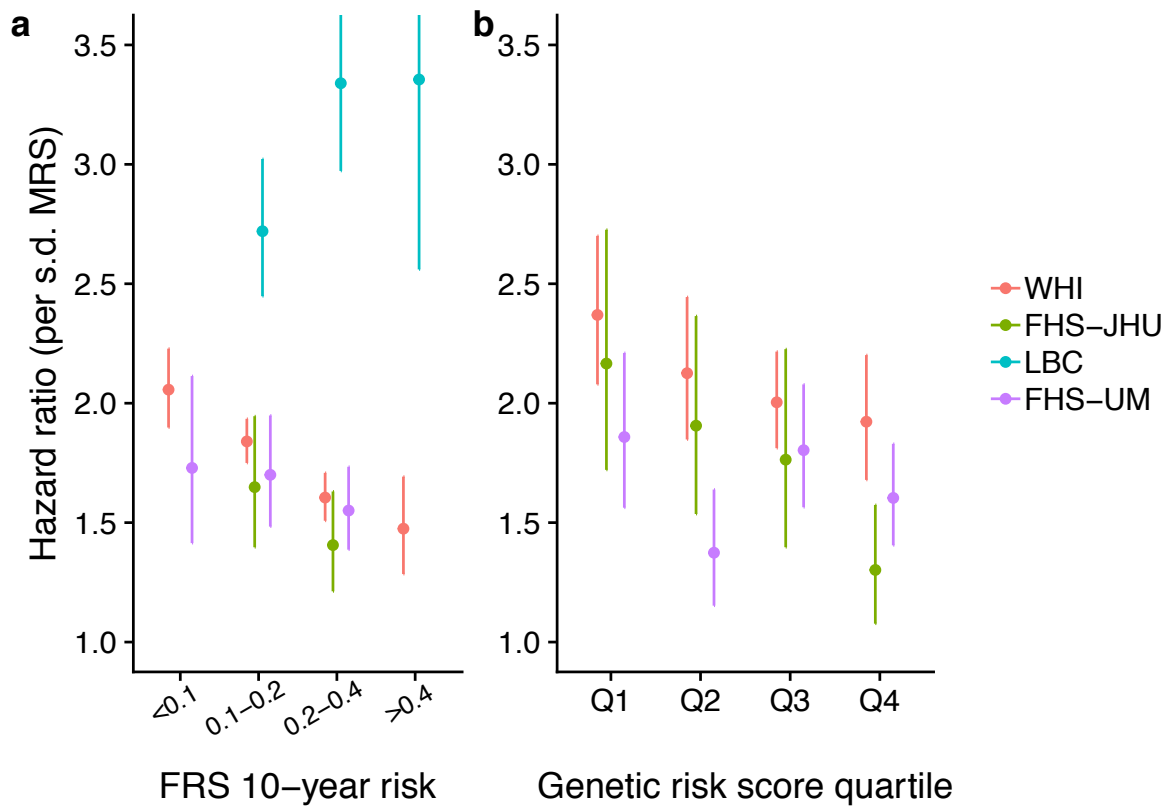
Sample size	391
Prior myocardial infarction	50.1%
Ancestry (% European)	100%
Age	63.2 (6.9)
Sex (% female)	48.6
Smoking	21.5%
Body mass index	28.5 (4.8)
LDL cholesterol	127 (26)
HDL cholesterol	50.0 (10.5)
Systolic blood pressure	135 (18)
Diabetes prevalence	24.7%

Supp. Table S5: Validation of Framingham Risk Score

Study	HR_per_SD	p
WHI	1.50	4.7e-61
FHS-JHU	1.42	4.8e-06
FHS-UM	1.62	3.7e-21
LBC	0.97	6.6e-01

Supp. Table S6: Validation of genetic risk score

Study	OR_per_SD	p
WHI	1.28	0.0000011
FHS-JHU	1.09	0.3779815
FHS-UM	1.04	0.5575529



Supp. Fig. S1: Interactions of MRS with other biomarkers of CVD risk. *a*) Hazard ratios for the MRS within subsets of 10-year generalized CVD risk according to the Framingham Risk Score. *b*) Hazard ratios for the MRS within quartiles of a genetic cardiovascular risk score (in white participants only for WHI). Hazard ratios are estimated using the final MRS, which was trained using each of these datasets. Stratum-specific Cox regressions were adjusted for age, sex, and estimated cell subtype fractions. Estimates for strata with less than 25 incident events are not shown. Error bars represent standard errors for the hazard ratio estimates (cut off above in panel (a) for ease of visualization of other points).

Supp. Table S7: Risk factor-stratified MRS performance in the REGICOR dataset

Risk factor group	OR per std. dev.	N
G1	4.97 [1.71-14.48]	119
G2	1.12 [0.62-2.01]	90
G3	2.92 [1.07-7.96]	60
G4	1.36 [0.31-5.98]	55

3.1.7 References

- Anderson, Garnet L, Steven R Cummings, Laurence S Freedman, Curt Furberg, Maureen M Henderson, Susan R Johnson, Lewis H Kuller, et al. 1998. "Design of the Women's Health Initiative Clinical Trial and Observational Study." *Controlled Clinical Trials* 19 (1): 61–109. [https://doi.org/10.1016/S0197-2456\(97\)00078-0](https://doi.org/10.1016/S0197-2456(97)00078-0).
- Aryee, Martin J., Andrew E. Jaffe, Hector Corrada-Bravo, Christine Ladd-Acosta, Andrew P. Feinberg, Kasper D. Hansen, and Rafael A. Irizarry. 2014. "Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays." *Bioinformatics* 30 (10). Oxford University Press: 1363–9. <https://doi.org/10.1093/bioinformatics/btu049>.
- Aslibekyan, Stella, Golareh Agha, Elena Colicino, Anh N. Do, Jari Lahti, Symen Ligthart, Riccardo E. Marioni, et al. 2018. "Association of Methylation Signals With Incident Coronary Heart Disease in an Epigenome-Wide Assessment of Circulating Tumor Necrosis Factor α ." *JAMA Cardiology* 3 (6): 463–72. <https://doi.org/10.1001/JAMACARDIO.2018.0510>.
- Baccarelli, A, R Wright, V Bollati, A Litonjua, A Zanobetti, L Tarantini, D Sparrow, P Vokonas, and J Schwartz. 2010. "Ischemic heart disease and stroke in relation to blood DNA methylation." *Epidemiology* 21 (6): 819–28. <https://doi.org/10.1097/EDE.0b013e3181f20457>.
- Bacos, Karl, Linn Gillberg, Petr Volkov, Anders H Olsson, Torben Hansen, Oluf Pedersen, Anette Prior Gjesing, et al. 2016. "Blood-based biomarkers of age-associated epigenetic changes in human islets associate with insulin secretion and diabetes." *Nature Communications* 7. Nature Publishing Group: 11089. <https://doi.org/10.1038/ncomms11089>.
- Bibikova, Marina, Bret Barnes, Chan Tsan, Vincent Ho, Brandy Klotzle, Jennie M. Le, David Delano, et al. 2011. "High density DNA methylation array with single CpG site resolution." *Genomics* 98 (4): 288–95. <https://doi.org/10.1016/j.ygeno.2011.07.007>.
- Bonder, Marc Jan, René Luijk, Daria V Zhernakova, Matthijs Moed, Patrick Deelen, Martijn Vermaat, Maarten van IJterson, et al. 2016. "Disease variants alter transcription factor levels and methylation of their binding sites." *Nature Genetics* 49 (1): 131–38. <https://doi.org/10.1038/ng.3721>.
- Bose, Maitreyee, Chong Wu, James S. Pankow, Ellen W. Demerath, Jan Bressler, Myriam Fornage, Megan L. Grove, et al. 2014. "Evaluation of microarray-based DNA methylation measurement using technical replicates: The atherosclerosis risk in communities (ARIC) study." *BMC Bioinformatics*. <https://doi.org/10.1186/1471-2105-15-312>.
- Chang, Christopher C., Carson C. Chow, Laurent C.A.M. Tellier, Shashaank Vattikuti, Shaun M. Purcell, and James J. Lee. 2015. "Second-generation PLINK: Rising to the challenge of larger and richer datasets." *GigaScience*. <https://doi.org/10.1186/s13742-015-0047-8>.
- D'Agostino, Ralph B., Ramachandran S. Vasan, Michael J. Pencina, Philip A. Wolf, Mark Cobain, Joseph M. Massaro, and William B. Kannel. 2008. "General cardiovascular risk profile for use in primary care: The Framingham heart study." *Circulation* 117 (6): 743–53. <https://doi.org/10.1161/CIRCULATIONAHA.107.699579>.
- Davis, Sean, Pan Du, Sven Bilke, Timothy Triche, and Oiz Bootwalla. 2019. "methylumi: Handle Illumina methylation data." <https://doi.org/10.1016/j.neuroimage.2007.06.039>.

Deary, Ian J., Alan J. Gow, Alison Pattie, and John M. Starr. 2012. "Cohort Profile: The Lothian Birth Cohorts of 1921 and 1936." *International Journal of Epidemiology* 41 (6): 1576–84. <https://doi.org/10.1093/ije/dyr197>.

Fernández-Sanlés, Alba, Sergi Sayols-Baixeras, Isaac Subirana, Irene R Degano, and Roberto Elosua. 2017. "Association between DNA methylation and coronary heart disease or other atherosclerotic events: A systematic review." <https://doi.org/10.1016/j.atherosclerosis.2017.05.022>.

Fortin, Jean-Philippe, Timothy J Triche, and Kasper D Hansen. 2016. "Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi." *Bioinformatics* 33 (4). Oxford University Press: btw691. <https://doi.org/10.1093/bioinformatics/btw691>.

Goh, Wilson Wen Bin, Wei Wang, and Limsoon Wong. 2017. "Why Batch Effects Matter in Omics Data, and How to Avoid Them." <https://doi.org/10.1016/j.tibtech.2017.02.012>.

Guarrera, Simonetta, Giovanni Fiorito, N Charlotte Onland-Moret, Alessia Russo, Claudia Agnoli, Alessandra Allione, Cornelia Di Gaetano, et al. 2015. "Gene-specific DNA methylation profiles and LINE-1 hypomethylation are associated with myocardial infarction risk." *Clinical Epigenetics* 7 (1): 133. <https://doi.org/10.1186/s13148-015-0164-3>.

Hannum, Gregory, Justin Guinney, Ling Zhao, Li Zhang, Guy Hughes, Srinivas Sadda, Brandy Klotzle, et al. 2013. "Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates." *Molecular Cell* 49 (2): 359–67. <https://doi.org/10.1016/j.molcel.2012.10.016>.

Hao, Xiaoke, Huiyan Luo, Michal Krawczyk, Wei Wei, Wenqiu Wang, Juan Wang, Ken Flagg, et al. 2017. "DNA methylation markers for diagnosis and prognosis of common cancers." *Proceedings of the National Academy of Sciences*. <https://doi.org/10.1073/pnas.1703577114>.

Hedman, Åsa K., Michael M. Mendelson, Riccardo E. Marioni, Stefan Gustafsson, Roby Joehanes, Marguerite R. Irvin, Degui Zhi, et al. 2017. "Epigenetic Patterns in Blood Associated With Lipid Traits Predict Incident Coronary Heart Disease Events and Are Enriched for Results From Genome-Wide Association Studies." *Circulation: Cardiovascular Genetics* 10 (1): e001487. <https://doi.org/10.1161/CIRCGENETICS.116.001487>.

Horvath, Steve. 2013. "DNA methylation age of human tissues and cell types." *Genome Biology* 14 (10): R115. <https://doi.org/10.1186/gb-2013-14-10-r115>.

Houseman, Eugene Andres, William P Accomando, Devin C Koestler, Brock C Christensen, Carmen J Marsit, Heather H Nelson, John K Wiencke, and Karl T Kelsey. 2012. "DNA methylation arrays as surrogate measures of cell mixture distribution." *BMC Bioinformatics* 13 (1): 86. <https://doi.org/10.1186/1471-2105-13-86>.

Joehanes, Roby, Saixia Ying, Tianxiao Huan, Andrew D. Johnson, Nalini Raghavachari, Richard Wang, Poching Liu, et al. 2013. "Gene Expression Signatures of Coronary Heart Disease." *Arteriosclerosis, Thrombosis, and Vascular Biology* 33 (6): 1418–26. <https://doi.org/10.1161/ATVBAHA.112.301169>.

Johnson, W. Evan, Cheng Li, and Ariel Rabinovic. 2007. "Adjusting batch effects in microarray expression data using empirical Bayes methods." *Biostatistics* 8 (1). Oxford University Press: 118–27. <https://doi.org/10.1093/biostatistics/kxj037>.

Kannel, William B, Manning Feinleib, Patricia M. Mcnamara, Robert J Garrison, and William P Castelli. 1979. "An investigation of coronary heart disease in families: The framingham offspring study." *American Journal of Epidemiology* 110 (3): 281–90. <https://doi.org/10.1093/oxfordjournals.aje.a112813>.

Khera, Amit V., Mark Chaffin, Krishna G. Aragam, Mary E. Haas, Carolina Roselli, Seung Hoan Choi, Pradeep Natarajan, et al. 2018. "Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations." <https://doi.org/10.1038/s41588-018-0183-z>.

Kim, Myungjin, Tiffany I. Long, Kazuko Arakawa, Renwei Wang, Mimi C. Yu, and Peter W. Laird. 2010. "DNA Methylation as a Biomarker for Cardiovascular Disease Risk." Edited by Joel S. Bader. *PLoS ONE* 5 (3): e9692. <https://doi.org/10.1371/journal.pone.0009692>.

Leek, Jeffrey T., and John D. Storey. 2007. "Capturing heterogeneity in gene expression studies by surrogate variable analysis." *PLoS Genetics*. <https://doi.org/10.1371/journal.pgen.0030161>.

Levine, Morgan E., Ake T. Lu, Austin Quach, Brian H. Chen, Themistocles L. Assimes, Stefania Bandinelli, Lifang Hou, et al. 2018. "An epigenetic biomarker of aging for lifespan and healthspan." *Aging* 10 (4): 573–91. <https://doi.org/10.18632/aging.101414>.

Lu, Ake T., Austin Quach, James G. Wilson, Alex P. Reiner, Abraham Aviv, Kenneth Raj, Lifang Hou, et al. 2019. "DNA methylation GrimAge strongly predicts lifespan and healthspan." *Aging* 11 (2): 303–27. <https://doi.org/10.18632/aging.101684>.

Lund, Jesper Beltoft, Shuxia Li, Jan Baumbach, Anne Marie Svane, Jacob Hjelmberg, Lene Christiansen, Kaare Christensen, et al. 2019. "DNA methylome profiling of all-cause mortality in comparison with age-associated methylation patterns." *Clinical Epigenetics* 11 (1). BioMed Central: 23. <https://doi.org/10.1186/s13148-019-0622-4>.

Patil, Prasad, and Giovanni Parmigiani. 2018. "Training replicable predictors in multiple studies." *Proceedings of the National Academy of Sciences*. <https://doi.org/10.1073/pnas.1708283115>.

Pidsley, Ruth, Chloe C Y Wong, Manuela Volta, Katie Lunnon, Jonathan Mill, and Leonard C Schalkwyk. 2013. "A data-driven approach to preprocessing Illumina 450K methylation array data." *BMC Genomics* 14 (1). BioMed Central: 293. <https://doi.org/10.1186/1471-2164-14-293>.

Pidsley, Ruth, Elena Zotenko, Timothy J Peters, Mitchell G Lawrence, Gail P Risbridger, Peter Molloy, Susan Van Djik, et al. 2016. "Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling." *Genome Biology* 17 (1): 208. <https://doi.org/10.1186/s13059-016-1066-1>.

Richardson, Tom G, Jie Zheng, George Davey Smith, Nicholas J Timpson, Tom R Gaunt, Caroline L Relton, and Gibran Hemani. 2017. "Mendelian Randomization Analysis Identifies CpG Sites as Putative Mediators for Genetic Influences on Cardiovascular Disease Risk." *American Journal of Human Genetics* 101 (4): 590–602. <https://doi.org/10.1016/j.ajhg.2017.09.003>.

Rogers, W. H. 1993. "Regression standard errors in clustered samples." *Stata Technical Bulletin* 13: 19–23.

Short, John D., Sina Tavakoli, Huynh Nga Nguyen, Ana Carrera, Chelbee Farnen, Laura A. Cox, and Reto Asmis. 2017. "Dyslipidemic Diet-Induced Monocyte 'Priming' and Dysfunction in Non-Human Primates Is Triggered by Elevated Plasma Cholesterol and Accompanied by Altered

Histone Acetylation.” *Frontiers in Immunology* 8 (August). Frontiers: 958. <https://doi.org/10.3389/fimmu.2017.00958>.

Teschendorff, Andrew E, Francesco Marabita, Matthias Lechner, Thomas Bartlett, Jesper Tegner, David Gomez-Cabrero, and Stephan Beck. 2013. “A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data.” *Bioinformatics* 29 (2): 189–96. <https://doi.org/10.1093/bioinformatics/bts680>.

Tobi, Elmar W., Roderick C. Sliker, René Luijk, Koen F. Dekkers, Aryeh D. Stein, Kate M. Xu, P. Eline Slagboom, et al. 2018. “DNA methylation as a mediator of the association between prenatal adversity and risk factors for metabolic disease in adulthood.” *Science Advances*. <https://doi.org/10.1126/sciadv.aao4364>.

Wahl, Simone, Alexander Drong, Benjamin Lehne, Marie Loh, William R. Scott, Sonja Kunze, Pei-Chien Tsai, et al. 2017. “Epigenome-wide association study of body mass index and the adverse outcomes of adiposity.” *Nature* 541 (7635): 81–86. <https://doi.org/10.1038/nature20784>.

Westerman, Kenneth, Paola Sebastiani, Paul Jacques, Simin Liu, Dawn DeMeo, and José M Ordovás. 2018. “DNA methylation modules associate with incident cardiovascular disease and cumulative risk factor exposure.” *bioRxiv*, January, 471722. <https://doi.org/10.1101/471722>.

Zhang, Yuqing, Christoph Bernau, Giovanni Parmigiani, and Levi Waldron. 2018. “The impact of different sources of heterogeneity on loss of accuracy from genomic prediction models.” *Biostatistics*, September. <https://doi.org/10.1093/biostatistics/kxy044>.

Zhuang, Joanna, Martin Widschwendter, and Andrew E. Teschendorff. 2012. “A comparison of feature selection and classification methods in DNA methylation studies using the Illumina Infinium platform.” *BMC Bioinformatics*. <https://doi.org/10.1186/1471-2105-13-59>.

Chapter 4

Genomic and Epigenomic Predictors of Diet Response

4.1 A gene-diet interaction-based score predicts response to dietary fat in the Women's Health Initiative

In preparation for submission to the American Journal of Clinical Nutrition

4.1.1 Abstract

While diet response prediction for cardiometabolic risk factors (CRFs) has been demonstrated using single SNPs and main-effect genetic risk scores, little investigation has gone into the development of genome-wide diet response scores. Here, we leveraged the multi-study setup of the Women's Health Initiative cohort to generate and test genetic scores for response to dietary fat. A genome-wide interaction study was undertaken for each of six CRFs in women ($n \sim 10000$) not participating in the Dietary Modification (DM) trial, which focused on the reduction of dietary fat. Genetic scores based on these analyses were developed and tested for the prediction of one-year CRF changes in DM trial participants ($n \sim 5000$). Only one of these genetic scores, for LDL-cholesterol (LDL-C), predicted changes in the associated CRF. This 1760-variant score explained 3.4% of the variance in one-year LDL-C changes in the intervention arm, but was unassociated with changes in the control arm. In contrast, a main-effect genetic risk score for LDL-C was not useful for predicting dietary fat response. Further investigation of this score with respect to downstream disease outcomes revealed suggestive differential associations across DM trial arms, especially with respect to stroke subtypes. These results lay the foundation for the combination of many genome-wide gene-diet interactions for diet response prediction while highlighting the need for further research in order to achieve robust biomarkers for use in personalized nutrition.

4.1.2 Introduction

Nutrigenetics approaches (in which genetic information is used to predict response to dietary inputs) are a central component to the emerging promise of personalized nutrition for cardiometabolic risk reduction. Inter-individual differences in food preferences, metabolism, detoxification, excretion, etc. affect our responses to diet, in a similar way to the well-studied field of pharmacogenomics (Ma and Lu 2011). Ideally, genotype-based nutrigenetic investigations would be conducted in large-scale dietary interventions. Two notable examples are the PREDIMED and POUNDS LOST trial, with significant findings including the interaction of a *TCF7L2* variant with a Mediterranean diet pattern for glycemic traits (Corella et al. 2013) and the interaction of a *PCSK9* variant with dietary carbohydrate for insulin resistance (Huang et al. 2015). However, these are only able to examine one dietary change (whether food, nutrient, or pattern) at a time, and are often limited to lower sample sizes (Ordovas et al. 2018).

To allow for more flexibility and greater statistical power, gene-diet interactions (GDIs) are more commonly investigated in observational datasets. There is a rich literature of GDI discovery in the cardiometabolic realm. Typically, these focus on cardiometabolic risk factors based on biology-based candidate genes/variants (Corella et al. 2009; Cuda et al. 2011), but some have looked at actual outcomes (e.g. MI (Cornelis et al. 2006)). Other

approaches use main-effect genetic risk scores, such as that for obesity interacting with sugar-sweetened beverage intake to influence anthropometric traits (Qi et al. 2012; Olsen et al. 2016).

Characterization of individuals based on single or small groups of single nucleotide polymorphisms (SNPs) may neglect important signal elsewhere in the genome, especially when dealing with highly polygenic cardiometabolic traits. Thus, for effective personalized nutrition approaches to be realized, it will be necessary to integrate signals across the genome. A few investigations have explored GDIs genome-wide, such as for dairy and BMI (Smith et al. 2018) and for various dietary components and colorectal cancer (Figueiredo et al. 2014). However, genome-wide interaction studies (GWIS) can be problematic due to the lower statistical power inherent in gene-environment interaction analyses (Dempfle et al. 2008). Furthermore, the potential for confounding and reverse causation (i.e. cardiometabolic risk impacting dietary behavior) in statistical interactions from observational data means that GDIs may not always predict modification of the risk factor in question after a dietary intervention.

In order to provide proof-of-concept for use of GDIs in developing comprehensive diet response genetic scores, we sought to develop a genome-wide, GDI-based dietary fat response score for each of a series of cardiometabolic risk factors (CRFs). We performed genome-wide interaction studies prioritizing sites showing nominal main effects in prior genome-wide association studies, and used these intermediate results to derive fat response scores (FRS) for each CRF. We tested the performance of these scores in the fat reduction-focused Women's Health Initiative Dietary Modification trial, finding that a FRS for LDL-cholesterol (LDL-C) predicts 1-year LDL-C changes selectively in the intervention arm. Furthermore, we found that despite trends in the expected direction, the LDL-C FRS did not predict incident coronary heart disease or other chronic disease outcomes over approximately 20 years of follow-up.

4.1.3 Results

Dietary fat responder score development

A series of genome-wide interaction studies (GWIS) were undertaken in women from the Women's Health Initiative studies who did not participate in the dietary modification (DM) trial, using imputed genotypes along with baseline self-reported diet intakes (from food frequency questionnaires) and fasting blood biomarkers. Baseline characteristics of these women, along with those participating in the DM trial, are shown in Table 1.

Preliminary power calculations were undertaken, based on parameter assumptions including a modest SNP main effect (0.5%) under an additive model and a binary environment with 50% prevalence explaining 10% of the outcome phenotypic variance. The results showed that, at the sample sizes available for European ancestry non-DM participants (7000-10000 individuals for each cardiometabolic risk factor (CRF)), this analysis was powered to detect only moderately large interaction effects (GxE variance explained of greater than approximately 0.5%) at genome-wide significance (Supp. Table S1). Based on these results, a filter for nominal main-effect SNPs ($p < 0.05$ in existing

large-scale GWAS for each CRF) was applied before score development in order to balance the contribution of variants with smaller interaction effect sizes and the inclusion of spurious genetic relationships.

Table 1: WHI Population Description (European ancestry only)

	WHI DM trial	WHI non-DM trial
Sample size	6173	11131
Age	66 (61-70)	68 (64-72)
Current smoking	7%	9%
Lipid-lowering medication	11%	13%
Hypertension medication	38%	36%
Diabetes medication	5%	5%
Body mass index (BMI; kg/m ²)	28.7 (25.2-32.8)	26.9 (23.7-30.9)
Systolic blood pressure (SBP)	128.5 (117.5-140)	144 (132-156)
LDL cholesterol (LDL-C; mg/dL)	150.8 (127.5-175)	161 (134-191)
HDL cholesterol (HDL-C; mg/dL)	50 (42.5-58)	51 (44-60)
Triglycerides (TG; mg/dL)	137 (100-194)	128 (92-179.6)
Fasting glucose (FG; mg/dL)	97 (90-107)	97 (90-113.3)

* Continuous values shown as: median (interquartile range)

Dietary fat response scores were generated for each CRF using results from the corresponding GWIS analysis. Linear regression models were fit for log-transformed CRF values (other than LDL-C), incorporating gene-dietary fat interactions (dietary fat represented as a binary % of total calories above or below the median) while adjusting for age, total calories, and five genetic principal components. Q-Q plots of the GWIS results showed that genomic inflation was fairly well-controlled (Supp. Fig. S1). For each CRF, the associated summary statistics (corresponding to the fat-genotype interaction term estimates) were filtered to include only those with nominal main-effect associations in large-scale published GWAS. A pruning-and-thresholding method was used to generate six FRS from these individual sets of summary statistics along with genotypes from the WHI DM participants as a linkage disequilibrium (LD) reference. Using parameters of seed p -value=0.05 and LD r^2 <0.5, six sets of score weights were generated, with relevant SNP set sizes ranging from 1536 (SBP) to 6042 (BMI). Scores were then calculated as the weighted sum of allele counts across SNPs, normalized by the number of non-missing SNPs per individual.

Dietary fat responder score assessment

As the scores were developed to predict a positive interaction with dietary fat intake, the expected direction of the FRS effect on risk factors in the present fat-reduction trial would be negative. Of the fat response scores examined, only the LDL-C fat response score (LDL-FRS) was predictive at p <0.05 of the associated CRF change in DM trial participants in the fat-reduction arm (passing a Bonferroni correction for the 6 CRFs tested in baseline-adjusted sensitivity models). For this score, the standardized effect size was -0.18 (corresponding to a 5.1 mg/dL greater decrease in LDL-C per score standard deviation;

$p=0.026$). We note that the sample size of European-ancestry DM trial participants with follow-up measurements was much smaller for biochemical variables ($n\sim 150$) compared to BMI and SBP ($n\sim 2000$). Using the score developed in European-ancestry individuals, score performance was then tested in a combined-ancestry group including Black and Hispanic individuals, which almost doubled the sample size (Supp. Table S2). While some traits showed strong relationships (e.g. SBP), the signs of many were in a counter-intuitive direction, including a flip in sign for the previously-strong LDL-C relationship, suggesting that these results reflect primarily differences in ancestry rather than the intended biological differences. This observation was reinforced by the lack of association of the score with CRF changes in either Blacks or Hispanics alone.

Table 2: Responder score effects on CRF changes in DM trial participants

Risk factor	# SNPs in risk score	N	Univariate		Baseline-adjusted	
			Std. effect size	P-value	Std. effect size	P-value
BMI	6042	1988	0.03	0.221	0.02	0.253
SBP	1536	2004	0.03	0.196	0.04	0.061
LDL-C	1760	145	-0.18	0.026	-0.19	0.007
HDL-C	1731	150	-0.06	0.471	-0.06	0.483
TG	1774	150	-0.14	0.066	-0.14	0.064
FG	1924	281	0.02	0.791	0.03	0.570

* N = sample size available with 1-year follow-up measurements for the CRF in question

* Std. effect size represents the regression coefficient estimate in terms of CRF standard deviation per responder score standard deviation

Based on its observed association in European-ancestry participants, the LDL-FRS was investigated further. Linear models showed that the LDL-FRS accounted for 3.4% of the variance in 1-year LDL-C changes in the DM intervention arm. In baseline-adjusted models, this figure rose slightly to 3.9, based on the R^2 change from a baseline-only model. Additional models confirmed an interaction between DM trial arm and the LDL-FRS ($p=0.011$), supporting the specificity of this score for the fat-reduction arm. The LDL-FRS also showed specificity for LDL-C in that it did not predict changes in any other CRF (Supp. Table S3). The 1760 component SNPs were annotated using Annovar (Wang, Li, and Hakonarson 2010), revealing a predominance of intergenic and intronic variants and a set of genes with high numbers of independent SNPs contributing to the score (Fig. 1a-d). Top genes by number of contributing SNPs included *CSMD1*, *PTPRD*, and *RGS12*.

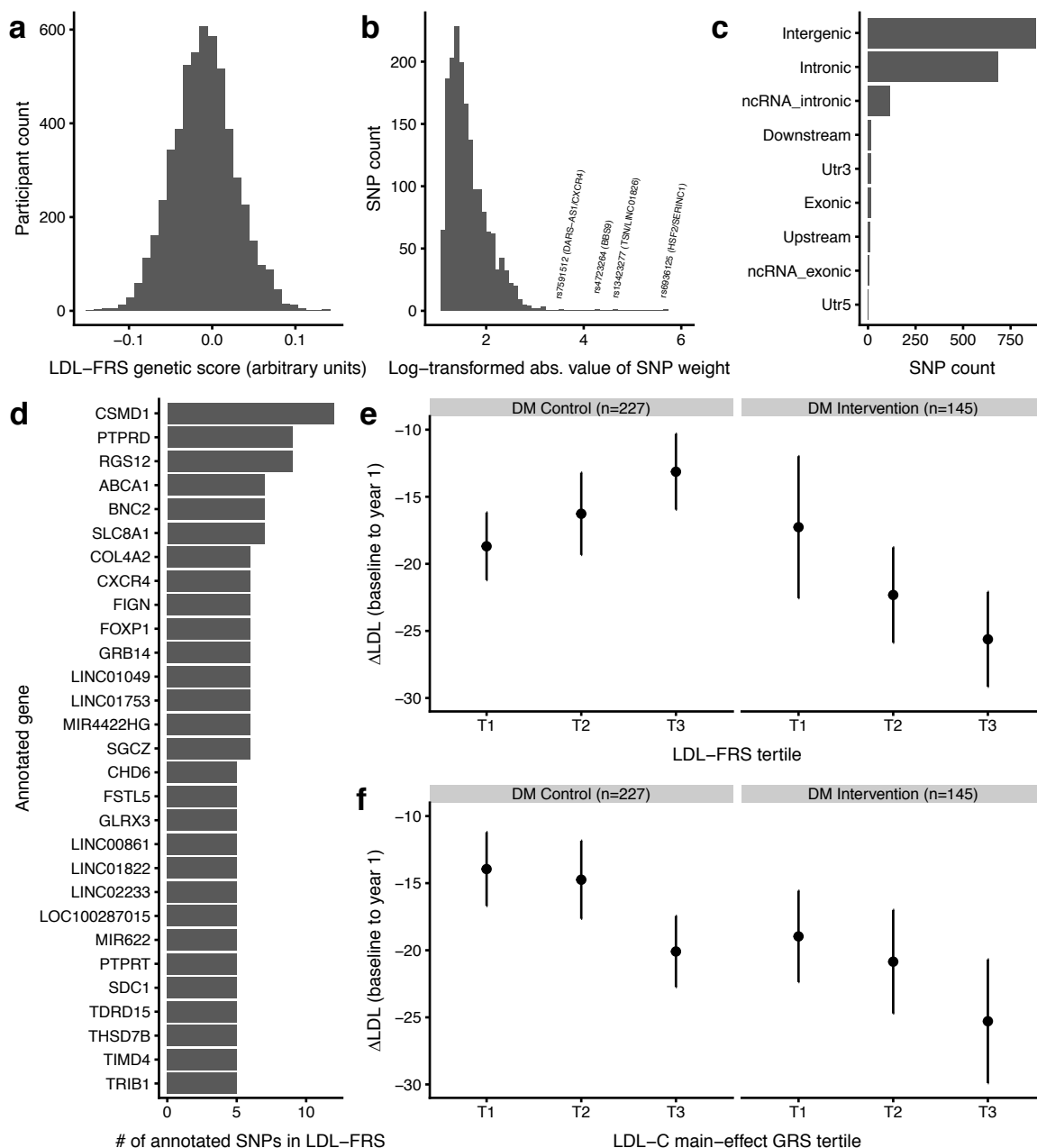


Figure 1: LDL-FRS characterization. a) Distribution of LDL-FRS in WHI DM trial participants. b) Distribution of SNP weights constituting the LDL-FRS (shown as the natural log-transformed absolute values of the true weights). c) SNP counts in different loci types for LDL-FRS constituent SNPs. d) Genes are summarized by the number of annotated SNPs in the LDL-FRS (genes with at least 5 component SNPs are shown). e-f) 1-year changes in LDL-C in DM trial participants as a function of genetic scores. Mean changes in LDL-C (y-axis) are shown as a function of either LDL-FRS (e) or LDL-GRS (f) tertile (x-axis). Error bars represent standard errors. LDL-FRS: LDL-C fat response genetic score, LDL-GRS: LDL-C main-effect genetic score.

Differences in mean LDL-C changes during the DM trial across genetic score strata are shown in Fig. 1e,f. As suggested by the regression results, those in the control arm trended

towards less strong LDL-C reductions in higher LDL-FRS strata, while those in the fat-reduction arm showed the opposite trend. Furthermore, isolation of individuals at the extremes of the score (top 10%) revealed an LDL-C reduction of almost double that of the rest of the DM intervention group (-39.0 versus -20.0 mg/dL). For comparison to the FRS, a main-effect genetic risk score (GRS) for LDL-C was developed using summary statistics from the Global Lipid Genetics Consortium meta-analysis (Willer et al. 2013) and an identical pruning-and-thresholding procedure to that used for the GDI-based scores. As expected, this score was strongly predictive of baseline LDL-C concentrations ($p=4.53 \times 10^{-27}$). However, unlike the GDI-based score, the GRS did not predict LDL-C changes in the DM intervention group ($p=0.15$; stratum-specific mean changes in Fig. 1f).

LDL-FRS association with chronic disease outcomes

Next, the LDL-FRS was tested for relationships with incident disease outcomes over approximately 22 years of follow-up using Cox proportional hazards models adjusted for age at baseline (Fig. 2). In addition to intervention versus control arm, another set of “per protocol-like” strata were produced by additionally filtering for FFQ-based self-reported fat reduction (in the intervention group) or fat increase (in the control group). While CHD and ischemic stroke qualitatively showed the expected interaction, i.e. a stronger inverse association between LDL-FRS and disease risk in the fat reduction group), these relationships contained a high amount of variability. Hemorrhagic stroke, while having a low number of events (44 in total), showed a positive association only in the fat reduction group ($p = 0.019$). Results for diabetes mirrored those for CHD and ischemic stroke, while those for non-CVD death were equivocal and did not vary across groups. Hemorrhagic stroke showed significant positive associations with the LDL-FRS only in the fat reduction group. None of these comparisons between trial arms reached nominal statistical significance ($p < 0.05$) in interaction models (score x arm).

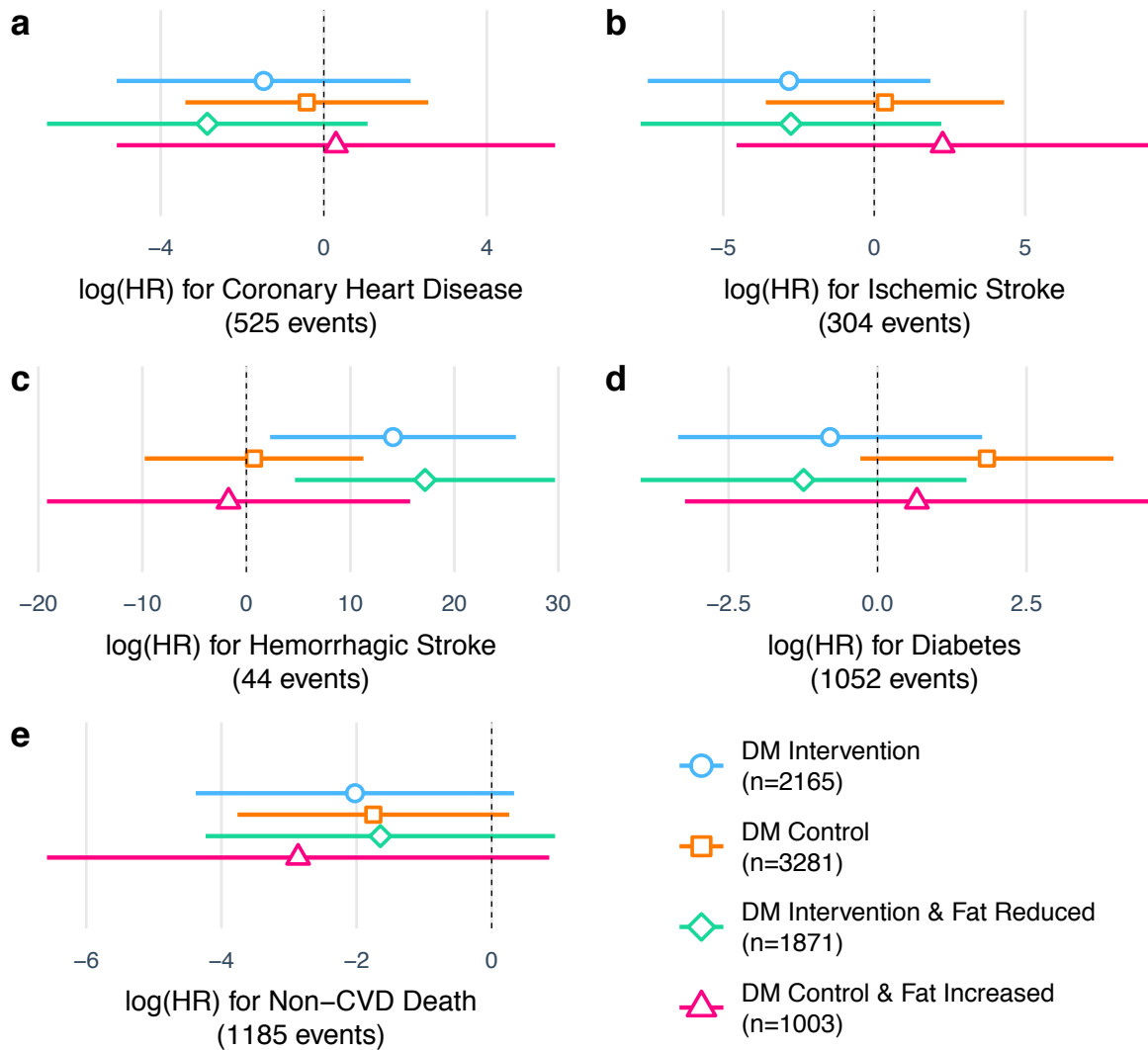


Figure 2: LDL-FRS prediction of chronic disease development. X-axis shows log-hazard ratio estimates for the LDL-FRS from Cox proportional hazards regression for a) coronary heart disease, b) ischemic stroke c) hemorrhagic stroke, d) diabetes, and e) non-CVD death. Separate estimates are shown for DM trial intervention arm, control arm, and the same strata filtered for FFQ-reported fat reduction or increase, respectively. Error bars represent 95% confidence interval estimates for the regression coefficients.

4.1.4 Discussion

Diet response scores have shown some success in predicting the response of cardiovascular risk factors (CRFs) to nutritional interventions, but they are often based solely on main effects or single GDI SNPs. Here, we explored the potential for gene diet interaction (GDI)-based diet responder score development, leveraging the multi-trial setup of the Women's Health Initiative. We developed what to our knowledge is the first example of a diet response score based on a hypothesis-free genome scan for each of six risk factors, and showed preliminary evidence for the viability of a LDL-C fat response score. The set of SNPs used for each score were limited to those showing nominal main effects in large-

scale GWAS as a compromise between discovery and utilization of prior information, but future investigations with larger sample sizes may have the power to use a fully unbiased approach.

Though FRS for six CRFs were developed and tested, only that for LDL-C showed nominal significance in predicting 1-year changes in the corresponding CRF. Multiple factors could explain this lack of predictive performance in general. First, analysis in observational datasets carries with it the potential for confounding of the observed relationships. Second, FFQs are imprecise instruments for measuring dietary intake, and though the FFQs used in the present study were optimized for detection of dietary fat, there was potential for substantial misclassification of this environmental exposure. Third, power calculations shown here and elsewhere suggest that a cohort of this size may not be powered to detect many very small gene-environment interactions such as are sought in genome-wide approaches like the one used here.

CSMD1, *PTPRD*, and *RGS12* stood out as genes containing the highest number of SNPs in the LDL-FRS (11, 9, and 9, respectively, after LD-pruning for $r^2 < 0.5$). *CSMD1* variants are notably associated with LDL-C response to statin treatment (Thompson et al. 2009) as well as SBP response to a high-salt diet [Newton-Cheh2009]. *CSMD1* has also shown epigenetic associations with LDL-C (Bell et al. 2012) as well as response to modification of dietary fat composition (Perfilyev et al. 2017). *PTPRD* variants modulate the response of T2D patients to pioglitazone therapy (Pei et al. 2013) and show suggestive associations with eating behaviors (caloric intake at dinner) (Comuzzie et al. 2012). *RGS12* has been linked to LDL-C in GWAS (Spracklen et al. 2017). Altogether, these genes have literature evidence for relationships to dietary intake, response to cardiometabolic therapies, and LDL-C, but have not until now been shown to directly modify the LDL-C response to dietary fat proportions. We note that there is a bias towards identifying LDL-C-related variants in the LDL-FRS, as only nominally-associated main-effect SNPs were used as input to the score development algorithm.

A reasonable body of literature exists establishing GDIs for both dietary fat on CRFs (Cuda et al. 2012; Zheng et al. 2015) and general dietary exposures on LDL-C (Ordovás, Robertson, and Cléirigh 2011). Multiple studies have looked specifically at genetic variants modulating the LDL-C response to dietary fat. For example, a caloric restriction intervention in type 2 diabetics was more effective in reducing LDL-C in ApoE4 carriers (-15.6% versus -0.7%) (Saito et al. 2004). Two studies using the POUNDS LOST trial found relationships of specific polymorphisms with 2-year changes in response to a dietary fat intervention. Carriers of a risk allele at the *APOA5* variant rs964184 show a decrease of 7.5 mg/dL in LDL-C in low-fat but not high-fat interventions (Zhang et al. 2012), and carriers of the minor allele at CETP rs3764261 show an 8.9 mg/dL greater decrease in LDL-C on a low-fat diet (Xu et al. 2015). Our observed effect size of a 5.4 mg/dL decrease in LDL-C is of a similar magnitude to these findings, and emerged despite the multifactorial nature of the WHI DM trial (which incorporated additional nutrition recommendations). The observed variance explained of 3.4% for the LDL-FRS means that the score, while contributing meaningfully to the prediction, does not capture most of the interindividual variability in LDL-C response to the WHI DM trial intervention. Based on prior observations of an inflection point in the impact of various genetic risk scores near the 90th percentile (Khera et al. 2018), we additionally evaluated the impact of LDL-FRS

in the top 10%, finding almost double the LDL-C reduction in DM intervention participants with values at this extreme.

There has been interest in the past in using main-effect genetic risk scores (GRS) as genetic variables in order to improve statistical power to detect gene-environment interactions (Aschard 2016). Such interactions may be viewed from a lens in which genetic risk corresponds to a predisposition that is only triggered in certain environments (e.g. dietary behaviors). Here, we observed only a minor association of a main-effect GRS for LDL-C with greater LDL-C reductions in the DM trial ($p=4.5e-27$). This trend runs counter to a prior observation of greater lifestyle intervention effectiveness for LDL-C reduction in those with low genetic risk of hyperlipidemia (Zubair et al. 2019). This discrepancy may be due to differences between the DM trial and the personalized diet and lifestyle changes recommended in the intervention in question. Regardless, the meaningful increase in predictive power of the GDI-based FRS compared to the main-effect GRS for LDL-C indicates the value in using interaction-based scores rather than simple genetic predispositions for the development of personalized dietary recommendations.

A diet response score such as that developed here is most useful if its value extends beyond just risk factor changes and predicts downstream changes in chronic disease and mortality risk. A modest but non-significant interaction for CHD, ischemic stroke, and diabetes was qualitatively apparent across strata (Fig. 2a,b,d), corresponding to a decreased risk in fat-reduction participants whose LDL-C would be expected to drop more prominently according to the score. In contrast, hemorrhagic stroke showed the opposite trend, with a positive score-disease relationship only in the fat reduction group. This result is in line with existing evidence for the detrimental effects of low LDL-C on hemorrhagic stroke risk (Sun et al. 2019). Non-CVD death showed no major associations, which could be expected due to the dominance of this category by cancer outcomes and the equivocal associations of cancer with lipids (Koene et al. 2016). We note that all disease outcome relationships assessed here are subject to the major caveat that dietary evolution and decreased adherence likely developed over time in many subjects, diluting the utility of the randomization and 1-year changes used for stratification in Fig. 2.

The present study had the advantage of developing a GDI-based dietary fat score in almost 10,000 women and testing in a dietary intervention trial from the same population. However, nominal main-effect SNPs were prioritized to improve statistical power given this moderate sample size, an approach which may fail to identify interactions with effect directions opposite that of the main effect. Smaller fractions of alternate ancestries in this population also made development of ancestry-specific response scores unrealistic. Additionally, the DM trial intervention in which the scores were tested may not exactly match the intervention relevant to the purely fat reduction-focused score developed here; it included additional non-fat-related dietary recommendations that may have affected the interactions examined here, and did not ultimately achieve its intended 20% fat reduction. Finally, this study only examined women, despite the fact that CRF profiles and their genetic trait architectures are known to vary across sexes (Knopp et al. 2006).

In summary, we present a method for the development of diet response scores based on genome-wide gene-diet interaction study summary statistics. While the resulting dietary fat response scores are not all informative, a score focused on LDL-C is predictive of 1-

year LDL-C changes during a fat reduction trial. Its performance is risk factor-specific, and is superior to that of a main-effect genetic risk score for LDL-C. Furthermore, it displays suggestive relationships with chronic disease outcomes, adding a new perspective to known discrepancies between associations of LDL-C with cardiovascular disease and stroke subtypes. These findings support the utility of gene-diet interactions for personalized nutrition while highlighting the need for increased sample sizes and improved diet measures for the discovery of robust genetic predictors of diet response.

4.1.5 Methods

Women's Health Initiative Dataset

The Women's Health Initiative study consists of a series of substudies: three clinical trials (related to cancer, cardiovascular disease, and osteoporosis) and an observational study (Anderson et al. 1998). Over 160,000 participants were enrolled between 1993-1998, with the ability to enroll in up to 3 of the clinical trials simultaneously. For the purposes of this analysis, participants were categorized based only on whether or not they were enrolled in the dietary modification (DM) trial, which randomized almost 50,000 women to a low-fat diet or a control diet with no recommended dietary changes, with primary outcomes being incidence of breast and colorectal cancers and heart disease (Ritenbaugh et al. 2003). Participants were comprehensively screened at baseline, including physical measurements, blood sample collection, and questionnaire administration, while only a subset of participants provided blood samples or returned questionnaires during later visits. The food frequency questionnaire (FFQ) was designed specifically for the WHI study, emphasizing specific foods and preparation methods to maximize its sensitivity to changes in fat intake (Patterson et al. 1999).

Phenotype data were accessed from dbGaP (accession: phs000746.v2.p3). Values shown in Table 1 only pertain to women whose genotypes were measured in one of a series of follow-up studies. For gene-diet interaction analyses, SBP, LDL-C, and GLU were adjusted for medication use: LDL-C and GLU values were divided by 0.75 for those on lipid-lowering and anti-diabetic medication, respectively, and SBP values were increased by 15 mmHg for those on anti-hypertensive medication. Cardiovascular risk factors (CRFs) were winsorized at 5 standard deviations from the mean and those other than LDL-C (BMI, SBP, HDL-C, TG, and GLU) were log-transformed prior to analysis. Longitudinal risk factor changes were calculated in DM trial participants as the difference between baseline and year 1. Adjudicated time-to-event data for chronic disease outcomes (coronary heart disease, myocardial infarction, ischemic stroke, hemorrhagic stroke, and non-CVD death) were collected, while diabetes incidence was defined as the self-report of any of: diabetes pills, insulin treatment, or general treatment for diabetes. Follow-up data was available for approximately 22 years following enrollment. Phenotype data processing was performed using R version 3.4.3 (R Core Team 2017) and Python version 3.6.0.

Genotype data and preprocessing

Imputed genotype data were retrieved from dbGaP (accession: phs000746.v2.p3) as a harmonized set of imputation outputs from a series of genotyping studies involving WHI

participants. Prior to imputation, study-specific quality control steps had been undertaken on directly-genotyped SNPs, with filters based on sample and call rate, Hardy-Weinberg equilibrium, and minor allele frequency. Phasing had been performed for autosomes using BEAGLE, followed by imputation using minimac (MACH for the SHARe study subset). After download from dbGaP, variants were filtered for imputation R-squared > 0.3 and minor allele frequency (MAF) > 0.001 and annotated with rsIDs, loci, and allelic information using the 1000 Genomes Phase 3 download from dbSNP (download date: April 13, 2018). Only variants passing the imputation quality threshold in all genotyping sub-studies were included in the final dosage dataset. For score development and calculation, imputed variant dosages were converted to hard-calls and set to missing if the estimated dosage was not within 0.1 of an integer allele count. Post-imputation genotype data processing was performed using PLINK 2.0, while clumping and score calculation were performed using PLINK 1.9 (Chang et al. 2015).

Genome-wide interaction study

A genome-wide interaction study was performed for each of the six cardiometabolic risk factors. The genome-wide scan used an additive genotype model, adjusted for dietary fat (binary: % of kcals above or below the median), total kcals per day, age, and five ancestry principal components. The primary estimand of interest was the interaction term between dietary fat and minor allele count at the SNP of interest. Interaction analyses were carried out using PLINK 2.0 (Chang et al. 2015). Variants of interest were annotated to genes using the Annovar tool (Wang, Li, and Hakonarson 2010).

Gene-environment interaction power calculations for single SNPs were performed using the Quanto tool (Gauderman 2002). The following assumptions were made: additive model; variance explained by genotype alone = 0.5%; and binary environment with 50% prevalence and explaining 10% of variance. (Note: there is no effect of minor allele frequency in this case given that variances explained are directly specified.)

Genetic responder score construction and evaluation

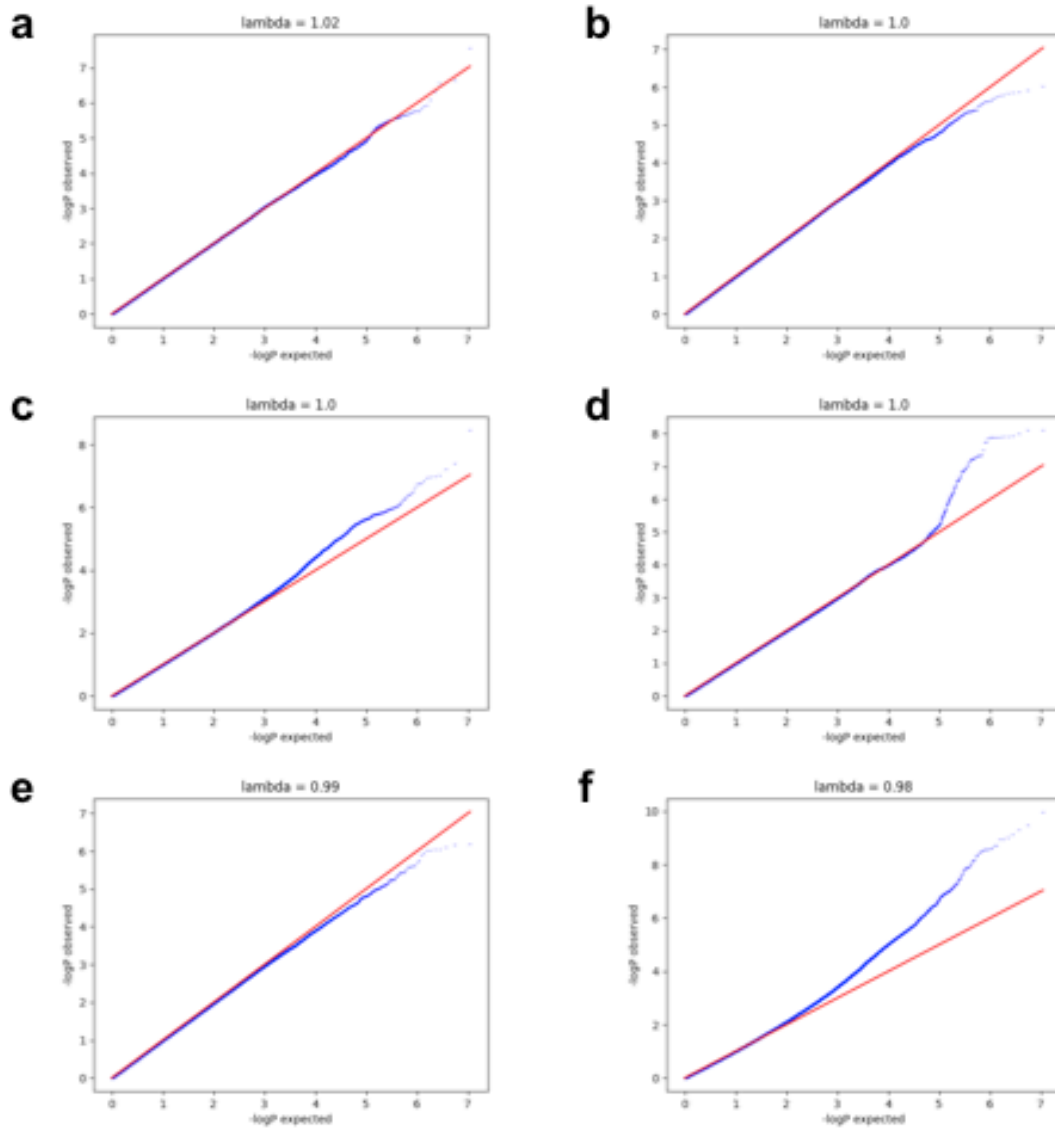
To prioritize variants for inclusion in genetic responder scores, nominal ($p < 0.05$) variants for main-effect on each risk factor were retrieved from large-scale meta-analyses: GIANT for BMI (Yengo et al. 2018); International Consortium for Blood Pressure for SBP (Ehret et al. 2011); Global Lipid Genetics Consortium (GLGC) for LDL-C, HDL-C, and TG (Willer et al. 2013); and MAGIC for fasting glucose (Dupuis et al. 2010). Each FRS was constructed using summary statistics for the diet-SNP interaction terms from the associated GWIS. SNPs were filtered for nominal main-effect relationships using the above meta-analyses, and interaction summary statistics were used as input to a pruning-and-thresholding (P&T) procedure (using the “-clump” function in PLINK 1.9), with a seed threshold of $p = 0.05$ and an LD threshold of $r^2 = 0.5$. The linkage disequilibrium references for the procedure was calculated from genotypes of the white DM trial participants. Genetic fat response scores (FRS) for each individual were then calculated as the weighted sum of allelic dosages for variants selected by the P&T procedure, with weights corresponding to the GWIS interaction term estimates. A genetic risk score for LDL-C was created using the GLGC LDL-C meta-analysis summary statistics and the same P&T method and parameters as was used for the interaction analyses, resulting in a 26467-SNP score.

FRS were used to test for discrimination of changes in CRFs over the first year of the DM trial. Risk factor changes were assessed using linear models in participants in the intervention arm, with and without adjustment for baseline CRF levels. As a sensitivity analysis, p-values were calculated in separate models for interaction of the genetic score with 1) trial arm (control vs. dietary modification), and 2) observed fat reduction (negative vs. positive 3-year change in FFQ-estimated dietary fat). GRS were further tested for prediction of chronic disease development during follow-up across DM trial strata. Time-to-event for each of coronary heart disease, myocardial infarction, ischemic stroke, hemorrhagic stroke, and non-CVD death were used to fit age-adjusted Cox proportional hazards models. Estimated log-hazard ratios were extracted from regressions conducted in the following strata: 1) DM trial intervention arm, 2) DM trial control arm, 3) DM trial intervention arm filtered for participants with 1-year fat reduction based on FFQ, and 4) DM trial control arm filtered for participants with 1-year fat increase based on FFQ.

4.1.6 Supplementary Data

Supp. Table S1: Sample size necessary to achieve power of 0.8

GxE variance explained (%)	N (nominal)	N (suggestive)	N (genome-wide)
0.05	14046	49488	70866
0.10	7021	24737	35423
0.50	1401	4936	7069
1.00	699	2461	3524



Supp. Fig. S1: Q-Q plots from individual CRF GWIS. The distribution of p -values from each GWIS is plotted against the expected uniform p -value distribution. Plots correspond to: a) BMI, b) SBP, c) LDL-C, d) HDL-C, e) TG, and f) FG.

Supp. Table S2: Responder score effects on CRF changes in DM trial participants across ancestries

CRF	All combined			Black			Hispanic		
	N	Std. effect size	P-value	N	Std. effect size	P-value	N	Std. effect size	P-value
BMI	3606	-0.02	0.24	1214	0	0.994	404	0.018	0.719
FG	572	0.08	0.043	214	-0.004	0.955	77	-0.014	0.901
HDL-C	430	-0.06	0.224	206	-0.096	0.182	74	-0.057	0.66
LDL-C	422	0.084	0.066	206	0.055	0.412	71	0.126	0.227
SBP	3645	0.041	0.015	1230	0	0.991	411	-0.008	0.868
TG	430	-0.075	0.112	206	-0.023	0.743	74	-0.098	0.433

Supp. Table S3: LDL-FRS effects on alternate CRF changes in DM trial participants

Outcome risk factor	# SNPs in risk score	Sample size	Std. effect size	P-value
BMI	1747	1988	-0.02	0.36
SBP	1747	2004	0.00	0.83
HDL-C	1747	150	-0.07	0.39
TG	1747	150	0.11	0.16
FG	1747	281	0.01	0.82

* Std. effect size represents the regression coefficient estimate in terms of CRF standard deviation per responder score standard deviation.

4.1.7 References

Anderson, Garnet L, Steven R Cummings, Laurence S Freedman, Curt Furberg, Maureen M Henderson, Susan R Johnson, Lewis H Kuller, et al. 1998. "Design of the Women's Health Initiative Clinical Trial and Observational Study." *Controlled Clinical Trials* 19 (1): 61–109. [https://doi.org/10.1016/S0197-2456\(97\)00078-0](https://doi.org/10.1016/S0197-2456(97)00078-0).

Aschard, Hugues. 2016. "A perspective on interaction effects in genetic association studies." *Genetic Epidemiology* 40 (8): 678–88. <https://doi.org/10.1002/gepi.21989>.

Bell, Jordana T., Pei-Chien Tsai, Tsun-Po Yang, Ruth Pidsley, James Nisbet, Daniel Glass, Massimo Mangino, et al. 2012. "Epigenome-Wide Scans Identify Differentially Methylated Regions for Age and Age-Related Phenotypes in a Healthy Ageing Population." Edited by Jun Li. *PLoS Genetics* 8 (4): e1002629. <https://doi.org/10.1371/journal.pgen.1002629>.

Chang, Christopher C., Carson C. Chow, Laurent C.A.M. Tellier, Shashaank Vattikuti, Shaun M. Purcell, and James J. Lee. 2015. "Second-generation PLINK: rising to the challenge of larger and richer datasets." *GigaScience* 4 (1): 7. <https://doi.org/10.1186/s13742-015-0047-8>.

Comuzzie, Anthony G, Shelley A Cole, Sandra L Laston, V Saroja Voruganti, Karin Haack, Richard A Gibbs, and Nancy F Butte. 2012. "Novel Genetic Loci Identified for the Pathophysiology

of Childhood Obesity in the Hispanic Population.” Edited by Dana C. Crawford. *PLoS ONE* 7 (12): e51954. <https://doi.org/10.1371/journal.pone.0051954>.

Corella, D., P. Carrasco, J. V. Sorli, R. Estruch, J. Rico-Sanz, M. A. Martinez-Gonzalez, J. Salas-Salvado, et al. 2013. “Mediterranean Diet Reduces the Adverse Effect of the TCF7L2-rs7903146 Polymorphism on Cardiovascular Risk Factors and Stroke Incidence: A randomized controlled trial in a high-cardiovascular-risk population.” *Diabetes Care*, dc13-0955 –. <https://doi.org/10.2337/dc13-0955>.

Corella, Dolores, Gina Peloso, Donna K. Arnett, Serkalem Demissie, L. Adrienne Cupples, Katherine Tucker, Chao Qiang Lai, et al. 2009. “APOA2, dietary fat, and body mass index: Replication of a gene-diet interaction in 3 independent populations.” *Archives of Internal Medicine*. <https://doi.org/10.1001/archinternmed.2009.343>.

Cornelis, Marilyn C., Ahmed El-Sohemy, Edmond K. Kabagambe, and Hannia Campos. 2006. “Coffee, CYP1A2 genotype, and risk of myocardial infarction.” *Journal of the American Medical Association*. <https://doi.org/10.1001/jama.295.10.1135>.

Cuda, Cristina, Alaa Badawi, Mohamed Karmali, and Ahmed El-Sohemy. 2011. “Polymorphisms in Toll-like receptor 4 are associated with factors of the metabolic syndrome and modify the association between dietary saturated fat and fasting high-density lipoprotein cholesterol.” *Metabolism* 60 (8): 1131–5. <https://doi.org/10.1016/j.metabol.2010.12.006>.

Cuda, Cristina, Bibiana Garcia-Bailo, Mohamed Karmali, Ahmed El-Sohemy, and Alaa Badawi. 2012. “A common polymorphism near the interleukin-6 gene modifies the association between dietary fat intake and insulin sensitivity.” *Journal of Inflammation Research*. <https://doi.org/10.2147/JIR.S27911>.

Dempfle, Astrid, André Scherag, Rebecca Hein, Lars Beckmann, Jenny Chang-Claude, and Helmut Schäfer. 2008. “Gene-environment interactions for complex traits: Definitions, methodological requirements and challenges.” <https://doi.org/10.1038/ejhg.2008.106>.

Dupuis, Josée, Claudia Langenberg, Inga Prokopenko, Richa Saxena, Nicole Soranzo, Anne U. Jackson, Eleanor Wheeler, et al. 2010. “New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk.” *Nature Genetics*. <https://doi.org/10.1038/ng.520>.

Ehret, Georg B., Patricia B. Munroe, Kenneth M. Rice, Murielle Bochud, Andrew D. Johnson, Daniel I. Chasman, Albert V. Smith, et al. 2011. “Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk.” *Nature* 478 (7367): 103–9. <https://doi.org/10.1038/nature10405>.

Figueiredo, Jane C., Li Hsu, Carolyn M. Hutter, Yi Lin, Peter T. Campbell, John A. Baron, Sonja I. Berndt, et al. 2014. “Genome-Wide Diet-Gene Interaction Analyses for Risk of Colorectal Cancer.” *PLoS Genetics*. <https://doi.org/10.1371/journal.pgen.1004228>.

Gauderman, W. J. 2002. “Sample Size Requirements for Association Studies of Gene-Gene Interaction.” *American Journal of Epidemiology* 155 (5): 478–84. <https://doi.org/10.1093/aje/155.5.478>.

Huang, Tao, Jinyan Huang, Qibin Qi, Yanping Li, George A. Bray, Jennifer Rood, Frank M. Sacks, and Lu Qi. 2015. “PCSK7 Genotype Modifies Effect of a Weight-Loss Diet on 2-Year Changes of Insulin Resistance: The POUNDS LOST Trial.” *Diabetes Care* 38 (3): 439–44. <https://doi.org/10.2337/dc14-0473>.

Khera, Amit V., Mark Chaffin, Krishna G. Aragam, Mary E. Haas, Carolina Roselli, Seung Hoan Choi, Pradeep Natarajan, et al. 2018. "Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations." *Nature Genetics* 50 (9): 1219–24. <https://doi.org/10.1038/s41588-018-0183-z>.

Knopp, Robert H., Pathmaja Paramsothy, Barbara M. Retzlaff, Brian Fish, Carolyn Walden, Alice Dowdy, Christine Tsunehara, Keiko Aikawa, and Marian C. Cheung. 2006. "Sex differences in lipoprotein metabolism and dietary response: Basis in hormonal differences and implications for cardiovascular disease." *Current Cardiology Reports* 8 (6): 452–59. <https://doi.org/10.1007/s11886-006-0104-0>.

Koene, Ryan J, Anna E Prizment, Anne Blaes, and Suma H Konety. 2016. "Shared Risk Factors in Cardiovascular Disease and Cancer." *Circulation* 133 (11): 1104–14. <https://doi.org/10.1161/CIRCULATIONAHA.115.020406>.

Ma, Q., and A. Y. H. Lu. 2011. "Pharmacogenetics, Pharmacogenomics, and Individualized Medicine." *Pharmacological Reviews* 63 (2): 437–59. <https://doi.org/10.1124/pr.110.003533>.

Olsen, Nanna J., Lars Ångquist, Sofus C. Larsen, Allan Linneberg, Tea Skaaby, Lise Lotte N. Husemoen, Ulla Toft, et al. 2016. "Interactions between genetic variants associated with adiposity traits and soft drinks in relation to longitudinal changes in body weight and waist circumference." *The American Journal of Clinical Nutrition* 104 (3): 816–26. <https://doi.org/10.3945/ajcn.115.122820>.

Ordovas, Jose M., Lynnette R. Ferguson, E. Shyong Tai, and John C. Mathers. 2018. "Personalised nutrition and health." *BMJ*, June, bmj.k2173. <https://doi.org/10.1136/bmj.k2173>.

Ordovás, José M., Ruairi Robertson, and Ellen Ní Cléirigh. 2011. "Gene–gene and gene–environment interactions defining lipid-related traits." *Current Opinion in Lipidology* 22 (2): 129–36. <https://doi.org/10.1097/MOL.0b013e32834477a9>.

Patterson, Ruth E., Alan R. Kristal, Lesley Fels Tinker, Rachel A. Carter, Mary Pat Bolton, and Tanya Agurs-Collins. 1999. "Measurement Characteristics of the Women's Health Initiative Food Frequency Questionnaire." *Annals of Epidemiology* 9 (3): 178–87. [https://doi.org/10.1016/S1047-2797\(98\)00055-6](https://doi.org/10.1016/S1047-2797(98)00055-6).

Pei, Qi, Qiong Huang, Guo-ping Yang, Ying-chun Zhao, Ji-ye Yin, Min Song, Yi Zheng, Zhao-hui Mo, Hong-hao Zhou, and Zhao-qian Liu. 2013. "PPAR- γ 2 and PTPRD gene polymorphisms influence type 2 diabetes patients' response to pioglitazone in China." *Acta Pharmacologica Sinica* 34 (2): 255–61. <https://doi.org/10.1038/aps.2012.144>.

Perfilyev, Alexander, Ingrid Dahlman, Linn Gillberg, Fredrik Rosqvist, David Iggman, Petr Volkov, Emma Nilsson, Ulf Risérus, and Charlotte Ling. 2017. "Impact of polyunsaturated and saturated fat overfeeding on the DNA-methylation pattern in human adipose tissue: a randomized controlled trial." *The American Journal of Clinical Nutrition* 105 (4). American Society for Nutrition: 991–1000. <https://doi.org/10.3945/ajcn.116.143164>.

Qi, Qibin, Audrey Y. Chu, Jae H. Kang, Majken K. Jensen, Gary C. Curhan, Louis R. Pasquale, Paul M. Ridker, et al. 2012. "Sugar-Sweetened Beverages and Genetic Risk of Obesity." *New England Journal of Medicine* 367 (15): 1387–96. <https://doi.org/10.1056/NEJMoa1203039>.

R Core Team. 2017. "R: A language and environment for statistical computing." Vienna, Austria: R Foundation for Statistical Computing. <https://www.r-project.org/>.

Ritenbaugh, Cheryl, Ruth E Patterson, Rowan T Chlebowski, Bette Caan, Lesley Fels-Tinker, Barbara Howard, and Judy Ockene. 2003. "The women's health initiative dietary modification trial: overview and baseline characteristics of participants." *Annals of Epidemiology*. [https://doi.org/10.1016/s1047-2797\(03\)00044-9](https://doi.org/10.1016/s1047-2797(03)00044-9).

Saito, Mieko, Masaaki Eto, Hayami Nitta, Yukiko Kanda, Makoto Shigeto, Katsura Nakayama, Kazuhito Tawaramoto, et al. 2004. "Effect of Apolipoprotein E4 Allele on Plasma LDL Cholesterol Response to Diet Therapy in Type 2 Diabetic Patients." *Diabetes Care* 27 (6): 1276–80. <https://doi.org/10.2337/diacare.27.6.1276>.

Smith, Caren E., Jack L. Follis, Hassan S. Dashti, Toshiko Tanaka, Mariaelisa Graff, Amanda M. Fretts, Tuomas O. Kilpeläinen, et al. 2018. "Genome-Wide Interactions with Dairy Intake for Body Mass Index in Adults of European Descent." *Molecular Nutrition and Food Research*. <https://doi.org/10.1002/mnfr.201700347>.

Spracklen, Cassandra N., Peng Chen, Young Jin Kim, Xu Wang, Hui Cai, Shengxu Li, Jirong Long, et al. 2017. "Association analyses of East Asian individuals and trans-ancestry analyses with European individuals reveal new loci associated with cholesterol and triglyceride levels." *Human Molecular Genetics* 26 (9): 1770–84. <https://doi.org/10.1093/hmg/ddx062>.

Sun, Luanluan, Robert Clarke, Derrick Bennett, Yu Guo, Robin G. Walters, Michael Hill, Sarah Parish, et al. 2019. "Causal associations of blood lipids with risk of ischemic stroke and intracerebral hemorrhage in Chinese adults." <https://doi.org/10.1038/s41591-019-0366-x>.

Thompson, John F., Craig L. Hyde, Linda S. Wood, Sara A. Paciga, David A. Hinds, David R. Cox, G. Kees Hovingh, and John J.P. Kastelein. 2009. "Comprehensive Whole-Genome and Candidate Gene Analysis for Response to Statin Therapy in the Treating to New Targets (TNT) Cohort." *Circulation: Cardiovascular Genetics* 2 (2): 173–81. <https://doi.org/10.1161/CIRCGENETICS.108.818062>.

Wang, Kai, Mingyao Li, and Hakon Hakonarson. 2010. "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data." *Nucleic Acids Research* 38 (16): e164–e164. <https://doi.org/10.1093/nar/gkq603>.

Willer, Cristen J., Ellen M. Schmidt, Sebanti Sengupta, Gina M. Peloso, Stefan Gustafsson, Stavroula Kanoni, Andrea Ganna, et al. 2013. "Discovery and refinement of loci associated with lipid levels." *Nature Genetics* 45 (11): 1274–83. <https://doi.org/10.1038/ng.2797>.

Xu, Min, San San Ng, George A Bray, Donna H Ryan, Frank M Sacks, Guang Ning, and Lu Qi. 2015. "Dietary Fat Intake Modifies the Effect of a Common Variant in the LIPC Gene on Changes in Serum Lipid Concentrations during a Long-Term Weight-Loss Intervention Trial." *The Journal of Nutrition* 145 (6): 1289–94. <https://doi.org/10.3945/jn.115.212514>.

Yengo, Loic, Julia Sidorenko, Kathryn E. Kemper, Zhili Zheng, Andrew R. Wood, Michael N. Weedon, Timothy M. Frayling, Joel Hirschhorn, Jian Yang, and Peter M. Visscher. 2018. "Meta-analysis of genome-wide association studies for height and body mass index in 700000 individuals of European ancestry." *Human Molecular Genetics* 27 (20): 3641–9. <https://doi.org/10.1093/hmg/ddy271>.

Zhang, Xiaomin, Qibin Qi, George A. Bray, Frank B. Hu, Frank M. Sacks, and Lu Qi. 2012. "APOA5 genotype modulates 2-y changes in lipid profile in response to weight-loss diet intervention: the Pounds Lost Trial." *The American Journal of Clinical Nutrition* 96 (4): 917–22. <https://doi.org/10.3945/ajcn.112.040907>.

Zheng, Yan, Tao Huang, Xiaomin Zhang, Jennifer Rood, George A Bray, Frank M Sacks, and Lu Qi. 2015. "Dietary Fat Modifies the Effects of FTO Genotype on Changes in Insulin Sensitivity." *The Journal of Nutrition* 145 (5): 977–82. <https://doi.org/10.3945/jn.115.210005>.

Zubair, Niha, Matthew P. Conomos, Leroy Hood, Gilbert S. Omenn, Nathan D. Price, Bonnie J. Spring, Andrew T. Magis, and Jennifer C. Lovejoy. 2019. "Genetic Predisposition Impacts Clinical Changes in a Lifestyle Coaching Program." *Scientific Reports* 9 (1): 6805. <https://doi.org/10.1038/s41598-019-43058-0>.

4.2 Epigenome-wide association study reveals a molecular signature of response to vitamin K supplementation

In preparation for submission to Epigenetics

4.2.1 Abstract

Evidence suggests there are roles for vitamin K in various chronic disease outcomes, but population-level diet and supplement recommendations are difficult to determine due to high levels of variability in measures of status and response to intake compared to other nutrients. In this preliminary investigation, a blood-based epigenome-wide association study (EWAS) was undertaken comparing responders and non-responders to vitamin K supplementation in order to better understand the molecular underpinnings of this observed variability. Twenty-four responders and twenty-four non-responders were identified in a prior 3-year phylloquinone supplementation trial based on their changes in plasma phylloquinone concentrations. Differential DNA methylation was identified at multiple novel loci with previously unknown relationships to phylloquinone absorption and metabolism. A hypothesis-driven analysis of lipid-related genes highlighted a site in the *NPC1L1* gene, supplementing existing evidence for its role in vitamin K absorption. This responder analysis was supported by a second EWAS for baseline plasma phylloquinone concentrations which showed a strong correlation between the epigenomic signatures of phylloquinone baseline status and response to supplementation. This work can guide future epigenomic research on vitamin K and contributes to the development of more personalized dietary recommendations for vitamin K.

4.2.2 Introduction

Vitamin K is a fat-soluble nutrient and essential cofactor for the post-translational γ -carboxylation of vitamin K dependent proteins. While it is best known for its role in coagulation, there is suggestive evidence for roles of vitamin K in various other health outcomes including inflammation, cardiovascular disease, and osteoarthritis (1,2).

Large inter-individual variability in measures of vitamin K status have been observed (3). Compared to circulating concentrations of the other fat-soluble vitamins, the major circulating form of vitamin K, phylloquinone, had the largest inter-individual variability (3). Stable isotope studies of phylloquinone have reported a bioavailability ranging from 2 to 26% (4). To date, biologically plausible factors have been examined, but most of the variability remains unaccounted for (5,6). A better understanding of the factors that influence biomarkers of vitamin K metabolism could help to clarify the contribution of vitamin K in health outcomes.

The bioavailability of other fat-soluble nutrients has been shown to be under partial genetic control (7). The genetic contribution to the inter-individual variability in circulating phylloquinone was examined in a small cross-sectional meta-analysis of genome-wide association studies (GWAS) (8). In this analysis, no single nucleotide polymorphism (SNP) reached genome-wide significance and no meaningful amount of variance in circulating phylloquinone concentrations could be explained by suggestive SNPs. The association of lipid-related candidate gene variants with the plasma phylloquinone response to

supplementation has also been examined. Though several variants identified were suggestive of mechanisms involved in phyloquinone metabolism, collectively the candidate SNPs did not explain the observed variability in the response to supplementation (9).

Epigenetic marks such as DNA methylation are emerging as a source of information that may be complementary to genetic variants in explaining trait variance (10). DNA methylation regulates gene expression through the modulation of transcription factor binding and chromatin accessibility, and displays a degree of temporal stability that allows for its use as a biomarker of modifiable phenotypic states (11). Furthermore, it has been demonstrated to predict response to intervention in both pharmaceutical (12) and nutritional (13) contexts. While these observations suggest that DNA methylation may be a valuable input for personalized nutrition approaches, to our knowledge, the role of DNA methylation in the variable response to vitamin K intake has not been examined.

The objective of this hypothesis-generating study was to characterize the epigenomic associations of DNA methylation with the plasma phyloquinone response to supplementation. To achieve this, we performed a microarray based, epigenome-wide association study (EWAS) using DNA isolated from 48 healthy older adults randomized to the treatment arm of a completed, 3-year, 500 µg/d phyloquinone supplementation trial. We identified differential methylation at both known and novel loci, and carried out a secondary EWAS of baseline plasma phyloquinone to demonstrate links between the epigenomic signatures of steady-state and dynamic phyloquinone concentrations.

4.2.3 Methods

Study Design and Participants

The present study was a secondary analysis of a completed, 3-year, double-blind, randomized, controlled phyloquinone supplementation trial (14). Based on a medical history questionnaire taken at baseline, the 452 community-dwelling men and women (60-80 years, 422 whites) who participated in the study were generally in good health, free of known coronary disease or osteoporosis, and not taking warfarin (a vitamin K antagonist). Equal numbers of participants were randomized to either the treatment (500 µg phyloquinone as part of a daily multivitamin formulation) or nontreatment (daily multivitamin without phyloquinone) group. All participants received a second daily tablet containing 600 mg calcium carbonate and 10 µg (400 IU) vitamin D as cholecalciferol. Protocol adherence was assessed by pill count. All participants provided written informed consent and the Institutional Review Board at Tufts University approved the protocol for this study (14). Consent for genomic data measurement was received for 416 participants. The trial was registered at clinicaltrials.gov as NCT00183001.

Biochemical Measures

Blood samples were drawn after a minimum of a 10-hour fast at baseline, 6, 12, 24, and 36 months (study completion). Samples were stored in individual cryogenic tubes at -80°C and protected from light until the time of analysis. Plasma phyloquinone was measured in the Vitamin K Laboratory at the Tufts University Human Nutrition Research

Center on Aging using high-performance liquid chromatography (HPLC), followed by fluorometric detection (15). Analysis was performed within 12 months of completion of the supplementation trial.

Responder and Non-responder Selection

For this hypothesis-generating pilot study, analysis was performed in 48 white participants randomized to the treatment group who consented for genotyping and demonstrated $\geq 85\%$ adherence to the protocol. The natural fluctuations (median of 4 follow-up measurements minus baseline measurement) in plasma phyloquinone over the 3 years in participants of the nontreatment group were determined (middle 95%) and used to classify participants in the treatment group with similar changes as non-responders. After meeting methylation analysis sample criteria (i.e. sufficient volume and minimum DNA concentration) and excluding participants with outsized variances in follow up measurements (standard deviation (SD) > 2 nmol/L), the 24 participants with the largest 3-year change in plasma phyloquinone were selected as responders. Of the participants classified as non-responders, the 24 with the smallest change in plasma phyloquinone who had baseline plasma phyloquinone measurements less than or equal to the maximum baseline plasma phyloquinone observed in the non-responders were selected for DNA methylation analysis.

DNA Methylation Data

DNA was isolated from blood and purified for PCR analysis using the QIAamp DNA Mini Kit (Qiagen Inc., Chatsworth, CA). Next, DNA was treated with bisulfite and amplified using PCR at the Yale Center for Genome Analysis. Methylation profiling was then performed using the Illumina MethylationEPIC microarray platform (16) and the data were downloaded as raw intensity files.

Methylation data preprocessing was performed using the *minfi* and *wateRmelon* packages for R (16,17). Sample-wise filters were as follows: robust overall signal in the main cluster based on visual inspection of an intensity plot, less than 10% of probes undetected at a detection threshold of $p < 1 \times 10^{-16}$ and a reported sex matching methylation-based sex prediction. Probes were removed using the following criteria: more than 10% of samples undetected at a detection threshold of $p < 1 \times 10^{-16}$, location in the X or Y chromosomes, non-CpG probes, cross-hybridizing probes, probes measuring SNPs, and probes with an annotated common SNP at the CpG site or in the single-base extension region. Samples were normalized using the Noob method for background correction and dye-bias normalization, followed by the BMIQ method for probe type correction (18,19). Blood cell fractions for 6 blood cell types (CD4⁺ T-cells, CD8⁺ T-cells, B-cells, natural killer cells, monocytes, and granulocytes) were estimated using a common reference-based method (20), and 5 of these (excluding granulocytes) were included in cell count fraction-adjusted statistical models. After quality control and filtering steps, 785,436 CpG sites were available for analysis, formatted as M-values (logit transform of the ratio of methylated signal to total microarray signal). Principal components analysis (PCA) was performed on the full set of M-values as well as on the set of control probes using code adapted from the CPACOR method (21). CpG annotations were retrieved using the MethylationEPIC

annotation file available from Illumina, and expression of genes of interest were accessed using the Human Protein Atlas (www.proteinatlas.org).

Statistical Analysis

Unless otherwise noted, statistical analyses were conducted using R version 3.4.2 (22). Differences in characteristics of the responders and non-responders were assessed with a *t*-test for continuous outcomes and a Chi-square test for dichotomous outcomes.

The phylloquinone responder EWAS was conducted to compare 24 responders and 24 non-responders, selected as described above. Univariate linear models were tested for each CpG site, with methylation M-value as the dependent variable and binary responder status as the independent variable. Model 1 (the primary model) was adjusted for 10 CPACOR control-probe principal components, which was found to be necessary to control the genomic inflation. Model 2 was adjusted additionally for age, sex, and 5 estimated cell count fractions (see above). Model 3 included all prior covariates as well as natural log-transformed plasma triglyceride concentrations at baseline. For each CpG, outliers were identified using a quantile-based method (M-values outside of 3 interquartile ranges from the 25th and 75th percentiles) and removed prior to regression analysis. Beta-value differences (shown in Table 2) were estimated for highlighted CpGs based on identical regressions (using Model 1) other than the replacement of methylation M-values with beta-values using an inverse logit transformation.

Because phylloquinone is transported on triglyceride-rich lipoproteins (23,24) and lipid-related SNPs have been linked to circulating phylloquinone (8), a candidate gene-site analysis was undertaken within this responder EWAS. CpG sites within the following lipid-related genes were examined: *APOA1*, *APOA5*, *APOB*, *APOE*, *LPL*, *NPC1L1*.

Under the presumption that variation in plasma phylloquinone concentrations due to a regular dietary intake would reflect differences in the same machinery responsible for variation in the plasma phylloquinone response to supplementation, an EWAS was conducted for baseline concentrations of phylloquinone to support the responder analysis. The independent variable was baseline plasma phylloquinone concentration (natural log-transformed) as a continuous value. Model 1 covariates included responder status and 5 full-dataset principal components (CPACOR control-probe principal components were not found to be helpful in controlling genomic inflation). Models 2 and 3 added covariates as in the responder analysis. One additional model was examined with adjustment for dietary phylloquinone intake at baseline, as assessed by the Willett semi-quantitative food-frequency questionnaire (FFQ).

Gene Ontology enrichment analyses were performed using the *missMethyl* package for R (*gometh* function). This method implements a hypergeometric test for gene set representation while accounting for the uneven representation of genes in the set of CpG sites on the MethylationEPIC microarray. Fully-adjusted models (Models 3 and 4 for responder and baseline analyses, respectively) were filtered for those CpGs with nominal p-values ($p < 0.05$) in both analyses at the regional and single-site level and categorized as either matching sign (of regression coefficients) or non-matching sign. Enrichment

analysis was then run for each of these sets independently to generate false discovery rates (FDRs) for Gene Ontology biological processes.

The Python-based tool Comb-p (25) was used to call differentially methylated regions (DMRs). The method uses EWAS p-values directly, removing the requirement for additional covariate adjustment. Comb-p first calculates an autocorrelation function (ACF), for which a maximum distance of 1kb and a step size of 50 bases were used. Next, it uses the ACF to adjust each p-value using a Stouffer-Liptak-Kechris correction (26), followed by identification of contiguous regions of sites with adjusted p-values below some threshold (here, $p < 0.1$ with no more than 500 bases between neighboring sites in a region). Finally, the ACF is recalculated out to the maximum region size (a step size of 50 was used here as well) and regional p-values are calculated using the Stouffer-Liptak test. For Šidák multiple testing correction of DMRs (27), Comb-p calculates the number of effective tests separately for each DMR as the number of loci tested divided by the number of loci in the region, thus approximating a correction for the total number of regions while accounting for region size.

4.2.4 Results

Phylloquinone supplementation responder identification

One hundred and four of the 229 individuals randomized to the treatment group met the conditions of this study (white, consented for genotyping, $\geq 85\%$ adherent to the protocol, with a SD < 2 nmol/L in the four follow up measures of plasma phylloquinone). Using the non-responder criteria derived in the nontreatment group, 40 of the 104 (38%) participants were classified as non-responders. Characteristics of the selected 24 responders and 24 non-responders are described in the Table 1. Baseline concentrations of plasma phylloquinone were not different between responders and non-responders, while a greater proportion of females were categorized as responders in this sample.

Table 1: Population description

	Responders (N=24)	Non-responders (N=24)	p-value
	Mean (SD)	Mean (SD)	
Age	66.8 (5.3)	69.2 (4.3)	0.10
Sex, N, (%F)	17 (70.8)	7 (29.2)	0.009
BMI	28.3 (4.9)	29.5 (6.0)	0.44
Smoker, N, (%Y)	1 (4.1)	0 (0.0)	1.0
Cholesterol-lowering medication use, N (%Y)	9 (37.5)	6 (25.0)	0.53
Baseline plasma TG (mg/dL)	112.9 (52.6)	92.3 (33.0)	0.11
Plasma phylloquinone (nmol/L)			
Baseline	0.48 (0.45)	0.67 (0.45)	0.15
Median 3-year change	4.6 (1.4)	1.5 (0.6)	< 0.001

EWAS for plasma phylloquinone supplementation response

After preprocessing and quality control steps, 785,436 CpG sites were available for analysis. While unadjusted genomic inflation was high, possibly due to technical biases arising from the randomization of responders and non-responders at the microarray chip level, adjustment for 10 control-probe principal components reduced the genomic inflation to a moderate level ($\lambda = 1.21$) for this preliminary analysis (see Methods). Genome-wide analysis revealed two differentially-methylated CpGs (DMCs) surpassing a genome-wide Bonferroni threshold (6.37×10^{-8}) (Table 2; Model 1). Adjustment for biological variables (Models 2 and 3) did not markedly change the results for cg10918016 when accounting for the loss of degrees of freedom, while the strength of the relationship at cg04894169 was moderately attenuated.

Table 2: Top hits from responder EWAS

CpG	Annotations			EWAS P-values		
	Chr	Pos	Gene	Model 1 (estimated β -value difference)	Model 2	Model 3
cg10918016	12	124809934	NCOR2	3.40E-08 (-0.9%)	5.93E-07	8.27E-07
cg04894169	3	14220407	XPC; LSM3	6.01E-08 (+9.5%)	1.61E-05	1.65E-05

* P-values are based on regressions using methylation M-values. β -value differences were estimated using identical regressions other than replacement of M-values with β -values.

Given the sample size, differentially-methylated regions (DMRs) were sought to improve power to detect biological signal. The Comb-p tool was used to call DMRs based on the p-values from Model 1 responder analysis. At a Šidák-corrected regional p-value threshold of 0.05, 80 regions were uncovered. The region with the strongest association was located on chromosome 12 at the *C12orf23* locus (regional $p = 2.99 \times 10^{-8}$). Regions with $p < 1 \times 10^{-5}$ are displayed in Table 3.

As a hypothesis-driven sub-analysis, a set of candidate lipid-related genes were selected for specific assessment. This included 254 CpGs in 6 genes (*APOA1*, *APOA5*, *APOB*, *APOE*, *LPL*, and *NPC1L1*). While this set did not show systematic enrichment for EWAS signal ($p = 0.81$; Kolmogorov-Smirnov test; Supp. Fig. S1), a single CpG site in *NPC1L1* showed a Benjamini-Hochberg false discovery rate of < 0.1 ($p = 2.9 \times 10^{-4}$). Further details on the results for all CpGs in *NPC1L1* as well as local genomic annotations are shown in Figure 1, while EWAS results for nominally significant ($p < 0.05$), lipid-related CpGs are shown in Supp. Table S1.

Table 3: DMRs in responder/non-responder EWAS from Comb-p with uncorrected $p < 1e-5$

Locus	# CpGs	Annotated gene	P	Adj. P
chr12:107348638-107351181	24	C12orf23	2.99E-08	9.24E-06
chr9:469262-470156	7		2.95E-07	2.60E-04
chr14:24778910-24780926	19	LTB4R2; LTB4R; CIDEA	1.33E-06	5.18E-04
chr12:56040024-56040601	6		5.46E-07	7.44E-04
chr7:74070893-74073064	10	GTF2I	2.28E-06	8.24E-04
chr5:100152959-100153624	3	MIR548P; ST8SIA4	1.15E-06	1.35E-03
chr7:27135565-27138974	20	HOXA1	8.58E-06	1.98E-03
chr7:140395229-140397380	23	LOC100134713; NDUFB2	5.46E-06	1.99E-03
chr9:32550954-32553215	14	TOPORS; TOPORS-AS1	8.12E-06	2.82E-03
chr19:47507398-47508119	5	GRLF1	3.05E-06	3.32E-03
chr10:120862958-120864350	9	FAM45B; FAM45A	7.47E-06	4.20E-03
chr1:110052009-110052825	10	AMIGO1	5.30E-06	5.09E-03
chr17:19314298-19314618	6	RNF112	2.25E-06	5.51E-03
chr5:96518720-96519644	12	RIOK2	7.02E-06	5.95E-03
chr12:104359247-104360312	11	TDG	9.66E-06	7.10E-03
chr1:63061967-63062724	4	DOCK7; ANGPTL3	7.23E-06	7.47E-03
chr6:28574470-28574981	8		5.48E-06	8.39E-03
chr1:151965090-151965283	3	S100A10	4.63E-06	1.86E-02
chr6:168580502-168580545	3		4.73E-06	8.28E-02
chr7:56183577-56183584	2	LOC389493	6.31E-06	5.07E-01

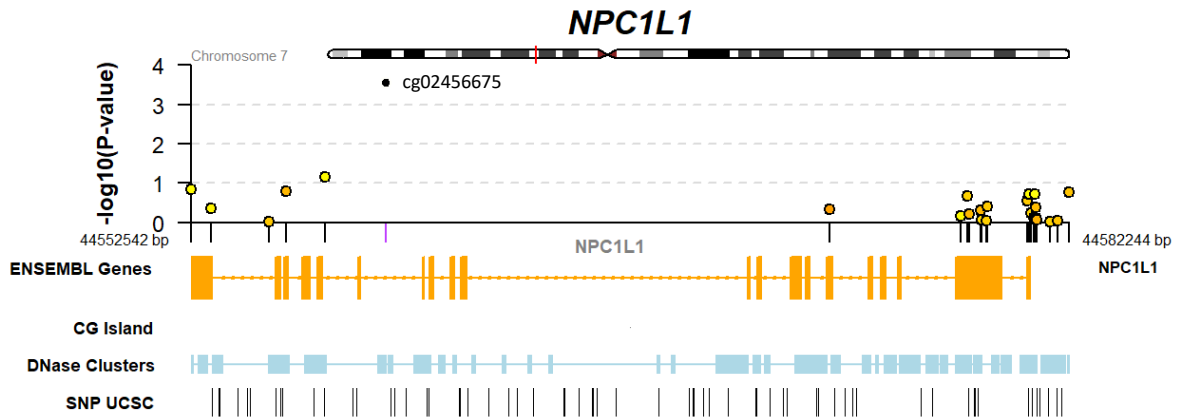


Figure 1: Genomic context of *NPC1L1* responder EWAS results. Upper panel shows a regional Manhattan plot for CpGs within the *NPC1L1* gene. Lower panel includes tracks indicating exonic and intronic regions of the gene, CG islands (none present within this region), DNase hypersensitivity regions, and SNPs retrieved from the UCSC genome browser database.

Since sex was imbalanced between responders and non-responders, its role in the observed relationships was investigated further for top CpGs genome-wide and for *NPC1L1*. Adjustment for sex did not strongly attenuate any of the relationships at these 3 CpGs (Table 2, Supp. Table S1). While visual inspection revealed patterns consistent with sex-specific effects (Supp. Fig. S2), incorporation of an interaction term between responder status and sex into Model 1 did not produce significant results. The available sample size was not sufficient to reasonably enable a stratified analysis.

EWAS for baseline plasma phylloquinone

A secondary EWAS for baseline plasma phylloquinone was performed to detect shared molecular signal between the plasma phylloquinone response to a regular dietary intake of phylloquinone and the response to a supplemental dose for these potentially-related quantities and to provide support for the primary responder analysis. With adjustment for 5 principal components (derived from the M-value matrix) as well as responder status, a genomic inflation in the baseline analysis of $\lambda = 0.99$ was observed. No DMCs were identified as passing a genome-wide Bonferroni threshold (Supplementary Table S2). Sensitivity analyses included two models adding identical variables to those from the responder analysis as well as a third with additional adjustment for estimated phylloquinone intake based on FFQ filled out at baseline. None of the models produced major changes in the results.

Using an analogous process to that for the responder analysis, 48 DMRs were uncovered at Šidák-corrected $p < 0.05$. The strongest baseline DMR was located on chromosome 20 at the *MATN4/RBPJL* locus (regional $p = 1.12 \times 10^{-17}$). No overlapping regions were found to show corrected $p < 0.05$ for both the responder and baseline analyses. Regions with $p < 1 \times 10^{-5}$ are displayed in Table 4.

A comparison of the epigenomic architectures of responder and baseline EWAS results is available in Figure 2. A set of moderate-confidence CpGs was generated by filtering for the group of CpG sites reaching nominal significance ($p < 0.05$) using minimally-adjusted models (Model 1) at both the regional and single-site level in both EWAS. For this set, the Pearson correlation between regression coefficients was low ($\rho = 0.01$, $p = 0.84$). However, when using biologically-plausible-adjusted models (Model 3), this correlation became stronger ($\rho = 0.23$, $p = 9.9 \times 10^{-5}$). With additional adjustment for plasma phylloquinone intake in the baseline analysis (Model 4), this correlation grew further ($\rho = 0.38$, $p = 1.7 \times 10^{-15}$). The set of CpGs with matching signs across fully-adjusted analyses was inputted to a gene enrichment analysis, highlighting biological processes related to cellular metabolism (e.g. nitrogen compound metabolic process; FDR = 1.44×10^{-6}). For comparison, the set of CpGs with non-matching signs returned only a single biological process, “homophilic cell adhesion via plasma membrane adhesion molecules” (FDR = 0.03).

Table 4: Baseline DMR results from Comb-p (regions with uncorrected $p < 1e-5$ shown)

Locus	# CpGs	Annotated gene	P	Adj. P
chr20:43936662-43938184	18	MATN4; RBPJL	1.12E-17	5.81E-15
chr20:62366418-62368565	16	ZGPAT	2.16E-07	7.89E-05
chr11:7041079-7042021	20	ZNF214; NLRP14	4.09E-07	3.41E-04
chr12:131616936-131617481	5	GPR133	3.77E-07	5.44E-04
chr19:39465820-39467435	15	FBXO17	1.25E-06	6.08E-04
chr16:2907516-2908715	13	PRSS22	1.43E-06	9.36E-04
chr19:55280671-55281274	5	KIR2DL1; KIR3DP1	7.94E-07	1.03E-03
chr6:33871421-33872861	9		4.98E-06	2.71E-03
chr15:22833148-22833975	12	TUBGCP5	3.09E-06	2.93E-03
chr22:37770943-37772106	7	ELFN2	4.94E-06	3.33E-03
chr21:22369309-22370864	11	NCAM2	6.80E-06	3.43E-03
chr14:97923970-97925676	11		7.60E-06	3.49E-03
chr12:113902261-113903973	11	LHX5	9.76E-06	4.47E-03
chr16:14402938-14403425	8	MIR365-1	3.04E-06	4.89E-03
chr21:47294738-47295375	4	PCBP3	4.69E-06	5.76E-03
chr2:99388232-99388712	4	LOC101927070	4.47E-06	7.28E-03
chr3:141086819-141087363	6	ZBTB38	6.38E-06	9.16E-03
chr11:44327868-44328423	3	ALX4	6.95E-06	9.78E-03
chr10:80605108-80605565	2		9.67E-06	1.65E-02
chr15:98103934-98104162	3	LOC101927310	5.72E-06	1.95E-02

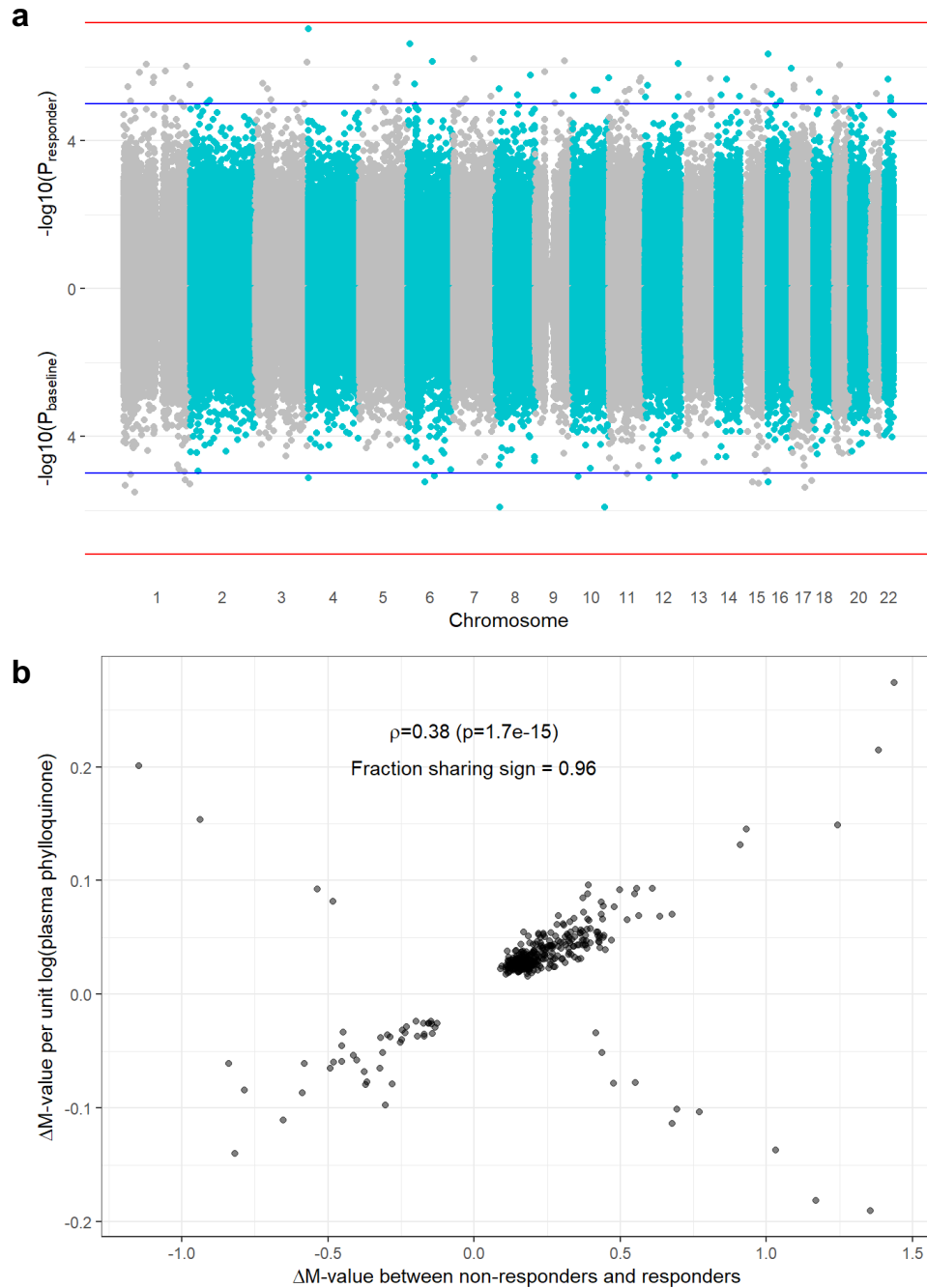


Figure 2: Overlap between responder and baseline EWAS results. *a*) Miami plot showing $-\log_{10}(p\text{-value})$ on the y-axis for the responder analysis (top) and the baseline analysis (bottom). *b*) Comparison of effect sizes (i.e. regression coefficients) between the responder (x-axis) and baseline (y-axis) results. Estimates shown are from fully-adjusted models, thus including adjustment for dietary phylloquinone intake for the baseline analysis. The annotated value is a Pearson correlation (ρ) coefficient between the set of results shown in the plot.

4.2.4 Discussion

Since vitamin K has potential roles in a range of chronic diseases (2), it is important to understand the interindividual variability in and be able to predict the response to dietary and supplemental vitamin K to better identify individuals who could benefit from greater intake. Here, we sought to use epigenomic data to examine the molecular basis for this variability. We found new potential epigenomic markers in whole blood and confirmed associations with known vitamin K biology. We also uncovered a strong correlation between the genomic architecture of the plasma phylloquinone response to supplementation and the phylloquinone concentrations outside of an intervention context. This suggests that the biology involved in phylloquinone metabolism with a typical dietary intake is also relevant when intake is in supplemental quantities.

Our primary analysis focused on epigenomic differences between 24 responders and 24 non-responders. Of single CpGs passing a genome-wide Bonferroni threshold using the minimally-adjusted model, one (in the *NCOR2* gene) was relatively unaffected by further adjustment for biological covariates. *NCOR2* (also known as *SMRT*) is recognized as a gene that codes for a nuclear receptor co-repressor involved in transcriptional silencing of certain genes (28). The protein, NCoR, is one of the most well-studied co-repressors with respect to the vitamin D receptor and also signals alongside the vitamin A-responsive RAR/RXR (29). Though we are not aware of any involvement of NCoR in phylloquinone metabolism, the present findings combined with its connection to fat-soluble vitamin signaling could warrant further investigation. It should be noted that despite the statistical significance of this relationship, the difference may not be biologically meaningful (estimate beta-value difference of 0.9%).

DMR detection was used to enhance statistical power to detect associations due to the aggregation of signal from neighboring CpG sites. The top responder DMR was in an enhancer upstream of the *C12orf23* gene, also known as *TMEM263*. This gene encodes a transmembrane helix protein that is highly and specifically expressed at the RNA and protein levels in the gastrointestinal tract (30). Beyond this mechanistic connection to absorptive processes, *TMEM263* variants have been linked in GWAS to bone mineral density (31), which is known to relate to vitamin K through the presence of vitamin K-dependent osteocalcin in bone and which was the primary outcome of the trial examined here.

Phylloquinone absorption and transport in circulation is known to be consistent with that of many other dietary lipids, involving intestinal uptake of mixed micelles, packaging into chylomicrons, export into the lymph, and lipoprotein transport (32). Age-related increases in plasma phylloquinone have been found to be secondary to increases in plasma triglycerides, and genomic signatures of plasma phylloquinone have been shown to change after adjustment for triglycerides (8). As such, we have previously prioritized lipid-related genes in seeking genetic variants that may modify response to vitamin K supplementation (9). While analysis of CpGs annotated to this subset of genes did not show a systematic enrichment of signal, a particularly strong relationship at cg02456675 in the *NPC1L1* gene was detected. This gene is traditionally known for its role in cholesterol and α -tocopherol transport (33,34). However, in a genetic analysis of the plasma phylloquinone response to supplementation, a SNP in this gene (rs2072183) was shown to be an important

determinant of the inter-individual variability in the response in men (9). Furthermore, there is preliminary evidence that this gene may have a role in the intestinal absorption of phylloquinone (35). While it is possible that this CpG is under genetic control, the SNP in question lies about 20kb upstream in the *NPC1L1* gene and this SNP-CpG combination has no established methylation-quantitative trait locus relationship in the mQTL Database (36). Sex-specificity of this relationship (along with those of the top genome-wide CpG sites) appeared to be present (Supp. Fig. S2), but statistical models testing this interaction did not return positive results.

The top baseline analysis DMR was at the *MATN4/RBPJL* locus, for which relationships with vitamin K metabolism or related phenotypes have not yet been published. Despite this lack of existing evidence, both of these genes are most strongly expressed in pancreas (30), providing a link to digestive processes that could modulate vitamin K absorption. When considering the set of CpGs that were nominally associated with both response and baseline at the regional and site-specific levels (a set more likely to be biologically relevant), we observed a strong correlation between EWAS coefficients for the two analyses. While the median intake among US men and women is near the adequate intake (90 µg/d for women and 120 µg/d for men) (37), phylloquinone intakes have been shown to vary by hundreds of µg/d across populations (38). Additionally, of the fat-soluble nutrients, daily dietary phylloquinone intakes have been shown to display the highest intra-individual variability (3). Thus, a supplemental dose of 500 µg/d is likely to be the dominant source of phylloquinone in supplemented individuals without representing a supra-physiological intake. Accordingly, we posit that the strong correlation of baseline EWAS results with those of the primary responder EWAS both supports the primary findings and suggests that similar molecular machinery is involved in responses to both diet and supplements at these levels of intake.

The preliminary analysis presented here is limited in statistical power due to its low sample size, affording the ability to detect only large differences in DNA methylation. Thus, our findings will need to be replicated in larger cohorts, which will also enable the construction of predictive models to assess the total ability of epigenomic factors to predict interindividual differences in response to phylloquinone supplementation. Additionally, while sex was adjusted for in sensitivity models and did not show significant interactions with responder status in our top hits, the sex imbalance in responders versus non-responders could have impacted the results observed here. The analysis of DNA methylation from blood limits the set of biological signals that can be uncovered; while genetically-controlled and leukocyte-related mechanisms can be plausibly found, the inclusion of other relevant tissues such as intestine could provide complementary information. Finally, much remains to be learned about optimal circulating phylloquinone concentrations. It is known that phylloquinone is the predominant form of vitamin K found in circulation, and phylloquinone concentrations are often compared to rank individuals' vitamin K status in population studies (38). However, the implications of high or low concentrations with respect to functionality in tissues are not yet clear.

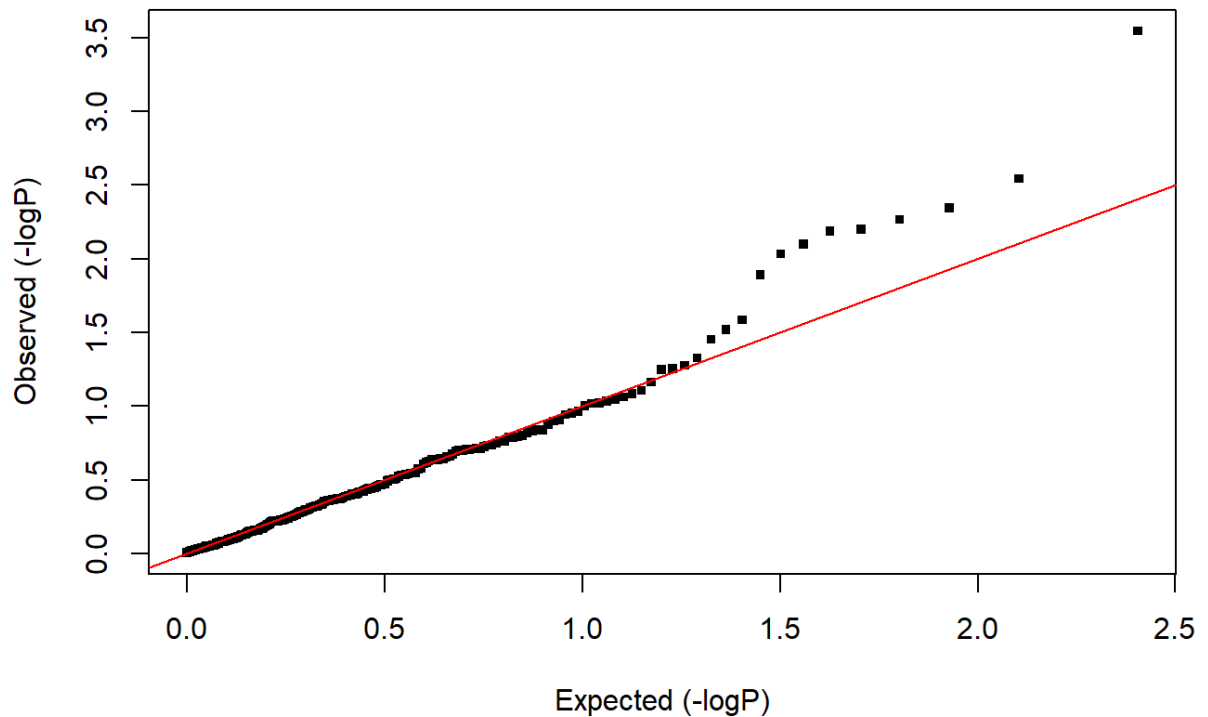
In summary, we present results from what is to our knowledge the first EWAS of vitamin K, both steady-state plasma phylloquinone concentrations and their dynamic changes in response to phylloquinone supplementation. We developed a framework for discriminating

responders and non-responders to phylloquinone supplementation, and showed differences in DNA methylation between these groups at both known and novel loci, highlighting the importance of the *NPC1L1* gene for further investigation. We also demonstrated a strong correlation between the epigenomic architectures of steady-state plasma phylloquinone and its response to supplementation, supporting the findings of the responder analysis and suggesting that similar molecular mechanisms may govern responses to food and supplementation. Taken together, this work can guide research related to the biology of vitamin K and aid the development of more personalized dietary recommendations.

4.2.5 Supplementary Materials

Supp. Table S1: Associations of CpGs in lipid-specific genes with plasma phylloquinone responder status (sites with $p < 0.05$ shown)

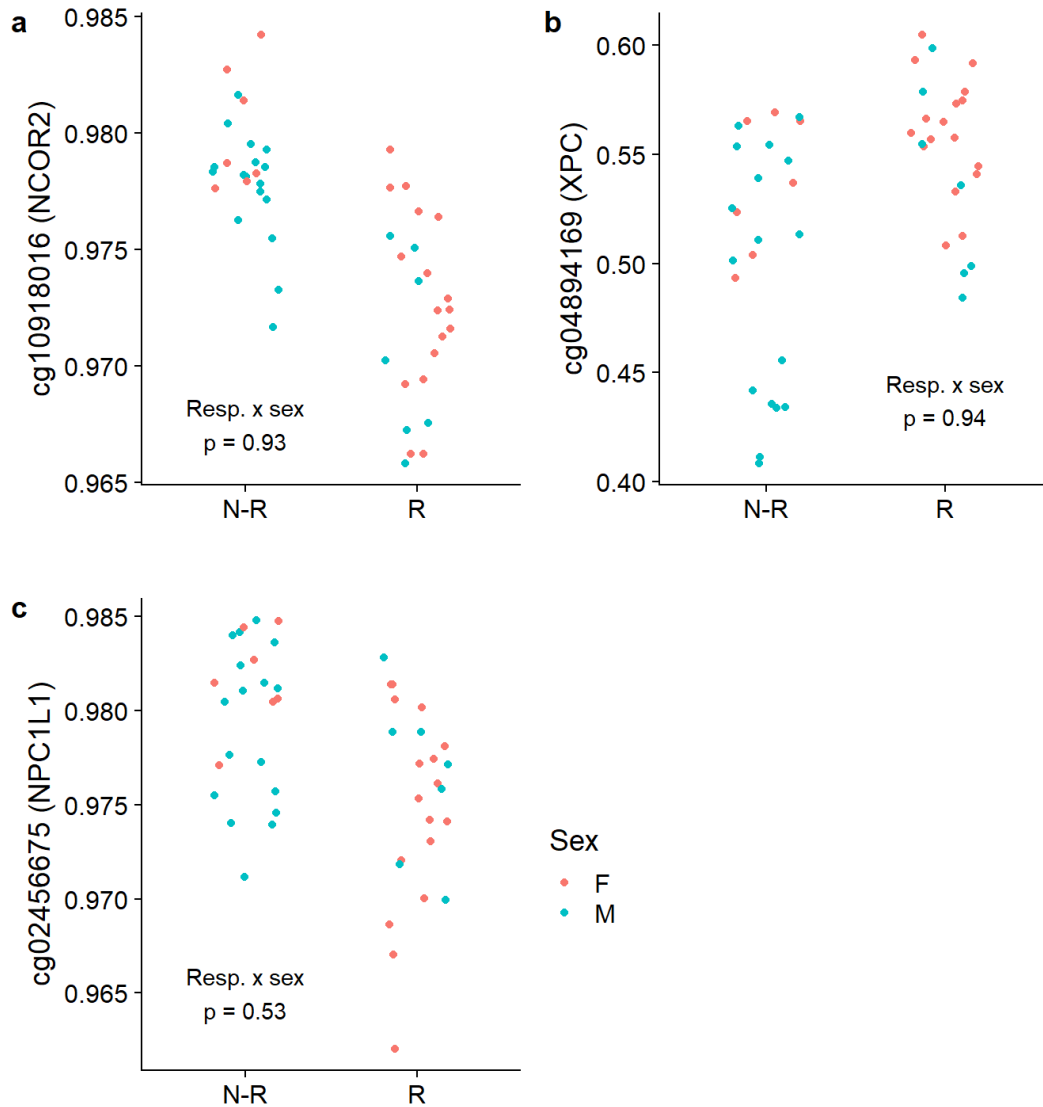
CpG	Annotations			EWAS P-values			
	Chr	Pos	Gene	Model 1	Model 2	Model 3	FDR
cg02456675	7	44559130	NPC1L1	2.85E-04	0.00241	0.002	0.07234
cg09822001	1	156561334	APOA1BP	0.00287	0.00772	0.00982	0.2769
cg12629632	8	19796651	LPL	0.00453	0.00821	0.00957	0.2769
cg08546856	1	183619416	APOBEC4; RGL1	0.00544	0.00931	0.0134	0.2769
cg01643229	11	116706810	APOA1; APOA1-AS	0.00635	0.00223	0.00343	0.2769
cg23117929	1	156561640	APOA1BP	0.00654	0.0088	0.011	0.2769
cg00359511	16	28509413	APOB48R	0.00796	0.00587	0.00823	0.2888
cg02051562	1	183616450	APOBEC4; RGL1	0.00931	0.0172	0.0226	0.2956
cg08124288	8	19796752	LPL	0.0129	0.0304	0.0321	0.3641
cg20579480	1	183622007	APOBEC4; RGL1	0.0261	0.0773	0.095	0.6630
cg18381283	1	156561353	APOA1BP	0.0303	0.0315	0.0363	0.6997
cg26000393	22	39378224	APOBEC3B	0.0354	0.0418	0.0527	0.7493
cg19324627	11	116708329	APOA1	0.0472	0.0312	0.0401	0.9017



Supp. Fig. S1: Q-Q plot for hypothesis-driven lipid-related CpG sites. The distribution of p -values from the responder EWAS (Model 1) is plotted against the uniform p -value distribution expected under the null hypothesis.

Supp. Table S2: Suggestive baseline EWAS results ($p < 1e-5$)

CpG	Annotations			EWAS P-values		
	Chr	Position	Gene	Model 1	Model 2	Model 3
cg00621646	11	123849141	OR10S1	8.38E-06	2.53E-05	3.81E-05
cg02006488	6	65443526	EYS	1.71E-07	2.68E-06	3.28E-06
cg02296029	1	66169873		9.19E-06	5.80E-05	8.84E-05
cg02766058	18	8500024		7.35E-06	6.81E-05	1.03E-04
cg03413355	15	56537579		7.41E-06	1.35E-05	2.11E-05
cg03982250	8	127337536	LOC101927657	5.44E-06	5.21E-05	7.72E-05
cg12155376	1	40723612	ZMPSTE24	5.23E-06	3.02E-06	4.57E-06
cg12753515	6	160699120		3.46E-06	2.70E-05	2.02E-05
cg20689135	13	113764225	F7	7.49E-06	1.43E-04	1.86E-04
cg21452219	2	28918652		2.69E-06	1.68E-05	7.45E-06



Supp Fig. S2: Beta-value plots for top CpGs. Scatter plots display beta-values (inverse logit-transformed M-values) as a function of responder status for top CpG sites from either the overall analysis (NCOR2 (a) and XPC (b)) or the lipid-specific analysis (NPC1L1 (c)). P-values shown are from minimally-adjusted EWAS models incorporating a responder-by-sex interaction term. N-R: non-responder; R: responder.

4.2.6 References

1. Booth SL. Roles for Vitamin K Beyond Coagulation. *Annu Rev Nutr.* 2009;29(1):89–110.
2. Harshman SG, Shea MK. The Role of Vitamin K in Chronic Aging Diseases: Inflammation, Cardiovascular Disease, and Osteoarthritis. *Curr Nutr Rep.* 2016;5(2):90–8.
3. Booth SL, Tucker KL, McKeown NM, Davidson KW, Dallal GE, Sadowski JA. Relationships between Dietary Intakes and Fasting Plasma Concentrations of Fat-Soluble Vitamins in Humans. *J Nutr.* 1997;127(4):587–92.
4. Jones KS, Bluck LJC, Wang LY, Coward WA. A stable isotope method for the simultaneous measurement of vitamin K1 (phyloquinone) kinetics and absorption. *Eur J Clin Nutr.* 2008 Nov;62(11):1273–81.
5. Shea MK, Benjamin EJ, Dupuis J, Massaro JM, Jacques PF, D’Agostino RB, et al. Genetic and non-genetic correlates of vitamins K and D. *Eur J Clin Nutr.* 2009;63(4):458–64.
6. Shea MK, Booth SL, Nettleton JA, Burke GL, Chen H, Kritchevsky SB. Circulating Phylloquinone Concentrations of Adults in the United States Differ According to Race and Ethnicity. *J Nutr.* 2012;142(6):1060–6.
7. Borel P, Desmarchelier C. Bioavailability of Fat-Soluble Vitamins and Phytochemicals in Humans: Effects of Genetic Variation. *Annu Rev Nutr.* 2018 Aug 21;38(1):69–96.
8. Dashti HS, Shea MK, Smith CE, Tanaka T, Hruby A, Richardson K, et al. Meta-analysis of genome-wide association studies for circulating phylloquinone concentrations. *Am J Clin Nutr.* 2014;100(6):1462–9.
9. Kelly JM, Ordovas JM, Matuszek G, Smith CE, Huggins GS, Dashti HS, et al. The contribution of lipids in vitamin K biomarker response to vitamin K supplementation. *Manuscr Submitt Publ Mol Nutr Food Res.* 2019;
10. Shah S, Bonder MJ, Marioni RE, Zhu Z, McRae AF, Zhernakova A, et al. Improving Phenotypic Prediction by Combining Genetic and Epigenetic Associations. *Am J Hum Genet.* 2015 Jul;97(1):75–85.
11. Feinberg AP, Irizarry RA, Fradin D, Aryee MJ, Murakami P, Aspelund T, et al. Personalized Epigenomic Signatures That Are Stable Over Time and Covary with Body Mass Index. *Sci Transl Med.* 2010 Sep 15;2(49):49ra67–49ra67.
12. Amatu A, Sartore-Bianchi A, Moutinho C, Belotti A, Bencardino K, Chirico G, et al. Promoter CpG Island Hypermethylation of the DNA Repair Enzyme MGMT Predicts Clinical Response to Dacarbazine in a Phase II Study for Metastatic Colorectal Cancer. *Clin Cancer Res.* 2013 Apr 15;19(8):2265–72.
13. Campión J, Milagro FI, Goyenechea E, Martínez JA. TNF- α Promoter Methylation as a Predictive Biomarker for Weight-loss Response. *Obesity.* 2009;
14. Booth SL, Dallal G, Shea MK, Gundberg C, Peterson JW, Dawson-Hughes B. Effect of Vitamin K Supplementation on Bone Loss in Elderly Men and Women. *J Clin Endocrinol Metab.* 2008 Apr;93(4):1217–23.

15. Davidson KW, Sadowski JA. Determination of vitamin K compounds in plasma or serum by high-performance liquid chromatography using postcolumn chemical reduction and fluorimetric detection. In: *Methods in Enzymology*. Elsevier; 1997. p. 408–21.
16. Pidsley R, Wong CC, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics*. 2013;14(1):293.
17. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*. 2014 May 15;30(10):1363–9.
18. Fortin J-P, Triche TJ, Hansen KD. Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics*. 2016 Dec 28;btw691.
19. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics*. 2013 Jan 15;29(2):189–96.
20. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*. 2012 Dec;13(1):86.
21. Lehne B, Drong AW, Loh M, Zhang W, Scott WR, Tan S-T, et al. A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies. *Genome Biol*. 2015 Dec;16(1):37.
22. R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
23. Lamon-Fava S, Sadowski JA, Davidson KW, O'Brien ME, McNamara JR, Schaefer EJ. Plasma lipoproteins as carriers of phyloquinone (vitamin K1) in humans. *Am J Clin Nutr*. 1998;67(6):1226–31.
24. Erkkilä AT, Lichtenstein AH, Dolnikowski GG, Grusak MA, Jalbert SM, Aquino KA, et al. Plasma transport of vitamin K in men using deuterium-labeled collard greens. *Metabolism*. 2004;53(2):215–21.
25. Pederson BS, Schwartz DA, Yang IV, Kechris KJ. Comb-p: software for combining, analyzing, grouping and correcting spatially correlated P-values. *Bioinformatics*. 2012;28(22):2986–8.
26. Kechris KJ, Biehs B, Kornberg TB. Generalizing Moving Averages for Tiling Arrays Using Combined P-Value Statistics. *Stat Appl Genet Mol Biol*. 2010;9(1).
27. Sidak Z. Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. *J Am Stat Assoc*. 1967;62(318):626–33.
28. Pei L, Leblanc M, Barish G, Atkins A, Nofsinger R, Whyte J, et al. Thyroid hormone receptor repression is linked to type I pneumocyte-associated respiratory distress syndrome. *Nat Med*. 2011 Nov;17(11):1466–72.
29. Teske KA, Yu O, Arnold LA. Inhibitors for the Vitamin D Receptor–Coregulator Interaction. In: *Vitamins & Hormones*. Elsevier; 2016. p. 45–82.

30. Uhlen M, Zhang C, Lee S, Sjöstedt E, Fagerberg L, Bidkhori G, et al. A pathology atlas of the human cancer transcriptome. *Science*. 2017;357.
31. Kichaev G, Bhatia G, Loh P-R, Gazal S, Burch K, Freund MK, et al. Leveraging Polygenic Functional Enrichment to Improve GWAS Power. *Am J Hum Genet*. 2019;104:65–75.
32. Shearer MJ, Fu X, Booth SL. Vitamin K Nutrition, Metabolism, and Requirements: Current Concepts and Future Research. *Adv Nutr*. 2012 Mar 1;3(2):182–95.
33. Davis HR, Zhu L, Hoos LM, Tetzloff G, Maguire M, Liu J, et al. Niemann-Pick C1 Like 1 (NPC1L1) Is the Intestinal Phytosterol and Cholesterol Transporter and a Key Modulator of Whole-body Cholesterol Homeostasis. *J Biol Chem*. 2004 Aug 6;279(32):33586–92.
34. Narushima K, Takada T, Yamanashi Y, Suzuki H. Niemann-Pick C1-Like 1 Mediates - Tocopherol Transport. *Mol Pharmacol*. 2008 Apr 24;74(1):42–9.
35. Takada T, Yamanashi Y, Konishi K, Yamamoto T, Toyoda Y, Masuo Y, et al. NPC1L1 is a key regulator of intestinal vitamin K absorption and a modulator of warfarin therapy. *Sci Transl Med*. 2015 Feb 18;7(275):275ra23-275ra23.
36. Gaunt TR, Shihab HA, Hemani G, Min JL, Woodward G, Lyttleton O, et al. Systematic identification of genetic influences on methylation across the human life course. *Genome Biol*. 2016 Dec;17(1):61.
37. Harshman SG, Finnan EG, Barger KJ, Bailey RL, Haytowitz DB, Gilhooly CH, et al. Vegetables and Mixed Dishes Are Top Contributors to Phylloquinone Intake in US Adults: Data from the 2011-2012 NHANES. *J Nutr*. 2017 Jul 1;147(7):1308–13.
38. Shea M, Booth S. Concepts and Controversies in Evaluating Vitamin K Status in Population-Based Studies. *Nutrients*. 2016;8(1):8.

Chapter 5

Summary and Discussion

5.1 Summary of results

This thesis consists of a series of investigations intended to further develop our understanding of the epigenomic basis for cardiovascular disease (CVD), the use of multi-omic metrics to improve risk prediction, and the use of epigenomics and genomics to resolve inter-individual differences in cardiometabolic risk. In Chapter 2, a module- and region-based epigenome-wide association study was undertaken to provide insights into cardiovascular pathophysiology and the cumulative effects of CVD risk factor exposure over time. In Chapter 3, an epigenomic risk modeling procedure extended these findings to quantify the accuracy of blood-based DNA methylation in predicting clinical CVD risk, introducing a novel domain for use of ensemble models for omics-based risk prediction, and providing proof-of-concept for the measurement of DNA methylation as a synergistic risk factors alongside genotypes and traditional clinical risk factors. Chapter 4 moved beyond direct risk prediction, investigating genomic and epigenomic signatures of diet response with implications for cardiometabolic health-focused personalized nutrition.

Chapter 2 involved an investigation of CVD-related epigenomics while attempting to overcome some of the common obstacles faced by epigenome-wide association studies. Typical cross-sectional studies of CVD are plagued by reverse causation because they cannot determine unambiguously whether any uncovered signal is a cause or consequence of disease. Furthermore, while differentially-methylated regions are a common approach to boosting signal, module-based techniques such as WGCNA allow for potentially more interpretable features that can represent holistic epigenomic programs that are predictive of disease. The inclusion of these two approaches allowed us to discover multiple novel signals predictive of incident CVD and in supportive of known CVD pathophysiology, such as differential methylation near immune function-related genes. We further capitalized on the longitudinal nature of the FHS dataset, discovering a previously unobserved association between cumulative CVD risk factor exposure and methylation of an immune-related module. This provides proof-of-concept for the use of DNA methylation as a “molecular archive” of cumulative risk factors over the life course (Heijmans et al. 2009), already established to be an independent risk factor for CVD but not always measurable without consistent, longitudinal risk factor measurements.

Chapter 3 expanded on the findings of the previous chapter in seeking to directly optimize a predictor of CVD risk, with the output constituting a methylation-based CVD risk score (MRS). By directly seeking predictive power for incident CVD events, this investigation was able to explore the existing limits of epigenomics for CVD risk prediction in comparison to (and alongside) risk scores from other dimensions of biology. We found that an MRS model trained across a series of longitudinal cohorts was able to add predictive power beyond both traditional and genomics-based risk scores. We also evaluated a recently-introduced “cross-study learning” method for training ensemble models in omic datasets, observing comparable performance to a standard analysis pipeline (simple merging of datasets followed by “batch” (i.e. study) adjustment using ComBat (Johnson, Li, and Rabinovic 2007)). In a preliminary analysis of interactions between the MRS and other CVD risk metrics based on traditional risk factors and genomics, we found that the MRS appears to perform better, as evidenced by higher hazard ratio estimates in held-out samples, in lower-risk regions of these alternative metrics. This pattern was especially apparent for traditional risk factors (as represented by the Framingham Risk Score for

generalized CVD). This finding suggests that epigenomic quantities may be useful as a complementary data source for clinical implementation risk stratification, by identifying individuals with “hidden” risk that would not otherwise have been prioritized for preventative treatment.

While the use of overlapping cohorts in Chapters 2 and 3 would suggest that similar epigenomic patterns would be found, we observed consistency across analytical approaches of the finding of biological signal related to immune function, cellular adhesion, and monocyte activation. This observation is biologically plausible, as DNA methylation measured in whole blood is tracking almost exclusively the methylation of DNA from leukocytes. Furthermore, monocyte activation is a recognized pathway leading to CVD risk in mammals, for which epigenetics is thought to play an important role (Short et al. 2017). Given the finding from Chapter 2 of the strong correlation between an immune-related methylation module and cumulative risk factor exposure, these cumulative exposures over time may result in greater monocyte “priming”, both mechanistically explaining and statistically predicting an increased risk of CVD. However, it should be noted that we did not observe evidence of mediation of the cumulative risk factor association with CVD by epigenomic module activation.

Chapter 4 focused on moving towards biomarkers of diet response that could be used to decrease disease risk in individuals prioritized as high-risk based on the approaches pursued in the previous Chapters. Through the use of genome-wide interaction study (GWIS) results, we developed a series of genome-wide genetic scores for response to dietary fat. When testing these scores in the fat reduction-focused Dietary Modification trial from the Women’s Health Initiative, only one of six scores (LDL-cholesterol (LDL-C) fat reduction score, or LDL-FRS) showed meaningful prediction of one-year changes in the corresponding cardiometabolic risk factor (CRF). While the poor performance of most of the response scores indicates that improved datasets and methodology for diet response prediction are necessary, the LDL-FRS was shown to explain a modest proportion of the variance in LDL-C changes while highlighting loci with potential involvement in the biological response to dietary fat. Further longitudinal investigation of the LDL-FRS revealed suggestive interaction patterns between DM trial control and intervention arm participants, but no definitive associations with long-term chronic disease outcomes. The results lay the foundation for the combination of many genome-wide gene-diet interactions for diet response prediction while highlighting the need for further research in order to achieve robust biomarkers for use in personalized nutrition.

An additional preliminary study described in Chapter 4 explored the epigenomic signature of response to vitamin K intake. After categorizing vitamin K responders and non-responders based on their changes in plasma phylloquinone during a supplementation trial, an epigenome-wide association study was performed to test differences between these groups. Differential DNA methylation was identified at multiple known and novel loci with respect to phylloquinone absorption and metabolism. The results were supported by a second EWAS for baseline plasma phylloquinone concentrations, which showed a strong correlation between the epigenomic signatures of phylloquinone baseline status and response to supplementation.

5.2 Implications

The body of work developed here, while not immediately clinically applicable, provides important proof-of-concept for a series of different lines of inquiry related to epigenomics and diet and their role in modifying CVD risk.

In order to implement a comprehensive system for CVD prevention, it is necessary to understand how to both effectively predict risk and prescribe appropriate actions to reduce that risk. Compared to many other diseases, CVD risk prediction in adults is quite robust. Predictive risk factors have been identified from many types of data sources, including demographics, anthropometrics, blood lipids, imaging, genomics, and multi-omics. However, to take advantage of these metrics in a holistic manner, it is important to understand how emerging data types (e.g. epigenomics) can provide complementary information for risk prediction, and how they may be used synergistically with other established risk factors. This work contains a preliminary analysis of interactions between epigenomic risk and metrics based on traditional risk factors (FRS) or genetic variants (GRS), finding that epigenomic risk may be helpful in identifying higher-risk individuals that would otherwise not be detected by other metrics. Given further evaluation of these effects in larger samples, this research may inform the incorporation of epigenomic measurements into clinical practice as a secondary method for screening otherwise low-risk individuals.

As discussed in more detail in Chapter 1, CVD develops as a downstream result of a series of inter-related biological processes. If a given individual is identified as being at risk for the development of CVD, pharmaceutical and lifestyle-based approaches are available for reducing that risk by targeting one or many of these pathways and processes. The epigenomic investigations presented here, especially the work in Chapter 2, helps to elucidate some of these biological mechanisms in ways that reinforce known biology and suggest new approaches to modifying these risk pathways. The confirmation of monocyte-related biology, especially that related to monocyte priming for response to biological insults, supports known mechanisms of immune activation and monocyte infiltration into the endothelial wall that contribute to the development of atherosclerosis. The recent success of immune-related therapies, such as canakinumab as an anti-inflammatory drug, for CVD supports the relevance of these mechanisms (Ridker et al. 2017). Furthermore, any relationships of this monocyte priming to DNA methylation-based mechanisms, as has been previously observed for other epigenetic quantities such as histone modifications, could suggest the possibility of new immune-focused therapies that could break the link between cumulative risk factor exposure and downstream CVD risk. Indeed, targeted DNA methylation-directed therapies are being enabled by advances in CRISPR-based technologies (Liu et al. 2016). However, it is important to note that the work done here did not support the role of DNA methylation in blood as a mediator between this risk factor exposure over time and the downstream increases in CVD. Thus, there is more to understand about the mechanisms and tissue-specificity of these epigenetic signals before this knowledge can be translated into therapies.

Our observation of links between cumulative risk factor exposure and CVD risk reinforces the importance of maintaining lower risk factor levels throughout the life course, rather than only in later life. As diet is known to be a critical modifiable behavior affecting these

risk factors (including body weight, blood lipids, and inflammation), it is important to resolve inter-individual differences in the response of these risk factors to dietary inputs. The diet response work presented in this thesis details preliminary approaches and provides proofs-of-concept for using molecular data to predict response to dietary inputs. As an example, Chapter 3 highlighted LDL-C as one of a series of CRFs whose cumulative values over time associated more strongly than present values with a CVD-associated epigenomic module. Chapter 4's exploration of the genetic underpinnings of the LDL-C response to dietary fat represents a step towards actionable and targeted strategies for the maintenance of low LDL-C levels. The integration of genomics and epigenomics, each of which were explored in Chapter 4, could in the near future create a powerful combination for the prediction of diet response in a way that may evolve over the life course. It should be noted that the effect of diet on health is complex, and that any failure of diet to control CRF levels on a population level is due to an interaction between these molecular underpinnings and behavioral factors (Spahn et al. 2010).

Lastly, risk or dietary response biomarkers would be implemented into clinical practice in the context of the complex existing systems for risk assessment, treatment, and behavior modification. Acknowledging this complexity, part of the research reported in Chapter 3 focuses on the intersection between epigenetic risk scores and other existing metrics of CVD risk. As mentioned above, we identified a potential synergy between traditional risk metrics and the epigenetic score for identifying hidden high-risk individuals. Furthermore, these multiple types of prognosis biomarkers can be used most effectively in conjunction with response biomarkers to put preventative measures in place at earlier timepoints in life. The present exploration of genotype-based diet response biomarkers moves towards enabling doctors and dieticians to more precisely weigh the expected benefits of a dietary change alongside the pharmaceutical and behavioral context that is already considered.

5.3 Methodological considerations and future directions

The main cardiovascular phenotype investigated in Chapters 2 and 3 had a broad definition including multiple subtypes of CVD; namely, most major manifestations other than heart failure. This was chosen as a compromise between biological homogeneity and sample size, but movement in either direction along this gradient could provide additional value. For example, investigation specifically of coronary heart disease or stroke individually could identify CVD subtype-specific epigenomic signals. Additionally, we chose to create epigenomic models agnostic to prior CVD events. While this was chosen consciously in order to create the most broadly-applicable CVD risk predictor, a distinction between primary and secondary prevention may be important for some subset of signals.

All investigations composing this body of research would benefit from expansion into more diverse populations. Sex was treated here primarily as a confounder, but known differences in CVD manifestation across sexes (Regitz-Zagrosek and Kararigas 2017) suggest that there could be value in performing stratified analyses to look for explicit differences in genomic or epigenomic signatures. Initial indications of such sex-specific interactions were apparent in the vitamin K responder analysis describe in Chapter 4. Integrated analysis of male-only cohorts along with the mixed and female-only cohorts used here would provide an opportunity to expand and strengthen these results considerably. Furthermore, ancestry

is well-recognized as being a critical factor in GWAS and polygenic risk score creation, as different populations have highly heterogeneous correlation structures (linkage disequilibrium) within chromosomes. This led to the focus on individuals of European ancestry in Chapter 4, but expansion into other populations could provide new biological insights and more accurate diet response scores. Ancestry is also a meaningful contributor to inter-individual variation in DNA methylation, although these associations can be confounded by the environmental impacts represented by cultural ethnicity (Galanter et al. 2017). The epigenomic analysis of WHI participants here included all ancestries together, but future investigations could parse any effect modification of ancestry on identified associations.

The biological analysis underlying the genomic and epigenomic datasets used in this work limit the breadth of conclusions that can be drawn. All epigenomic datasets were derived from blood, meaning that their primary interpretation must be related to the function of leukocytes. Though leukocytes, especially monocytes, do play critical roles in the pathogenesis of CVD, data from other tissues (e.g. artery endothelial cells) could provide information related to complementary disease processes. While several manuscripts have shown proof-of-concept that blood-based methylation can act as a proxy for other tissues (e.g. between blood and pancreatic beta cells (Bacos et al. 2016)), this is not well-established in general, with several studies showing poor cross-tissue consistency (e.g. between blood and brain (Davies et al. 2012; Ma et al. 2014)).

Additionally, both the genomic and epigenomic quantities analyzed here were based on microarray technologies. While providing accurate and cost-effective measurements for large-scale cohorts, these platforms only capture part of the total information. For DNA methylation, the Illumina HumanMethylation450 array captures only 1-2% of the CpG sites in the human genome. A sequencing-based technique such as reduced-representation bisulfite sequencing (RRBS) or whole-genome bisulfite sequencing (WGBS) could capture a much greater fraction of the CpG sites in the body, enabling more effective differentially-methylated region detection and potentially uncovering important loci that otherwise would have gone unmeasured. For genotyping, most arrays supporting the data used in Chapters 3 and 4 measured between 500,000 and 1 million variants, allowing imputation to about 10 million variants with reasonable confidence. While this does allow for genome-wide analysis of common variants, sequencing technologies may provide a critical boost to discovery here as well. Specifically, whole-genome sequencing (WGS) can uncover rare variants that have been subject to less selective pressure over time and thus can show much larger effect sizes, an intuition which should translate equally well from main-effect GWAS to studies of gene-environment interaction.

With respect specifically to the gene-diet interaction (GDI) analyses, the imprecision and biases inherent in food frequency questionnaire (FFQ)-based dietary measurement represent a critical limitation to the efficacy of GDI-based diet response score development. While FFQs have been shown to generally rank dietary intakes successfully, their accuracy has been questioned by a series of studies, including some performed in the Women's Health Initiative study (Tucker et al. 2013). GDI-based methods such as the one used in chapter 4 require only summary statistics and so could be developed using results from cohorts with alternate and complementary dietary measurement methods such as 24-hour recall. However, the benefits of such a strategy must be weighed against any potential

heterogeneity between cohorts that would induce a population-specificity to observed gene-diet interactions.

Beyond these biological and experimental considerations, there are opportunities for meaningfully extending the statistical analyses performed here. One of these relates to the methylation module-finding algorithm used in Chapter 2 (WGCNA), which was originally created for use with gene expression data. While it has been adapted to other data types here and elsewhere, methylation-specific optimizations could be applied to make it maximally relevant. For example, consideration of the genomic location of CpG sites could indicate the expected direction of association with gene expression (e.g. promoter CpGs being negatively associated vs. gene body CpGs being positively associated). Incorporating these directional associations would allow more effective “signed” clustering and thus more biologically informative modules. In Chapter 3, the cross-study learning method could be expanded to add more diverse sets of learning algorithms per study and to enable calibration of performance for absolute risk prediction. There are also many opportunities to further develop the statistical approaches to GDI score creation from Chapter 4. Altering the dietary phenotype definition (for dietary fat) to a continuous or more refined measure could highly impact the results, and there is an opportunity to use more sophisticated algorithms for responder score development that incorporate genotypic main effects.

Finally, all of the research performed here would benefit from expansion and replication in larger datasets. This would allow more robust methylation module identification, more confident assignment of clinical value to epigenomic risk scores, and more powerful assessment of gene-diet interactions for use in diet response score development. Such efforts would help this research form the foundation for more targeted and effective dietary approaches to the prevention of cardiovascular disease.

5.4 References

Bacos, Karl, Linn Gillberg, Petr Volkov, Anders H Olsson, Torben Hansen, Oluf Pedersen, Anette Prior Gjesing, et al. 2016. “Blood-based biomarkers of age-associated epigenetic changes in human islets associate with insulin secretion and diabetes.” *Nature Communications* 7. Nature Publishing Group: 11089. <https://doi.org/10.1038/ncomms11089>.

Davies, Matthew N, Manuela Volta, Ruth Pidsley, Katie Lunnon, Abhishek Dixit, Simon Lovestone, Cristian Coarfa, et al. 2012. “Functional annotation of the human brain methylome identifies tissue-specific epigenetic variation across brain and blood.” *Genome Biology* 13 (6): R43. <https://doi.org/10.1186/gb-2012-13-6-r43>.

Galanter, Joshua M, Christopher R Gignoux, Sam S Oh, Dara Torgerson, Maria Pino-Yanes, Neeta Thakur, Celeste Eng, et al. 2017. “Differential methylation between ethnic sub-groups reflects the effect of genetic ancestry and environmental exposures.” *eLife* 6 (January). <https://doi.org/10.7554/eLife.20532>.

Heijmans, Bastiaan T., Elmar W. Tobi, L.H. Lumey, and P. Eline Slagboom. 2009. “The epigenome: Archive of the prenatal environment.” *Epigenetics* 4 (8): 526–31. <https://doi.org/10.4161/epi.4.8.10265>.

- Johnson, W. Evan, Cheng Li, and Ariel Rabinovic. 2007. "Adjusting batch effects in microarray expression data using empirical Bayes methods." *Biostatistics* 8 (1). Oxford University Press: 118–27. <https://doi.org/10.1093/biostatistics/kxj037>.
- Liu, X. Shawn, Hao Wu, Xiong Ji, Yonatan Stelzer, Xuebing Wu, Szymon Czauderna, Jian Shu, Daniel Dadon, Richard A. Young, and Rudolf Jaenisch. 2016. "Editing DNA Methylation in the Mammalian Genome." *Cell* 167 (1): 233–247.e17. <https://doi.org/10.1016/j.cell.2016.08.056>.
- Ma, Baoshan, Elissa H. Wilker, Saffron A. G. Willis-Owen, Hyang-Min Byun, Kenny C. C. Wong, Valeria Motta, Andrea A. Baccarelli, et al. 2014. "Predicting DNA methylation level across human tissues." *Nucleic Acids Research* 42 (6): 3515–28. <https://doi.org/10.1093/nar/gkt1380>.
- Regitz-Zagrosek, Vera, and Georgios Kararigas. 2017. "Mechanistic Pathways of Sex Differences in Cardiovascular Disease." *Physiological Reviews* 97 (1): 1–37. <https://doi.org/10.1152/physrev.00021.2015>.
- Ridker, Paul M, Brendan M Everett, Tom Thuren, Jean G MacFadyen, William H Chang, Christie Ballantyne, Francisco Fonseca, et al. 2017. "Antiinflammatory Therapy with Canakinumab for Atherosclerotic Disease." *New England Journal of Medicine* 377 (12): 1119–31. <https://doi.org/10.1056/NEJMoa1707914>.
- Short, John D., Sina Tavakoli, Huynh Nga Nguyen, Ana Carrera, Chelbee Farnen, Laura A. Cox, and Reto Asmis. 2017. "Dyslipidemic Diet-Induced Monocyte 'Priming' and Dysfunction in Non-Human Primates Is Triggered by Elevated Plasma Cholesterol and Accompanied by Altered Histone Acetylation." *Frontiers in Immunology* 8 (August). Frontiers: 958. <https://doi.org/10.3389/fimmu.2017.00958>.
- Spahn, Joanne M., Rebecca S. Reeves, Kathryn S. Keim, Ida Laquatra, Molly Kellogg, Bonnie Jortberg, and Nicole A. Clark. 2010. "State of the Evidence Regarding Behavior Change Theories and Strategies in Nutrition Counseling to Facilitate Health and Food Behavior Change." *Journal of the American Dietetic Association* 110 (6): 879–91. <https://doi.org/10.1016/j.jada.2010.03.021>.
- Tucker, Katherine L., Caren E. Smith, Chao-Qiang Lai, and Jose M. Ordovas. 2013. "Quantifying Diet for Nutrigenomic Studies." *Annual Review of Nutrition* 33 (1): 349–71. <https://doi.org/10.1146/annurev-nutr-072610-145203>