MonsterLM: A method to estimate the variance explained by genome-wide interactions with environmental factors

By Mohammad Khan,

A Thesis Submitted to the School of Graduate Studies in the Partial Fulfillment of the Requirements for the Degree Master of Science

McMaster University

Master of Science (2020)

Hamilton, Ontario (Department of Statistics)

TITLE: MonsterLM: A method to estimate the variance explained by genome-wide interactions with environmental factors

AUTHOR: Mohammad Khan (McMaster University)

SUPERVISOR: Dr. Guillaume Pare

NUMBER OF PAGES: ix, 76

Abstract

Estimations of heritability and variance explained due to environmental exposures and interaction effects help in understanding complex diseases. Current methods to detect such interactions rely on variance component methods. These methods have been necessary due to the m » n problem, where the number of predictors (m) vastly outnumbers the number of observations (n). These methods are all computationally intensive, which is further exacerbated when considering gene-environment interactions, as the number of predictors increases from m to 2m+1 in the case of a single environmental exposure. Novel methods are thus needed to enable fast and unbiased calculations of the variance explained (R^2) for gene-environment interactions in very large samples on multiple traits. Taking advantage of the large number of participants in contemporary genetic studies, we herein propose a novel method for continuous trait R2 estimates that are up to 20 times faster than current methods. We have devised a novel method, monsterlm, that enables multiple linear regression on large regions encompassing tens of thousands of variants in hundreds of thousands of participants. We tested monsterlm with simulations using real genotypes from the UK Biobank. During simulations we verified the properties of monsterlm to estimate the variance explained by interaction terms. Our preliminary results showcase potential interactions between blood biochemistry biomarkers such as HbA1c, Triglycerides and ApoB with an environmental factor relating to obesity-related lifestyle factor: Waist-hip Ratio (WHR). We further investigate these results to reveal that more than 50% of the interaction variance calculated can be attributed to ∼5% of the single-nucleotide polymorphisms (SNPs) interacting with the environmental trait. Lastly, we showcase the impact of interactions on improving polygenic risk scores.

Acknowledgements

I'd like to thank first and foremost my thesis supervisor: Dr Guillaume Pare, for his guidance and support throughout my time spent in his lab. I'd like to thank Shihong Mao, Walter Nelson and Shuang Di for their support in developing tools that allowed me to work on my thesis project. I'd also like to thank Dr. Angelo Canty for his guidance in reviewing my simulation protocol. Finally, I'd also like to thank my fellow colleagues in the Math and Statistics department for their support and friendship that helped greatly in completing this thesis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

The genome is often the conduit through which environmental exposures convey their effects on health and disease. In understanding complex disease, there exists sensitivities to certain environmental exposures and pharmacotherapies that some people experience and others do not. It's hypothesized to be governed by genetic factors that interact with these exposures to determine risk. Sometimes, these genetic variations may be beneficial, allowing the individual to be sensitive to health-enhancing effects of specific activities; while other variations may exacerbate the detrimental effects of specific lifestyle choices. Development of methods that quantify gene-environment interactions may eventually yield data that helps guide health-related choices and medical interventions for complex diseases.

Current methods to detect such interactions rely on variance component methods. These methods have been necessary due to the m $\gg$ n problem, where the number of predictors (m) vastly outnumbers the number of observations (n). These methods are all computationally intensive from computations of the sum of squares between each set of predictors, which is further exacerbated when considering gene-environment

interactions, as we increase the number of predictors from m to 2m+1 in the case of a single environmental exposure. Novel methods are thus needed to enable fast and unbiased calculations of the variance explained (R^2) for gene-environment interactions in very large samples on multiple traits. Taking advantage of the large number of participants in contemporary genetic studies, we herein propose a novel method for continuous trait R^2 estimates that are up to 20 times faster than current methods.

Our method is based on the observation that the multiple regression coefficient of determination or variance explained (R^2) can be used to accurately estimate heritability, whereby the trait of interest is the dependent variable and genotypes from the targeted genetic region are the independent variables. By assuming a linear model relationship between a quantitative trait and genetic variables, the heritability; defined as the variance explained by genetics, becomes equivalent to the R^2. Extending this observation to include an environmental exposure variable and computing the interactions between the genotypes and environmental exposure allows us to examine the variance explained by genetic interactions with an environmental exposure. By partitioning the genome into non-overlapping regions, it is thus possible to estimate genome-wide interactions with environmental exposure variables. However, partition of the genome into a large number of small regions presents challenges. First, linkage disequilibrium (LD) spills at the junction can theoretically inflate heritability estimates if many such junctions exist. Second, any residual population stratification effects would be amplified if heritability at each region is even slightly overestimated. This effect is expected to be proportional to the number of regions. Third, calculation of R^2 on very large datasets can be challenging. Through usage of the conjugate gradient method and computation acceleration via a graphical-processing unit, we can estimate R^2 using multiple linear regression models at exponentially higher speeds. We have devised a novel method, monsterlm, that enables multiple linear regression on large regions encompassing tens of thousands of variants in hundreds of thousands of participants. Including up to 25 000 variants per region,

1 environmental factor and 25 000 interactions between SNP variants and environmental exposure, only 60 blocks are necessary for genome-wide analysis, limiting both the possiblity of LD spillage and inflation due to population stratification.

1.2 Background and definitions

Deoxyribonucleic acid(DNA) is a molecule, found typically in the nucleus of a cell, containing the genetic information of a living organism. This information codes for cellular growth, function and reproduction. DNA is made up of two strands of molecules called *nucleotides* coiled together to form a double helix structure. Each nucleotide is made up of one of the four nitrogen bases: adenine (A), thymine (T), cytosine (C) and guanine (G), combined with a sugar and phosphate molecule. Due to the Watson-Crick pairing rules, which states that A pairs with T and C pairs with G, each strand of DNA contains the exact same genetic information, with different coding. A sequence of DNA that produces a protein or other functional element is referred to as a *gene*. The *genome* is the complete set of genetic information in an organism. Genomes are stored in *chromosomes*, which are made up of long strands of DNA tightly coiled together around histone proteins. Typically, there are 23 pairs of chromosomes that make up the human genome. Through genetic recombination, where the genetic material between chromosomes of the parent are exchanged, each copy of the chromosome in the child is not identical to the parents chromosome they inherit. *Single-nucleotide polymorphisms* (SNPs) are the most common form of genetic variation among humans. SNPs represents a change that occurs at a specific locus of the genome at a single nucleotide. For example, at a specific locus in the genome where most of the population have a C nucleotide, there is a subset of the population that have a T nucleotide, meaning there is a SNP at that locus of the genome. While most SNPs have no effect on the human body, certain variations can have an effect on health. SNPs are studied to locate genes associated with diseases with a hereditary component, such as heart disease. Most SNPs are *biallelic*,

meaning that at a base position of the genome, there can be two possible nucleotide variations, however, there are cases where there can be up to 4 possible bases at that location. The variations of the nucleotide are called the alleles at that locus. The common variation at a base position is called the major allele, and the less common allele is called the minor allele.

To analyze the effects of SNPs on various quantitative traits, studies are conducted using genotype data. A SNP genotype dataset of n individuals and m studied SNPs is represented as a matrix $X_{n\times m}$. Typically genotype data comes from a gene array where the columns represent the SNPs coded onto the array in its manufacture. SNP genotype data can take on values of 0, 1, 2 to indicate the number of minor alleles in a single base pair. The SNP genotype data is used to test for association with an observable biological trait called the *phenotype*. The phenotypes of interest in this work are biomarkers related to cardiovascular disease: ApoB, Bilirubin, Cholesterol, C-reactive protein (CRP), High-Density Lipoprotein (HDL), Glycated Hemoglobin (HbA1c),Low-Density Lipoprotein (LDL), and Triglycerides (TG).

SNPs are not entirely independent of one another. The genotype of one SNP may be correlated with the genotype of other SNPs which are physically close to it in the genome. When the genotype at multiple SNPs is not independent, the SNPs are said to be in *linkage disequilibrium* (LD). SNPs in LD can affect the association between SNPs and a measured trait. If two SNPs are high in LD, but only one SNP is a significant contributor to a phenotype, the study can incorrectly overestimate the effects of the non-contributing SNP on the measured phenotype due to the correlation between the allele frequencies. It is an important aspect in studies of associations between SNPs and phenotypes that the LD between SNPs are accounted for.

1.3 Genome-Wide Environmental Interactions

The term $gene-environment\ interaction$ has different meanings to different biomedical researchers. However, in this thesis, we focus on the concept of $effect\ modification$ where the genetic and environmental exposures convey synergistic effects, or in other words, where the joint effects are more or less than additive and the estimated genetic effect on a trait differs in magnitude across the spectrum of an environmental exposure.

In the case of type-2 diabetes, it's often said that the disease is the consequence of gene-environment interactions (Wright 1932). Indeed, both the environment and the genome are involved in diabetes etiology, and there are many genetic and environmental risk factors for which robust evidence of association exist. But when epidemiologists and statisticians discuss interactions, they are referring to the synergistic relationship between the exposures, and there is limited empirical evidence for such effects in the study of cardio-metabolic disease. In non human obesity, a condition widely believed to result from a genetic predisposition triggered by exposure to adverse lifestyle factors, of the >200 human gene-lifestyle interaction studies reported since 1995, only a few examples of interactions have been adequately replicated (Ahmad et al. 2013).

In our studies, we examined gene-environmental interactions as the increased variance explained as a result of including interaction variables into our analysis model in $MonsterLM$. The environmental exposure of interest, was one that would relate best to obesity: the waist-hip ratio. An interaction variable between a SNP and the environmental exposure would be defined as the multiplicative product between the 2 variables.

1.4 Applications of Gene-Environment Interactions

Some of the earliest empirical examples of gene-environment interactions come from studies in *Drosophilia* that show that eye facet number varies both by genotype and

temperature; similar examples exist for other features of the fly (Krafka 1920). In agricultural genetics, the need to maintain or improve food security cultivated genetically engineered plants to maximize crop yields conditional on environmental characteristics. Studies in durum wheat, illustrate that in low crop yield regions, specific strains like the D3145 can perform well, whereas other strains produce much higher yields than D3414 in high yield regions (Annicchiarico and Mariani 1996). These studies emphasize the point that matching appropriate environments and medical interventions to genotype is likely to be necessary for the optimization of health phenotypes in humans.

Animal studies of obesity and diabetes also provide useful examples of interactions, where phenotypic differences between genetically engineered animals are augmented with interventions that perturb the molecular pathways upon which the gene of interest reside. For example, high-fat feeding is commonly used as an intervention to accentuate phenotypic differences between genetically distinct animals; in a study of glucose and lipid metabolism, the effects of 8-week high and low fat feeding regimes on metabolic phenotypes of five inbred mouse strains were compared; the study showed that metabolic sensitivity to dietary fat varied considerably by genotype. The non-obese dietary mouse strain has provided a longstanding mouse model for autoimmune type 1 diabetes due to its predisposition to early-onset disease; this mouse is especially susceptible when raised in a germ-free environment, but much less so when raised in 'dirty' cages. This phenomenon, which is not observed in wild-type mice is thought to reflect immune adaptations in the NOD mouse that require exposure to foreign microbes early in life (Singh and Rabinovitch 1993).

Modulation of disease by environmental exposure is not limited to mouse models either. Complex metabolic diseases such as non-autoimmune diabetes are often uncommon in indigenous populations living traditional substance farming or hunter-gatherer

lifestyles, yet genetically similar people living industrialized lifestyles are often disproportionately afflicted. These observations are consistent with the presence of susceptibility loci whose effects are triggered by environmental exposures. This phenomenon is most apparent in ethnic groups whose recent evolution is characterized by migration and frequent exposure to famine, cold and other metabolic stressors. This process might have led to enrichment of alleles that predispose to metabolic efficiency, specifically after meals. Other intriguing examples are those from certain populations that cope unusually well living at high altitudes, in nutrient deficient settings or in cold climates (Hancock 2011). Whilst these observations are especially prone to confounding, bias and reverse causality, they provide tentative support for gene-environment interactions in human disease.

1.5 Discovery Strategies

Through the advent of genome-wide association studies (GWAS) around 2005 facilitated a new era of genetic association studies and the rapid discovery of thousands of loci for many complex traits. GWAS lead to rapid advances in population genetics largely because it is agnostic to prior biological knowledge, which contrasts most previous gene discovery approaches. Following the identification of genetic loci associated at genome-wide significance, researchers were exploring if environmental exposures modified their effects. There is appeal to this approach because few statistical tests are performed, which helps preserve statistical power and is analytically simple. However, there are arguments for why loci derived from GWAS may not be good candidates for interactions. Heterogeneous SNP association signals are generally filtered out in standard GWAS meta-analyses, yet SNPs involvement in gene x environment interactions are likely to have different effects across populations. The need for $MonsterLM$ is accentuated from the fact that most comprehensive studies focused on determining whether established GWAS-derived loci interact with environmental risk factors or clinical interventions have

yielded predominantly null results (Franks 2011). With GWAS came the the possibility to conduct interaction tests at a much higher variant density, and in samples of unrelated individuals. The simplest approach involves testing all SNPs for interaction with one or more environmental exposures. Conventional genome-wide interactions required sample sizes that are often unachievable to be adequately powered. The advantage of $MonsterLM$ is to be able to estimate the overall variance that is due to interactions.

1.6 Biomarkers and Cardiovascular Disease (CVD)

Biomarkers are a measurable substance in an organism whose presence is indicative of some phenomenon such as disease or infection. Biomarkers are by definition objective, quantifiable characteristics of biological processes. They may correlate with a patient's experience and sense of wellbeing, and it is easy to imagine measurable biological characteristics that do not correspond to patients' clinical state, or whose variations are undetectable and without effect on health. It is also even easier to imagine measurable biological characteristics whose measurement or intra-individual variance among populations is so great as to render them all but useless as reliable predictors of disease or its absence. Nonetheless, biomarkers provide valuable information on understanding disease with respect to a patient.

The biomarkers examined throughout this project have substantial evidence relating them to cardiovascular disease (CVD). ApoB or Apolipoprotein B is a component of lipoproteins, which are particles that carry lipids in the blood. Specifically, this protein is a building block of very low-density lipoproteins (VLDLs), intermediate-density lipoproteins (IDLs), and low-density lipoproteins (LDLs). These related molecules all transport lipids, including cholesterol in the bloodstream.

LDLs are the primary carriers of cholesterol in the blood. When there is too much cholesterol in your blood, it builds up in the walls of your arteries, causing a process

called atherosclerosis, a form of heart disease. The arteries become narrowed and blood flow to the heart muscle is slowed down or blocked. The blood carries oxygen to the heart, and if not enough blood and oxygen reach your heart, you may suffer chest pain. If the blood supply to a portion of the heart is completely cut off by a blockage, the result is a heart attack.

There are two forms of cholesterol that many people are familiar with: Low-density lipoprotein (LDL or "bad" cholesterol) and high-density lipoprotein (HDL or "good" cholesterol.) These are the form in which cholesterol travels in the blood. Triglycerides are another form of lipids which travel in our bloodstream, which research is showing may also be linked to heart disease Feingold and Grunfeld 2018.

HbA1c is glycated hemoglobin, or commonly known as an individual's blood sugar levels. Research shows that higher HbA1c levels and glucose levels are generally linked with a higher risk of heart disease. A study published in 2017 found that the most ideal HbA1c level for people without diabetes is in the 5-6 percent range (Sujit 2017). Recent research has also suggested that patients with elevated levels of C-reactive protein, an inflammation related biomarker, also were at higher risk of diabetes, hypertension and CVD ((Pradhan 2001)).

1.7 Population Stratification

GWAS are an effective approach for identifying genetic variants associated to disease risk. However, a concerning issue is that they can be confounded by population stratification. Population stratification is defined as systematic ancestry differences underlying the genetic variance as noise, this observation comes from the first 2 principal components showing the stratification of individuals with Eastern and Western. ancestry. This method perform well in data sets in which population structure is the only kind of structure present, but are inadequate in data sets that also contain family structure or

cryptic relatedness. Fortunately in the data-set used in $MonsterLM$, the individuals are unrelated, so this issue is minimized and can be handled through the usage of genetic principal components.

GWAS have identified hundreds of common variants associated to disease risk or related traits. These studies have overcome the dangers of population stratification, which can produce spurious genetic associations if not properly corrected. However, accounting for population structure is more challenging when family structure or cryptic relatedness is also present, motivating the development of new methods. Because the associations that have been reported primarily occur at markers with allele frequency differences between subpopulations, it is critical to correct for stratification (Price et al. 2006).

The current method has been to use genetic principal components to measure the extent of inflation due to population stratification and to correct for stratification using methods that infer genetic ancestry by modeling our phenotypes and environmental exposure of interest with the genetic principal components, and then taking the residuals of that model to use in place of the old phenotypes. This is adjusting our phenotypes and exposures for population stratification effects through residualization.

The concept of residualization will be described in more detail in the statistical background through Chapter 2.

Chapter 2

Statistical Background

2.1 The Linear Model

The linear model has become the standard framework for many genetic analyses. They provide control for confounding factors, allow for aggregating genetic effects from multiple variants and enable the join analysis of multiple traits. While inference in linear models can be computationally demanding as the number of variants increases to large numbers, efficient implementations and manipulation of the genotype matrix enables use of linear models to large datasets.

A linear model describes a continuous output variable as a function of one or more input variables, which in the case of genetics is our SNP variables. Denoting the number of samples as N, the output variable for sample i as y_i and $[x_{i1}, ..., x_{iM}]$ M input variables or observations for sample i, the linear model can then be cast as:

$$y_i = \sum_{m=1}^{M} x_{im}\beta_m + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, \sigma_e^2) \tag{2.1}$$

The residual term ϵ_i accounts for the stochastic relationship between x and y, there is measurement noise or other unmodelled factors. ϵ_i is assumed to follow a normal

distribution with mean 0 and variance σ_e^2, and to be independent across samples. In equation (2.1), β_m denotes the effect size of the input variable m.

Introducing the output vector $\mathbf{y}$, the input matrix $\mathbf{X}$, the effects vector β and the residual vector $\mathbf{e}$ as

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1M} \\ x_{21} & x_{22} & \dots & x_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{NM} \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_M \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix} \tag{2.2}$$

the linear model in (2.1) can be expressed in matrix form as

$$\mathbf{y} = \mathbf{X}\beta + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma_e^2 \mathbf{I}_N), \tag{2.3}$$

where $\mathbf{I}_N$ denotes the $N \times N$ identity matrix.

2.2 Incorporating an interaction variable

An interaction may arise when considering the relationship in which the effect of one causal variable depends on the state of a second causal variable. An interaction variable or feature is a variable constructed from an original set of variables to try to represent either all of the interaction present or some part of it. It is common to use the products of original variables as the basis of testing whether interaction is present (Southwood 1978). In this case, we consider an additive model that includes an interaction between the genotype matrix X and an environmental vector E.

In this full model with the interaction; we consider Y as a continuous phenotype trait, X as the genotype matrix and E as the environmental exposure vector. The

model becomes:

$$Y_{phenotype} = \beta_X X_{genotype} + \beta_E E_{environment} + \beta_{GxE}(X \times E)_{interaction} \tag{2.4}$$

Where $G \times E$ is the product between each individual SNP variant and an environmental exposure, resulting in another $N \times M$ matrix the same size as X. For the purposes of analysis, we will include all variants and interactions as a single matrix U, which is a $N \times 2M$ matrix without the exposure, as described in the methods section, will be residualized from the traits of interest.

2.3 Ordinary Least Squares

To estimate the effects vector β, the ordinary least squares method is commonly used. OLS chooses the parameters of a linear function of a set of input variables by the principle of least squares: minimizing the sum of squares of the difference between the observed dependent variable and those predicted by the linear function, it was first derived by Gauss in 1795 (Bretscher 1995) .

Given the linear system:

$$U\beta = \mathbf{y}, \tag{2.5}$$

the goal is to find coefficients β which fit the equation best in minimizing the sum of squares function $S(\beta)$, where S is defined as:

$$S(\beta) = ||\mathbf{y} - U\beta||^2 \tag{2.6}$$

This minimization problem has a unique solution, provided that the $2M$ variables are linearly independent (which is the case through the use of LD pruning), giving the

normal equations:

$$(U^T U)\hat{\beta} = U^T y \tag{2.7}$$

Normally, this results in the solution for $\hat{\beta}$ as:

$$\hat{\beta} = (U^T U)^{-1} U^T y \tag{2.8}$$

However, as the number of variables increases, computing the inverse of the $U^T U$ matrix takes an extraordinary longer time; therefore, an alternative method using the Conjugate Gradient Method was employed to speed up calculations, as it will be described in Chapter 3.

2.4 The Coefficient of Determination R^2

The coefficient of determination has many definitions. Generally, R^2 is defined as the proportion of variability (measured by the sum of squares S function) in a data set accounted for by a multiple regression model. This interpretation is usually presented at the conclusion of a multiple regression analysis.

Calculation of the coefficient depends on the ratio between 2 aspects, the total sum of squares of the model, and the sum of squares of the regression model. The total sum of squares (proportional to the total variance of the data) is defined as:

$$SS_{tot} = \sum_i (y_i - \bar{y})^2, \tag{2.9}$$

But since, $\bar{y}$ is mean-centered, it becomes 0, whereas the regression sum of squares is defined as:

$$SS_{reg} = \sum_i (\hat{y}_i - \bar{y})^2 \tag{2.10}$$

Hence, the R^2 is calculated as:

$$R^2 = \frac{SS_{reg}}{SS_{tot}} \tag{2.11}$$

Based on the definition, R^2 is a goodness-of-fit statistic for the overall performance of a multiple linear regression model. It can represent how well the regression line approximates the observed data points, as a measure of variability of the residuals. It lies between 0 and 1, where the close it is to 1, the better the linear relationship between the response variable and predictors. The closer to 0, the worse the linear relationship; indicating no linear relationship between the variables.

There is controversy regarding R^2 as a goodness of fit statistic, as it always increases even when a non-predictive regressor is added in a linear model. This is dealt with by adjusting the R-squared variable, including a penalty for the number of predictors in a model, the adjusted R^2 increases only if the added predictor improves the model more than would be expected by chance. The adjusted R^2 is calculated as:

$$R^2_{adj} = 1 - (1 - R^2)\frac{N - 1}{N - (2M) - 1} \tag{2.12}$$

where N and 2M are the dimensions from the corresponding covariate matrix U with size $N \times 2M$. R^2 is an important concept in genetics, as it described the variance explained by an additive SNP-based linear model. This variance is more commonly known as SNP-based heritability, which is how predictable a trait of interest can be, given all the genetic information known about the trait.

Besides evaluating the overall performance of a multiple regression model, R^2, can also be used as a general measure of determining the relative importance of predictors in MLR (Azen et al. 2001). After having built a model with a chosen set of predictors, one may want to know a relatively important set of predictors, or rank the predictors according to their contributions in predicting the outcome. This is especially important in the case

of interaction variables, and determining if an added interaction matrix improves the fit of a linear model. Budescu (Budescu 1993) suggested that an appropriately general measure of importance should satisfy three conditions:

- Importance should be defined in terms of a variable's 'reduction of error' in predicting the outcome
- The method should allow for direct comparison of relative importance instead of relying on inferred measures
- Importance should reflect a variable's direct effect, total effect and partial effect.

Based on these criteria, importance became defined as the squared semipartial correlation, or the difference between 2 models' R^2 from nested models.

2.5 Properties of R^2

The variance of R^2 was originally derived by Wishart (Wishart 1931) as:

$$var(R^2) = \frac{4\rho^2(1-\rho^2)^2(n-m-1)^2}{(n^2-1)(n+3)} \tag{2.13}$$

When n is large and makes $2m+5$ small relative to n, the expression further reduces to:

$$var(R^2) \approx \frac{4\rho^2(1-\rho^2)^2}{n} \tag{2.14}$$

Hence, the 100(1-α)% confidence limits for ρ^2 by using the Wald method is given by

$$100(1-\alpha)\%CI = R^2 \pm z_{\alpha/2}\sqrt{\hat{var}(R^2)}, \tag{2.15}$$

where $\hat{var}(R^2)$ is the sample estimate of var(R^2) with ρ^2 replaced by R^2, $z_{\alpha/2}$ is the $100(\alpha/2)$ upper percentile point of a standard normal distribution.

It is important to recognize that when the coefficient of determination is 0, the variance estimate is also 0, thus the Wald method cannot be used for that specific case should it occur during analysis. Furthermore, this method may give confidence limits out of range of 0 to 1, due the skewness of the sampling distribution of R^2.

2.6 Confidence Interval of a Semipartial R^2

Comparisons of R^2 may arise in the case of two R^2s between 2 models, where 1 is nested in the other. Specifically, determining whether an additional set of predictors provides a significant improvement in prediction response. It means whether an additional parameter set B provides improvement over parameter set A alone in predicting y. Normally the equivalence between the R^2s between the 2 parameter sets AB and A through an F statistic comparing a full model with a reduced model in regression analysis. However, for $MonsterLM$ it will be important to have CI estimates for each fit of our partitioned genotype matrix with and without interactions.

The best way to describe an estimator of an unknown parameter of interest is to obtain the density function of that estimator's sampling distribution. However, the exact marginal distribution for the semipartial R^2 is extremely complex (Fisher 1928). Many researchers provide asymptotic solutions for the joint distribution of multiple correlations. Olkin and Siotani (Olkin and Siotani 1976) provided the asymptotic distributions of multiple correlations, which based on central limit theorem is approximately multivariate normal. However, the sampling distribution of a single R^2 has been known to be so skewed that the asymptotical normality of two or more R^2s is only suitable for large samples for n $>$ 200 (Cohen and Cohen 1983). Fortunately, the partitioning of the genotype matrix allows such conditions to be fulfilled.

A confidence interval is usually the most important index when making statistical

inferences on this parameter. Olkin and Finn Olkin and Finn 1995 suggested Wald-type confidence intervals for differences between two R^2, defined as ΔR^2_{AB-A}, which is constructed through directly estimating the variance of the difference between these R^2 values.

The Wald-type confidence interval is defined as:

$$100(1-\alpha)CI(\Delta R^2_{AB-A}) = (R^2_{AB} - R^2_A) \pm z_{\alpha/2}\hat{\sigma}_1/\sqrt{n} \tag{2.16}$$

where σ_1 is the asymptotic standard error of the difference ΔR^2, and $\hat{\sigma}_1$ is the sample estimate with population correlations replaced by their corresponding sample correlations, given by:

$$\hat{\sigma}_1 = \sqrt{\hat{var}(\Delta R^2_{AB-A})} \tag{2.17}$$

The standard error estimator can be obtained using the delta method as suggested by Olkin and Finn.

The way the delta method works here is by first declaring our function of interest:

$$f(R_A, R_{AB}) = R^2_{AB} - R^2_A \tag{2.18}$$

Where AB represents the full parameter set with interactions, and A represents the parameter set just including the genotype matrix.

The variance of the difference is obtained by the delta method from the expression:

$$Var_\infty(R^2_{AB} - R^2_A) = a\Phi a' \tag{2.19}$$

where a is a row vector comprising of 2 partial derivatives, in which the correlations are evaluated at their parameter values. These two partial derivatives are:

$$\frac{\delta}{\delta r_A}(R^2_{AB} - R^2_A) = 2r_A \tag{2.20}$$

and

$$\frac{\delta}{\delta r_{AB}}(R^2_{AB} - R^2_A) = 2r_{AB} \tag{2.21}$$

Thus, we get:

$$a_1 = 2\rho_{AB}a_2 = -2\rho_A \tag{2.22}$$

Next, the matrix Φ is the 2x2 variance-covariance matrix for r_{AB} and r_A. From Olkin and Finn (Olkin and Finn 1995),

$$\Phi_{11} = Var_\infty(r_{AB}) = (1-\rho^2_{AB})^2/n \Phi_{22} = Var_\infty(r_A) = (1-\rho^2_A)^2/n \tag{2.23}$$

and lastly the other components of Φ as:

$$\begin{aligned}\Phi_{12} = \Phi_{21} &= Cov(R_A, R_{AB}) \\ &= (\frac{1}{2}(2\rho_{AB,A} - \rho_{AB}\rho_A)(1 - \rho^2_{AB} - \rho^2_A - \rho^2_{AB,A}) + \rho^3_{AB,A})/n\end{aligned} \tag{2.24}$$

Now we will let b, c, d equal the corresponding terms of Φ to rewrite the following expression:

$$a\Phi a' = \begin{bmatrix} 2\rho_{AB} & 2\rho_A \end{bmatrix} \begin{bmatrix} b & c \\ c & d \end{bmatrix} \begin{bmatrix} 2\rho_{AB} \\ 2\rho_A \end{bmatrix} \tag{2.25}$$

Calculating the triple product yields:

$$a\Phi a' = 4b\rho^2_{AB} + 4d\rho^2_A - 8c\rho_{AB}\rho_A \tag{2.26}$$

Re-substituting b,c,d into their respective equations then finally derives the asymptotic variance of the semipartial $R^2_{AB,A}$ as described by Alf Jr and Graf 1999:

$$Var_\infty(\Delta R^2_{AB-A}) = Var_\infty(R^2_{AB}) + Var_\infty(R^2_A) - 2Cov_\infty(R^2_{AB}, R^2_A) \tag{2.27}$$

Combining all parts together from the previous equations, we obtain an expression that can use to calculate this asymptotic variance:

$$\begin{aligned} \hat{var}(\Delta R^2_{AB-A}) = \\ & 4R^2_{AB}(1 - R^2_{AB})^2/n + \\ & 4R^2_A(1 - R^2_A)^2/n - \\ & 8R_{AB}R_A(0.5(2R_{AB,A} - R_{AB}R_A)(1 - R^2_{AB} - R^2_A - R^2_{AB,A}) + R^3_{AB,A})/n \end{aligned}$$

where the term $R_{AB,A}$ is calculated as:

$$R_{AB,A} = \sqrt{\frac{R^2_A}{R^2_{AB}}} \tag{2.28}$$

Through these equations, we can compute the $100(1-\alpha/2)$ CI for the semipartial coefficient ΔR^2_{AB-A}, where AB represents the full model with interactions and A represents the model without interactions, i.e. the genotype matrix only.

2.7 Quantile Normalization

It's extremely important to recognize that many assumptions of our linear model rely on data having mean 0 and variance equal to 1, as the model does not use an intercept term. Quantile normalization is a technique for making two distributions identical in statistical properties. To quantile-normalize a test distribution to a reference distribution of the

same length, sort the test distribution and sort the reference distribution. The highest entry in the test distribution then takes the value of the highest entry in the reference distribution, the next highest entry in the reference distribution, and so on, until the test distribution is a perturbation of the reference distribution Bolstad et al. 2003.

To quantile normalize two or more distributions to each other, without a reference distribution, sort as before, then set to the average (usually, arithmetic mean) of the distributions. So the highest value in all cases becomes the mean of the highest values, the second highest value becomes the mean of the second highest values, and so on. Fortunately, in our analyses, the trait and variables including interactions have a reference distribution to a Gaussian normal.

Quantile normalization is conducted in R using a standard normal reference distribution with the following command:

```
quantNorm = function(x){qnorm(rank(x,ties.method = "average")/(length(x)+1))}
```

This function would be applied to all data including the biomarker traits ($\mathbf{y}$), the genotype matrix ($\mathbf{X}$), the environmental exposure vector ($\mathbf{E}$) and the interaction matrix ($X \times E$).

2.8 Residualization

Adjusting for the effects of confounding variables is an important procedure in most statistical genetic analyses. The idea is to remove the effects a covariate that could be confounding to the relationship between the dependent and independent variables (Richard 2012).

Consider the linear model, with a phenotype y, covariate matrix U, and the true effects vector β:

$$y = U\beta + \epsilon \tag{2.29}$$

Given the phenotype trait, and variables to be adjusted for, we can estimate $\hat{\beta}$ as stated in the linear model section, by using those variables in the model, and computing the predicted $\hat{y}$ values as:

$$\hat{y} = U\hat{\beta} \tag{2.30}$$

The predicted values $\hat{y}$ contain the information that the variables to be adjusted for have on the phenotype y. Next consider the residuals e:

$$e = y - \hat{y} \tag{2.31}$$

The residuals can be considered the remaining information about the phenotype y after the information about the variables involved with $\hat{y}$ is removed through subtraction.

This method is useful in adjusting for the confounding variables in regressional analyses; but has an even more important effect in $MonsterLM$.

$MonsterLM$ involves partitioning the genome into multiple genetic blocks, and then fitting the SNP blocks with their associated interactions with an environmental exposure. However, by including the environmental exposure during each fit with a SNP block, the overall R^2 suffers from inflation.

This can be circumvented by residualizing the biomarker phenotypes with the environmental exposure variable, alongside adjusting for population stratification effects by including the genetic principal components. Through this residualization, there will be no need to include the environmental exposure during the fitting of each SNP block and associated interactions.

The model we'll be concerned with then would be:

$$e = X_{genotype}\beta_G + (X \times E)_{interaction}\beta_{GxE} + \epsilon \tag{2.32}$$

Where e is the residualized phenotype after adjusting for the environmental exposure and population stratification.

With this residualized phenotype, there will be no need to be concerned with the potential inflation and spillover of the variance explained by the environmental exposure on the trait of interest and in the interaction terms.

Chapter 3

MonsterLM: Methodology

3.1 Data and Pre-processing

The data used in analysis comes from the UK Biobank, specifically a sample of 325 991 unrelated British individuals with their genotyping data available (1 031 125 variants included), alongside phenotypic traits such as blood biomarkers: Triglycerides, ApoB, HbA1c, CRP, total cholesterol, total bilirubin, total glucose, HDL, LDL and environmental exposures such as Body Mass Index (BMI) and Waist-Hip Ratio (WHR) (Bycroft et al. 2018). The UK Biobank is a large population-based study which includes over 500,000 participants living in the United Kingdom. Men and Women aged 40–69 years were recruited between 2006 and 2010 and extensive phenotypic and genotypic data about the participants was collected, including ethnicity and history of GDM. Details of this study are available online (https://www.ukbiobank.ac.uk) (Bycroft et al. 2018). It was important in analysis to adjust for individuals receiving medication that would affect glucose and cholesterol levels.

The genotype data was also processed by removing highly correlated SNPs that had a linkage disequillibrium value of more than 0.9 (a process known as LD pruning with a r^2 threshold of 0.9), and removing rare SNPs with a minor allele frequency of less than

0.05 as the process was examining common variants. This process cut the SNP variants analyzed to 1 031 135 variants. After pruning, the raw genotype data was normalized to have mean = 0 and variance = 1.

To make sure data matched between each biomarker, the genotype data and the environmental exposure; mean imputation was employed for missing data in the biomarkers. The underlying assumption for mean imputation requires the missing values to be the mean values in reality, this is rarely the case, however due to the relatively small number of missing values, the effect was minimal. Mean imputation is a method where given a dataset with missing values, the mean of the variable is computed without the missing values, then each missing value is replaced with the computed mean. Simple simulation tests were done to confirm mean imputation did not have a noticeable effect on the adjusted R^2 computations.

The rationale for using WHR as the environmental exposure is that WHR is a measurement of obesity; which is the accumulation of both genetic factors and lifestyle choices (lack of exercise, poor eating habits, etc.). Specifically, it's a measure of central obesity, which is widely accepted as the type of adiposity mainly driving adverse metabolic consequences, including risk of diabetes and CVD. Furthermore, a recent study concluded that life expectancy in the United States will decrease in the coming years due to the dramatic increase in obesity (Stewart et al. 2009). The public health implications of WHR are a strong driver in choosing WHR as the environmental exposure in this study. The WHO states that abdominal obesity is defined as a waist-hip ratio above 0.9 for males and above 0.85 for females. Furthermore, WHR has been shown to be a better predictor of CVD than waist circumference and body-mass index as it also takes into account of the differences in body structure (Mørkedal et al. 2011). Hence it's possible for two individuals to have vastly different waist-hip ratios despite having the same BMI. There is a genetic component to obesity; however, from our simulation

study, we examined the cases where the exposure was partially determined by genetics (either by the same SNPs that are causal to the trait of interest, or different SNPs). In both cases, the estimated interaction variance between SNPs and the exposure for the residualized simulated phenotype did not differ from the true interaction variance. The heritability of the waist-hip ratio can range from 20 percent to 60 percent in different studies (Heid et al. 2010); in the UKB it is reported to be 22 percent (Bulik-Sullivan et al. 2015, which is relevant as our simulation tested environmental exposures that would have a heritability of around 30 percent.

3.2 MonsterLM: the Model

To enable MLR on extremely large datasets, we developed the monsterlm method. Monsterlm parallelizes the calculation of least squares regression including the interaction terms between the genotypes and environmental factor. The calculation is done such that the only practical limitation is the inversion of the m x m matrix, where m is the combined number of SNPs, environmental factors and interaction terms, but without a limit on the number of participants (n) included in the analysis. This limitation is circumvented through the use of the conjugate gradient method and GPU acceleration. This method allows for the calculation of R^2 interaction estimates for multiple traits, leading to appreciable gains in speed.

Briefly, the n x m genotype matrix G represents genotypes at m variants on n individuals, the n x 1 vector E represents the environmental exposure on n individuals. Lastly, the m x n interaction matrix GxE represents m interaction effects for 1 exposure on n individuals. All genotypes and environmental exposures have been quantile normalized to have mean 0 and variance 1, and adjusted for population stratification effects through residualizing the biomarkers of the first 20 genetic principal components. Given a quantitative trait y, the combined covariate matrix U, we describe the least squares

estimate for $\hat{\beta}$ as a manipulation of the quadratic form of the normal equation (described in Section 3.4). After computing $\hat{\beta}$, the predicted values ($\hat{y}$), can be computed as:

$$\hat{y} = U\hat{\beta} \tag{3.1}$$

, where U is the combined matrix of variants for the genotype block of interest and interaction terms. We can compute the same model except the interaction terms are no longer present.

Then, from each model we can calculatethe varianced explained by interactions as the difference between the variance explained for the full model and the model without interactions for each monsterblock:

$$R^2_{interaction} = R^2_{fullmodel} - R^2_{SNPsonly} \tag{3.2}$$

Under the assumption that the variants selected for analysis are not highly correlated (based on pruning for LD), we can then estimate the total interaction variance explained across the pruned genotype-wide environmental interaction matrix R^2_{GWEI} as:

$$R^2_{GWEI} = \sum_{i}^{j} R^2_{interaction} \tag{3.3}$$

where j represents the total number of monster blocks generated, which for this analysis is 60.

To compute the $100(1-\alpha)$ confidence intervals for R^2_{GWEI}, we take the calculated variance of each block's difference as $R^2_{fullmodel} - R^2_{genotype}$, and based on the assumption of low to moderate correlation between blocks (not very we can calculate the total variance as:

$$Var_{\infty}(\Delta R^2_{GWEI}) = \sum_{i}^{j}(Var_{\infty}(\Delta R^2_i)) \tag{3.4}$$

Then with the total variance calculated, we can use the Wald method to compute the interval as:

$$100(1-\alpha)CI = \Delta R^2_{GWEI} \pm z_{\alpha/2}\sqrt{Var_\infty(\Delta R^2_{GWEI})} \tag{3.5}$$

3.3 Simulation Study

We tested MonsterLM with simulations using LD-pruned and MAF filtered genotypes from the UK Biobank (Bycroft et al. 2018. Specifically, we used chromosome 22 alongside the full UKB sample (325K individuals) to generate a single block of 20 000 SNPs. We then simulated the effects (β) from a standard normal distribution, simulated the error assuming it belonged to an i.i.d. normal distribution and randomly sampled 20% of the total SNPs to have a causal genetic effect in simulation our trait of interest: Y. To sample the GxE effects to be causal, we randomly subset a sample of the SNPs that had causal effects, and set 50% of these SNPs to have GxE effects causal to the trait (10% of the total SNPs). The simulated trait Y is computed through the following equation:

During simulations we considered multiple scenarios: 1) The environmental factor is not dependent on the genotypes of individuals (noted as EFND). 2) The environmental factor is dependent on the genotype of individuals. Further exploring this scenario, we considered 2 sub-scenarios. First, we simulated the EF assuming that the SNPs causal to the phenotype, would be the same SNPs that the EF depended on. The second sub-scenario simulated the EF assuming that the SNPs causal to the phenotype, were not the same SNPs that the EF depended on. The heritability was set to 0.025, exposure R2 set to 0.2 and interaction R2 set to 0.005 to simulate real trait scenarios. . The simulated trait Y is computed through the following equation:

$$Y_{phenotype} = \beta_G X_{genotype} + \beta_E E_{environment} + \beta_{GxE}(X \times E)_{interaction} + \epsilon \tag{3.6}$$

In the cases where we simulated an environmental exposure that was dependent on the SNPs, we'd use the following equation:

$$Exposure = \beta_{SNPs} X_{genotype} + \epsilon \tag{3.7}$$

where the genetic effect of SNPs in determining the exposure would explain 20% of the variance in the exposure trait, an extreme estimate for a single chromosome simulation; however this setting proves the robustness of our model to high correlation between SNPs and the environmental exposure.

The simulations were based on a single chromosome to simplify results. The heritability was set to 0.025, the environmental exposure R^2 set to 0.01, GxE interaction R^2 set to 0.005 and the error variance was set to 0.96 to accurately simulate real trait scenarios. Adjusting the variance explained by the genetic effect was done by adjusting the product of the β and U matrix or vector after mean-variance standardization or vector:

$$Effect_{adj} = \sqrt{R^2} \frac{(U\beta - \hat{\mu}_{U\beta}}{\hat{\sigma}_{U\beta}} \tag{3.8}$$

Where R^2 is the variance explained set for the simulation experiment for each effect. This process was repeated for all effects including the environmental exposure effect and GxE interaction effects.

Our simulation results (Fig. 3.1) confirm that monsterlm can estimate interaction effects with notable precision and accuracy regardless of the dependency of the environmental factor (with slight bias). It may be a surprising result that despite the environmental exposure being dependent on SNPs, the distribution of interaction variance narrowed slightly; however a study had derived that correlation between SNPs and environmental exposure would have minimal effects on the interaction variance Sulc et al. 2020.

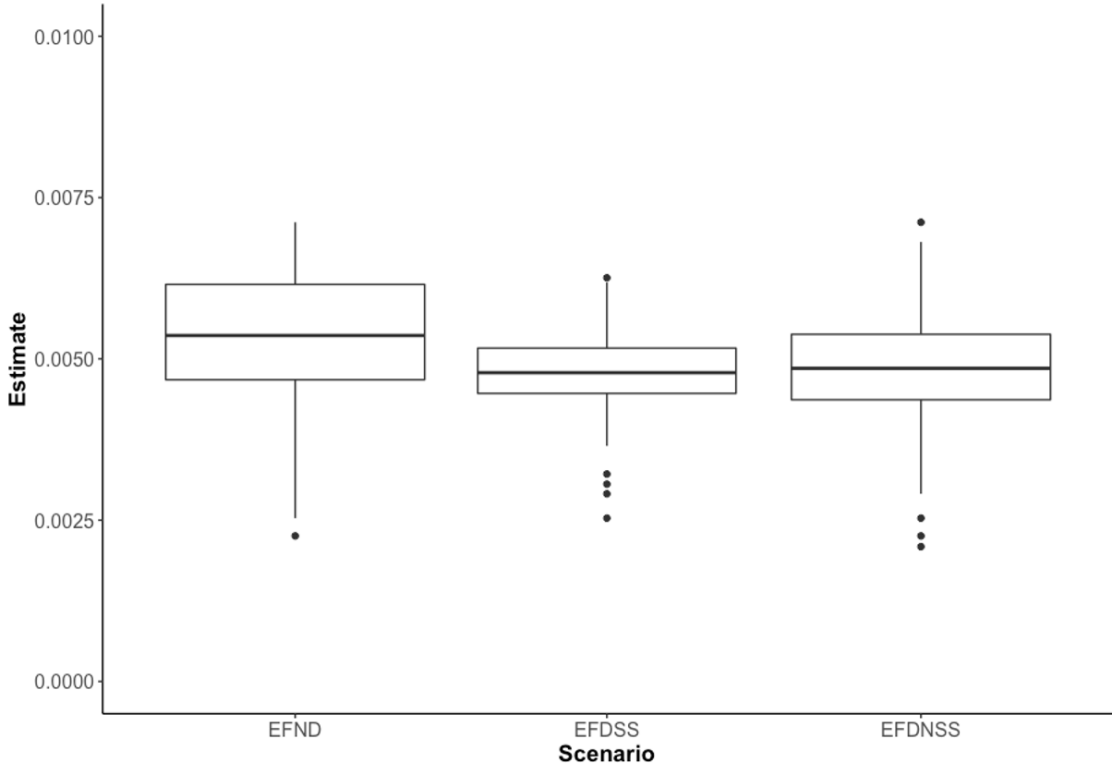

FIGURE 3.1: Distribution of Estimated R^2 for the 3 cases described in Simulations

3.4 GPU Acceleration through the Conjugate Gradient Method

Solving the normal equation for $\hat{\beta}$ using equation (2.7), which we will rewrite as:

$$A\hat{\beta} = b \tag{3.9}$$

where $A = X^T X$ and $b = X^T y$, computing the least squares solution for $\hat{\beta}$ can still be quite a computationally intensive task for matrices of size 50 000 x 50 000. Knowing that matrix A is a positive-definite matrix, we can actually solve $\hat{\beta}$ directly from using the Conjugate Gradient method. The Conjugate Gradient method is one of the most popular methods for solving systems with definite matrices (Bohacek et al. 2019). It produces the

exact solution of $\hat{\beta}$ after a finite number of iterations (Shewchuk 1994). Through the use of graphical processing units, the conjugate gradient method can compute estimates for the multiple linear regression model up to 170 times faster for matrices of up to 50 000 features than the usual method solving the normal equation through matrix inversion (Fig. 3.2). The experiment described simulates the $n \times p$ covariate matrix, the $p \times 1$ coefficient vector β, and a noise term ϵ. From here, $\hat{\beta}$ was estimated using both general matrix inversion in base R (B/L) and using the proposed solution (PPSD), and the error was computed as the mean squared error between the estimated $\hat{\beta}$ and the true β.

# Feature	time_B/L	time_PPSD	error_B/L	error_PPSD
1000	0.033	0.014	0.0001	0.0001
2000	0.110	0.013	5.6191e-05	5.6191e-05
4000	0.526	0.023	2.8837e-05	2.8837e-05
8000	3.530	0.065	1.3793e-05	1.3793e-05
12000	10.364	0.129	9.1160e-06	9.1160e-06
16000	22.208	0.228	6.9269e-06	6.9269e-06
24000	69.410	0.496	4.6462e-06	4.6462e-06
32000	169.099	0.904	3.4748e-06	3.4748e-06

FIGURE 3.2: Comparison of GPU-based acceleration vs. baseline inversion of matrices in R

The theory and algorithm of the conjugate gradient method will be described in the supplementary section; but a brief outline will be given here.

Given the normal equation, we can rewrite the equation as a quadratic function to be minimized:

$$F(\hat{\beta}) = \frac{1}{2}\hat{\beta}^T A\hat{\beta} - \hat{\beta}^T b \tag{3.10}$$

The purpose of the CG method is to minimize quadratic functions similar to the normal equation, an initial solution vector $\hat{\beta}_0$ is initialized (usually the inverse of the main diagonal of the A matrix), and then that vector is iterated going in the direction of the gradient of $F(\hat{\beta})$ until the minimum. Let $\hat{\beta}_k$ be the solutions vector after the k-th iteration. The residual vector after the k-th iteration is:

$$r_k = y - X\hat{\beta}_k, \tag{3.11}$$

and the negative gradient is:

$$g_k = -\nabla F(\hat{\beta}_k) = X^T r_k \tag{3.12}$$

Resulting in the following algorithm (Algorithm 1), The speed of this method comes

Algorithm 1 Conjugate Gradient Algorithm - Initiate $\hat{\beta}_0$

$r_0 = y - X\hat{\beta}_0$
$g_0 = X^T r_0$
for $i = 1$ to m(
if $g_{i-1} = 0$ then return x_{i-1}
if (i > 1) then $\hat{\beta}_i = ||g_{i-1}||^2/||g_{i-2}||^2$
if i $= 1$, then $p_i = g_0$, else $p_i = g_{i-1} - \hat{\beta}_i p_{i-1}$
$\alpha_i = ||g_{i-1}||^2/||Xp_{i-1}||^2$
$\hat{\beta}_i = \hat{\beta}_{i-1} + \alpha_i p_i$
$r_i = r_{i-1} - \alpha_i X p_i$
$g_i = X^T r_i$)
return x_m

from skipping the computation of the inverse of a large matrix, but instead just finding a vector over many iterations that solves the normal equation, and only requires one multiplication of X with an m-size vector and one multiplication of X^T with an n-vector. The multiple iterations can be computed in parallel via GPU systems.

3.5 Analysis Workflow

With $MonsterLM$ tested and verified through simulations, the analysis workflow would consist of 3 stages:

1. **Primary Analysis**: Calculate total variance explained by interactions across all 325 991 individuals and all 60 blocks for the 9 biomarkers and environmental exposure: WHR.

2. **Secondary Analysis**: Split the UKB data into two sets, 70 percent for model building (known as the discovery set), and 30 percent for model testing (known as the validation set). This would be followed with 2 separate analyses

   (a) Analysis 1: Conduct univariate regression on the discovery set with just SNPs and select a set of SNPs based on a set of p-value threshold (0.01, 0.001, 0.0001, 0.00001). Then test those SNPs and their associated interactions in the validation set and calculate heritability and interaction variance explained (and compare to the Primary Analysis Results).

   (b) Analysis 2: Conduct univariate regression similar to Analysis 1, however based on interaction p-values instead of SNPs, then conduct $MonsterLM$ with a proportion of interactions passing the same p-value thresholds on the validation set and calculate interaction variance explained.

3. **Polygenic Risk Scores** Guided by the results of Analysis 1 and 2, we would compute polygenic risk scores and examine if the incorporation of interactions can improve the predictiveness of PRS scores on the biomarkers.

The purpose of primary analysis was to get a measure of the variance explained by common variants across the genome interacting with an exposure that reflects lifestyle choices (i.e. obesity through waist-hip ratio).

3.6 Univariate Regression

The purpose of secondary analysis was to answer three questions:

1. Do G x E interactions arise from SNPs associated strongly to the trait of interest or all SNPs (including SNPs not significant to the trait).

2. Do a few strong G x E interactions explain the large variance explained by interactions, or is it a large number of small interactions.

Initially, significant SNPs were to be evaluated using a p-value threshold based on multivariable regression p-values. However, due to the large number of variables, the standard error on each β_i coefficient would become too large to be deemed signficant. Thus univariate regression was employed to evaluate the individual significance of a SNP or interaction for secondary analysis.

Upon completion of univariate SNP-based regression on the discovery set, four sets of SNPs were generated with each sets corresponding to a set of SNPs passing one of the following p-value thresholds in the validation set: $p < 0.01$, $p < 0.001$, $p < 0.0001$, $p < 0.00001$.

Each block would then be refitted to a linear model using $MonsterLM$, and then the heritability and interaction variance would be calculated for each of biomarkers. Then the proportion of variance explained by the set of SNPs would be compared to the total variance explained defined as:

$$h^2_{fraction} = \frac{h^2_{SNPset}}{h^2_{fullgenotype}} \tag{3.13}$$

for heritability, and for interaction variance:

$$\Delta R^2_{fraction} = \frac{\Delta R^2_{intset}}{\Delta R^2_{fullint}} \tag{3.14}$$

The process for interaction-based univariate analysis would occur as follows:

1. Regression would be conducted on each single SNP and its associated interaction effect from the discovery set for each block of SNPs.

2. SNPs would be selected if the interaction p-value passed the significance threshold, as stated in univariate SNP-based regression.

3. The significant interactions would then be included for each block of SNPs in the validation set, where $MonsterLM$ would fit each SNP block including the interactions passing the significance threshold used.

4. Once the fitting was complete, the fraction for interaction would be calculated as in Equation (3.12).

3.7 Polygenic Risk Scores

The final analysis of the project was to examine if inclusion of G x E interactions improve polygenic risk score's predictiveness.

A PRS is a quantitative measure that serves to predict the risk of a certain trait in individuals based on their genetics. Polygenic scores look at the relationship between an individual's genetic variation, represented by the SNP genotype data, and a certain phenotype. Once constructed, a polygenic risk scores can be used to predict the trait of an individual based on his or her genetics.

For a given phenotype, the polygenic risk score is defined to be a linear combination of a subset of the given coded genotypes of the SNPs and their associated weights that reflect the association between each individual SNP and the trait of interest. Let $[x_{i,j}]_{j=1}^{M}$ represent the genotypes of the M SNPs used in the PRS for individual i, the PRS is

defined to be:

$$S_i = \sum_{j=1}^{M} x_{i,j} W_j \tag{3.15}$$

One of the most common ways to construct the weights of a PRS is to use the regression coefficients from an external GWAS on the phenotype as the weights in the risk score. Under the GWAS approach, the PRS is calculated to be:

$$PRS_i = \sum_{j=1}^{M} x_{i,j} \hat{\beta}_j \tag{3.16}$$

A major issue with this GWAS approach is that since GWAS regression is univariate, it will not the resulting model will not be one that fits the optimal linear combination of SNPs to the desired polygenic trait. Another issue is that the analysis does not account for SNPs in strong linkage disequilibrium. Suppose that a region on the genome has one very high contributing SNP, and a number of non-contributing SNPs that are in high LD with the contributing SNP. Due to being in LD with the contributing SNP, the polygenic risk score could incorrectly over-emphasize the effects of the non-contributing SNPs on the measured trait. Thirdly, due to the size of M, it is computationally infeasible to fit any higher-order interaction terms, resulting in a model that does not take into account any possible interaction effects between SNPs that could either increase or decrease the effects of the interacting SNPs on the phenotype. Fortunately, through $MonsterLM$ and the use of LD pruning, these issues can be solved and more accurate PRS calculations can be computed through using the β coefficients outputted from $monsterLM$

3.7.1 Applications of PRS

The first successful application of a polygenic risk score in humans using GWAS data was in the International Schizophrenia Consortium (2009) Consortium 2009. Using the pruning and threshold method, the study first conducted extensive pruning of the SNPs independent of their association with the measured trait. After pruning the SNPs,

the study tested setting the significance threshold at p-values of 0.1, 0.2, 0.3, 0.4, and 0.5 percent levels. International Schizophrenia Consortium (2009) performed a case-control study of 3,322 cases and 3,587 controls of individuals of European descent, the study constructed a PRS using up to approximately one million SNPs associated with risk of schizophrenia. The best resulting PRS used a threshold of 0.5 percent, and was able to explain approximately 3 of the total variability in the risk of developing schizophrenia. The study also showed that the PRS for schizophrenia is also associated with risk of bipolar disorder, but not associated with non-psychiatric diseases such as coronary artery disease, Crohn's disease, or T1/T2D. In addition, they were the first to show in a simulation study that the effectiveness of polygenic risk scores increases as sample size increases, showing that the explained variance increases from approximately 3 percent to over 20 percent by simulating a sample of 20,000 case/control pairs.

Since then, studies have applied PRS to several complex traits to test for both association and prediction of traits, with varying results. Simonson et al. (2011) (Simonson et al. 2011) aimed to use GWAS to identify common SNPs associated with cardiovascular disease (CVD) as well as to develop a PRS using the unadjusted GWAS coefficients to predict CVD risk. The results of the study was ultimately inconclusive, with no SNPs reaching the genome-wide significance level of 5×10^{-8}, while the polygenic risk score showed low predictive power, only explaining 1 percent of the variance in risk of CVD.

While many studies have used GWAS and polygenic scores, quite often the results were statistically inconclusive. Dudbridge (Dudbridge 2013) showed that previous studies with mixed to negative results were not due to the limitations of the polygenic risk score, but rather could be attributed to studies having too few participants to achieve any substantial test power or pre- dictive accuracy. Furthermore, Dudbridge (2013) went on to show that an increase in sample size is sufficient to improve both the power of association tests and the predictive accuracy of the polygenic risk scores. These conclusions

serve to further reinforce the claims made in Inter- national Schizophrenia Consortium with regards to sample size which demonstrated that predicting quantitative traits will become more precise as sample data grows in size, observing an increase in explained variance from 3 to 20. However, the sample size required was orders of magnitude greater than what was available at the time, requiring 20,000 case control pairs compared to the original study of 3,322 cases and 3,587 controls.

Despite its prevalence in current construction of polygenic risk scores, the P+T method is not without its drawbacks. The main shortfall of P+T is that SNPs that are discarded from the GWAS for not being statistically significant enough could still contribute to the overall polygenic trait, resulting in a score with reduced predictive power. Simonson et al noted in their study using polygenic models to predict CVD that "Thousands of small effects that are indistinguishable from background noise when performing a traditional GWAS can collectively account for a large proportion of the risk variation". Yang et al (Yang et al. 2010) was able to show that a model that takes into account all SNPs was able to explain a much larger pro- portion of heritability in human height as opposed to a model using only the SNPs that meet a certain significance value, showing that the most accurate polygenic scores should be built using all the SNPs simultaneously. The conclusion made by these studies is further reinforced by Ware et al (Ware et al. 2017) in a study conducted on the population-based longitudinal panel study Health and Retirement Study across four different polygenic traits: height, BMI, educational attainment, and depression. This study showed polygenic risk scores that include all available SNPs were able to either achieve a higher explained variance, or had non-inferior performance than a risk score that did not use all SNPs.

3.7.2 PRS Workflow for MonsterLM

The process for computing PRS first without interactions is as follows:

1. Select SNPs based on univariate p-values from the discovery set based on p-value thresholds as in Secondary Analysis 1).

2. Re-run $MonsterLM$ regression on the discovery set and obtain the β coefficients.

3. Using the β coefficients from 2. calculate the PRS in the validation set using equation 3.14

4. Fit a linear model to the biomarker and PRS, calculating the R^2 for the model.

This process would be repeated for the 4 sets of p-value thresholds and each biomarker, and then the set which would give the PRS the highest predictability (based on R^2), and use that set of SNPs as the basis for computing PRS with interactions.

For the computation of PRS with interactions:

1. Select significant interaction terms based on the univariate interaction p-values from the discovery set using p-value thresholds for the SNP set that passes p-value thresholds from univariate SNP analysis.

2. Re-run $MonsterLM$ regression on the discovery set with the set of SNPs from PRS computation with the selected interactions and associated SNPS, then obtain the β coefficients.

3. Using the β coefficients from 2. calculate the PRS in the validation set using equation 3.14

4. Fit a linear model to the biomarker and PRS, calculating the R^2 for the model.

This process would also be repeated for the 4 sets of p-value thresholds and each biomarker. Then the highest R^2 would be examined to see if the incorporation of interactions improve the PRS.

To calculate the PRS score with interactions the following equation was used, once the β_X and $\beta_{X\times E}$ coefficients were calculated from the discovery set:

$$PRS_{X\times E} = \sum_i (\beta_{X,i} \cdot X_i) + \sum_j (\beta_{X\times E,j} \cdot (X \times E)_j)) \tag{3.17}$$

where X_i and $(X \times E)_j$ are the individual SNPs and SNP interactions features from the validation set selected based on univariate p-values from the discovery set. By using a validation set for prediction purposes, we mitigate the potential bias that occurs from overfitting the PRS after continuously re-using the same dataset for univariate analysis, beta-coefficient computation and then re-using the coefficients on that set.

Chapter 4

Results

4.1 The Importance of Quantile Normalization

After the UKB genotype data was pruned and partitioned into 60 blocks of 25 000 contiguous SNPs per block, it was then normalized prior to fitting in $MonsterLM$.

An important assumption in $MonsterLM$ is that the model requires all covariates (including interaction terms) to have mean = 0, and variance = 1, as the model does not use an intercept term.

If the interaction terms were not quantile normalized to have mean = 0, variance = 1; then the model was observed to have a negative bias with respect to R^2 for each block. An initial test with residualized triglycerides shows the negative bias below (Table 4.1):

Table (4.2) below shows the correct estimates after properly quantile normalizing the interaction terms, as can be seen from the results; the negative bias present in each block is no longer present.

After identifying the error to be attributed to lack of normalization for interactions,

monsterLMs data-processing pipeline included a feature to quantile-normalize interactions included in analysis.

4.2 Total Interaction Variance for the Biomarkers

The 9 biomarkers: ApoB, Bilirubin, Cholesterol, Glucose, High-Density Lipoproteins, HbA1c, Low-Density Lipoproteins and Triglycerides were subsequently analyzed using *MonsterLM*. Through the use of GPU-based parallel processing, the 9 traits could be analyzed in parallel and obtaining the R^2_{GWEI} for each trait could be done within a week for a total of 2 062 250 covariates (1 031 125 variants $\times$ 2, for 1 environmental exposure) and 325 991 individuals. (Fig. 4.1), showcases the calculated interaction variances and heritabilities across all LD-pruned common variants with confidence intervals.

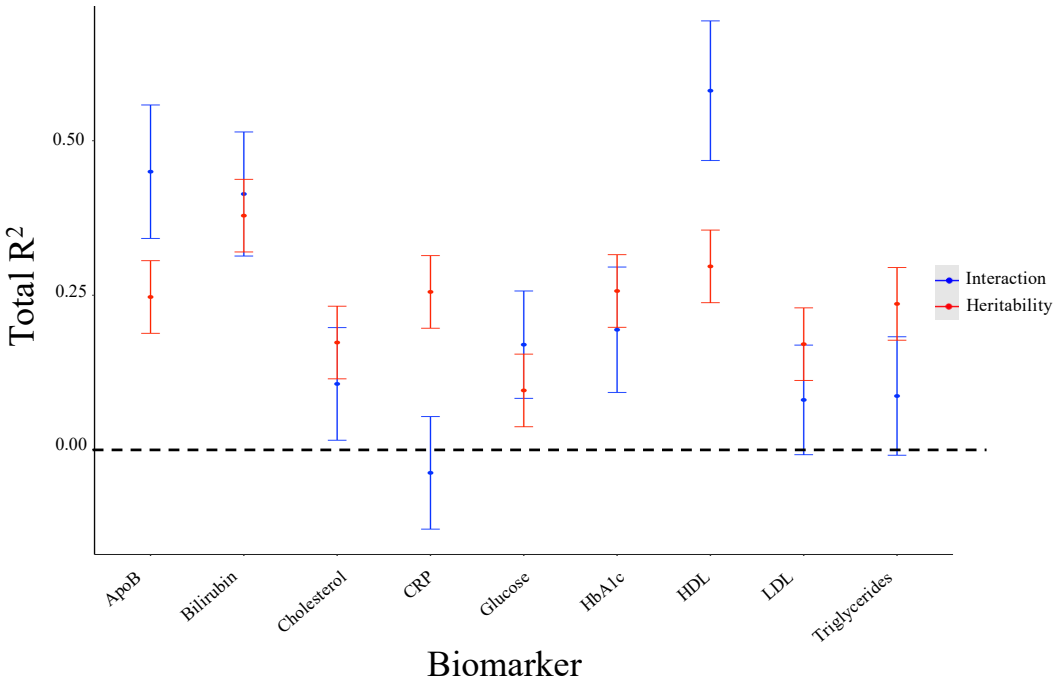

FIGURE 4.1: Interaction Variance and Heritability computed using *MonsterLM* for the residualized biomarkers, adjusted for age, sex and stratification effects, with confidence intervals calculated as in Chapter 2.

The table below (Table 4.4) summarizes the heritability and full R^2 values computed through $MonsterLM$. Although, these estimates of interaction R^2 can appear to be quite high for specific biomarkers (like ApoB and HDL); however under the results of our simulation study and based on the CI's, we are quite confident in these estimates under the assumptions outlined. However, it is important to note that CRP did not have interaction variance > 0 after fitting through $MonsterLM$ with waist-hip ratio. Since no significant interaction variance was found, it was decided that CRP wouldn't be examined further in secondary analysis. It is important to recognize this result could be purely due to chance, as the 95% CI of CRP contains positive interaction values in its range.

4.3 Univariate SNP-based Analysis

Following the results of Primary Analysis, a few questions were posed regarding the interaction variance estimates. Now that it was shown that 8 of the 9 biomarkers had rather large, significant estimates in interaction variance, we wanted to understand where this variance came from. If it was attributed to the SNPs that were associated to the trait, or perhaps unique SNPs that weren't associated with the trait were the ones interacting with waist-hip ratio.

Following the protocol outlined in Chapter 3, we split the UKB data into 2 sets: the discovery and validation set. Then, univariate regression was conducted onto the discovery set to determine which SNPs were significant to the trait of interest. Initially, multivariable regression via $MonsterLM$ was used to select for significant SNPs, but an important caveat was discovered in the use of multivariable regression in genome-based complex trait analysis. When examining the standard error for the effect size estimates (β_i), it is equal to:

$$SE = \hat{\sigma}^2(X^T X)^{-1} \tag{4.1}$$

As the number of parameters in the model becomes quite large in $MonsterLM$, the standard error on the estimates also becomes large, resulting in most SNPs (and potential interactions) having very low p-values, and very few pass even the 0.01 significance threshold. The multivariable results are shown in Table (4.5), and these SNPs captured a minimal amount of the variance explained by the MAF/LD pruned SNP set used in primary analysis.

After conducting the analysis based on the univariate regression p-values, 4 sets of SNP blocks were created in the validation set, and using $MonsterLM$ protocol to estimate heritability, especially for SNP sets that were greater than 25K SNPs in size, requiring multiple blocks. After heritability was estimated, all possible interactions with the SNPs involved were added to the model, and the total R^2_{full} was computed, and then the R^2_γ was calculated afterwards. The results are shown in Figure (4.2).

4.4 Univariate Interaction-based Analysis

From the previous analysis, we identified that the SNPs contributing to the interaction variance estimated in primary analysis were not mostly not the same SNPs that were causal to the biomarker. This prompted the next analysis described in Chapter 3.6, using univariate interaction p-values to select a set of GxE interactions passing a significance threshold in the discovery set.

With the residualized phenotype, we conducted regression with the following model:

$$e \sim \beta_{SNP} X_{SNP} + \beta_{X\times E}(X_{SNP} \times E) \tag{4.2}$$

where X was a single SNP, and the single interaction between SNP and WHR was included in the model; then interactions were selected based on the p-values that tested

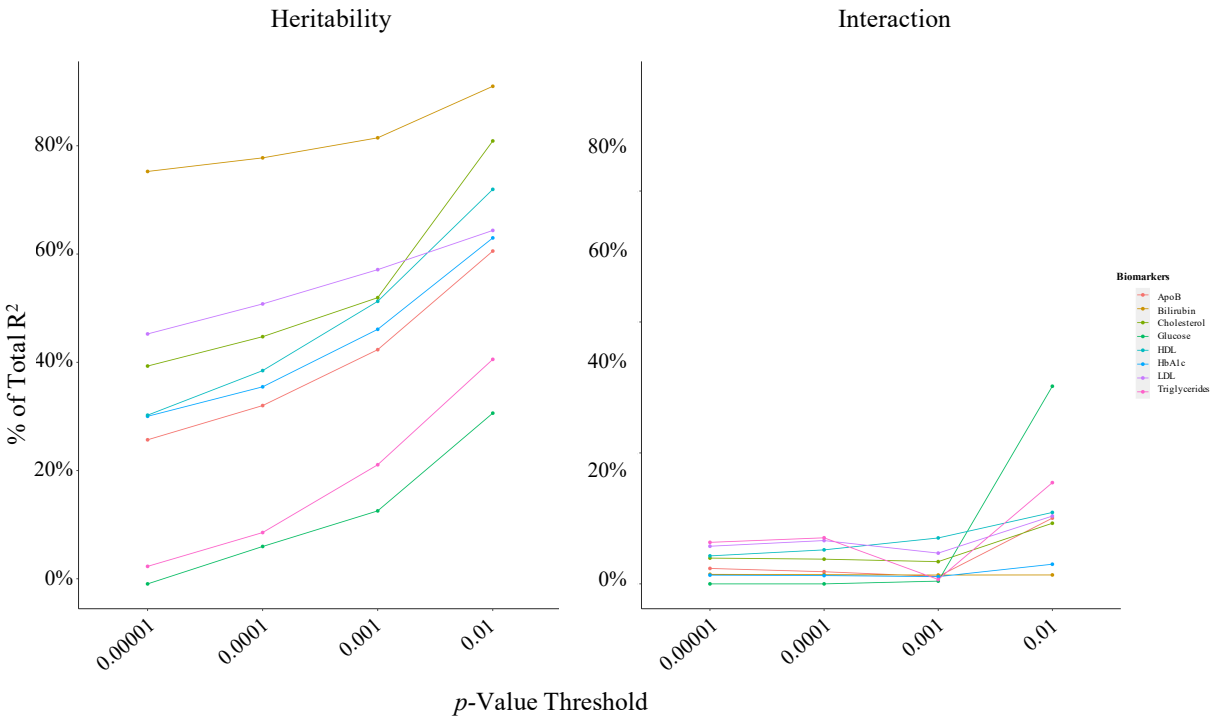

FIGURE 4.2: Fraction of Heritability and Interaction Variance Explained as a function of Significance levels for each of the 8 biomarkers used in follow-up analyses based on the validation sample

the hypotheses that:

$$H_0 : \hat{\beta}_{X\times E} = 0$$

$$H_A : \hat{\beta}_{X\times E} \neq 0$$

After selecting for these interactions based on 4 univariate interaction p-value thresholds: 0.01, 0.001, 0.0001 and 0.00001; we would create 60 new blocks to be computed through $MonsterLM$. Each block consisted of all the SNPs associated with the corresponding genetic block, and then the corresponding interactions selected from the univariate p-values would be added to the genetic SNP block, creating a new set of 60 blocks with SNPs and a proportion of interactions in the validation set. Then after fitting $MonsterLM$ to the blocks, the R^2_{GWEI} would be computed for each biomarker,

with the results displayed in Figure (4.3).

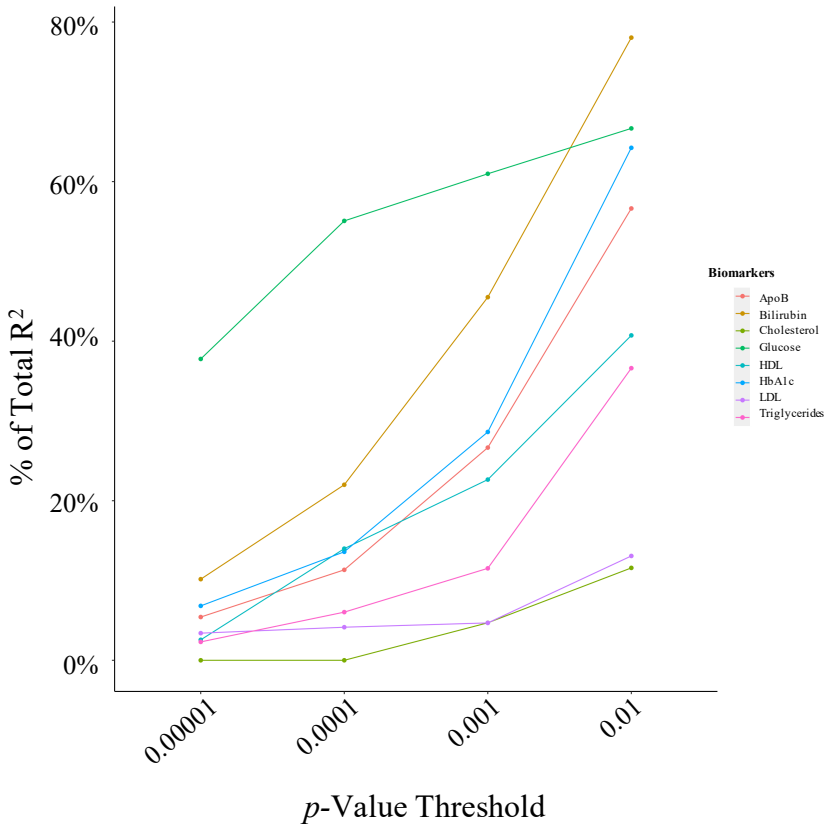

FIGURE 4.3: Fraction of Interaction Variance Explained as a function of Significance Levels for each of the 8 biomarkers

4.5 Incorporating Interactions into Polygenic Risk Scores

The final step of this project was to compute polygenic risk scores and examine if these scores could have improved predictive capabilities by incorporating interactions into computation of the PRS. Due to its relatively high interaction variance, ApoB was selected for PRS analysis including interactions with WHR.

The results for this analysis are displayed in Table (4.6). We've observed improvements in every case of incorporating interactions into the PRS score resulting in a %

increase ranging from 0.8% to 8% relative to the base PRS using only genotyping data. % increase was calculated as:

$$\%increase = \frac{R^2_{PRS+INT} - R^2_{PRS}}{R^2_{PRS}} \tag{4.3}$$

TABLE 4.1: Preliminary Results for Triglycerides without Quantile Normalization of Interactions

chr	block	adj $R^2 Full$	adj $R^2 Geno$	Interacction R^2
1	1	$1.73 \cdot 10^{-3}$	$3.76 \cdot 10^{-3}$	$-2.03 \cdot 10^{-3}$
	2	$4.15 \cdot 10^{-3}$	$6.64 \cdot 10^{-3}$	$-2.49 \cdot 10^{-3}$
	3	$6.92 \cdot 10^{-4}$	$2.64 \cdot 10^{-3}$	$-1.94 \cdot 10^{-3}$
	4	$9.61 \cdot 10^{-4}$	$3.39 \cdot 10^{-3}$	$-2.43 \cdot 10^{-3}$
2	1	$1.04 \cdot 10^{-2}$	$1.25 \cdot 10^{-2}$	$-2.12 \cdot 10^{-3}$
	2	$-1.18 \cdot 10^{-3}$	$1.2 \cdot 10^{-3}$	$-2.38 \cdot 10^{-3}$
	3	$1.75 \cdot 10^{-3}$	$4.27 \cdot 10^{-3}$	$-2.53 \cdot 10^{-3}$
	4	$5.94 \cdot 10^{-4}$	$3.42 \cdot 10^{-3}$	$-2.82 \cdot 10^{-3}$
3	1	$7.22 \cdot 10^{-4}$	$2 \cdot 10^{-3}$	$-1.28 \cdot 10^{-3}$
	2	$-1.1 \cdot 10^{-3}$	$1.17 \cdot 10^{-3}$	$-2.28 \cdot 10^{-3}$
	3	$3.38 \cdot 10^{-4}$	$1.7 \cdot 10^{-3}$	$-1.36 \cdot 10^{-3}$
	4	$6.09 \cdot 10^{-4}$	$2.13 \cdot 10^{-3}$	$-1.52 \cdot 10^{-3}$
4	1	$7.71 \cdot 10^{-4}$	$3.39 \cdot 10^{-3}$	$-2.62 \cdot 10^{-3}$
	2	$2.07 \cdot 10^{-3}$	$3.88 \cdot 10^{-3}$	$-1.81 \cdot 10^{-3}$
	3	$-1.57 \cdot 10^{-3}$	$1.25 \cdot 10^{-3}$	$-2.83 \cdot 10^{-3}$
15	1	$4.82 \cdot 10^{-3}$	$6.52 \cdot 10^{-3}$	$-1.71 \cdot 10^{-3}$
	2	$1.73 \cdot 10^{-3}$	$3.26 \cdot 10^{-3}$	$-1.53 \cdot 10^{-3}$
16	1	$1.37 \cdot 10^{-3}$	$2.15 \cdot 10^{-3}$	$-7.78 \cdot 10^{-4}$
	2	$2.32 \cdot 10^{-3}$	$4.67 \cdot 10^{-3}$	$-2.34 \cdot 10^{-3}$
17	1	$1.43 \cdot 10^{-3}$	$2.44 \cdot 10^{-3}$	$-1.01 \cdot 10^{-3}$
	2	$4.1 \cdot 10^{-3}$	$6.62 \cdot 10^{-3}$	$-2.53 \cdot 10^{-3}$
18	1	$-2.14 \cdot 10^{-4}$	$2.6 \cdot 10^{-4}$	$-4.74 \cdot 10^{-4}$
	2	$-4.33 \cdot 10^{-4}$	$2.6 \cdot 10^{-3}$	$-3.03 \cdot 10^{-3}$
19	1	$6.85 \cdot 10^{-3}$	$7.97 \cdot 10^{-3}$	$-1.12 \cdot 10^{-3}$
	2	$9.89 \cdot 10^{-3}$	$1.05 \cdot 10^{-2}$	$-5.83 \cdot 10^{-4}$
20	1	$2.58 \cdot 10^{-4}$	$1.52 \cdot 10^{-3}$	$-1.27 \cdot 10^{-3}$
	2	$2.48 \cdot 10^{-3}$	$4.28 \cdot 10^{-3}$	$-1.81 \cdot 10^{-3}$
21	1	$5.41 \cdot 10^{-4}$	$1.59 \cdot 10^{-3}$	$-1.05 \cdot 10^{-3}$
22	1	$-3.76 \cdot 10^{-4}$	$1.61 \cdot 10^{-3}$	$-1.99 \cdot 10^{-3}$
TOTAL		0.14	0.24	$-5.37 \cdot 10^{-2}$

TABLE 4.2: Preliminary Results for Triglycerides without Quantile Normalization of Interactions

chr	block	adj $R^2 Full$	adj $R^2 Geno$	Interacction R^2
1	1	$4.21 \cdot 10^{-3}$	$1.84 \cdot 10^{-3}$	$2.38 \cdot 10^{-3}$
	2	$6.09 \cdot 10^{-3}$	$3.78 \cdot 10^{-3}$	$2.32 \cdot 10^{-3}$
	3	$4.33 \cdot 10^{-3}$	$3.74 \cdot 10^{-3}$	$5.89 \cdot 10^{-4}$
	4	$3.82 \cdot 10^{-3}$	$2.99 \cdot 10^{-3}$	$8.26 \cdot 10^{-4}$
2	1	$5.53 \cdot 10^{-3}$	$4.09 \cdot 10^{-3}$	$1.44 \cdot 10^{-3}$
	2	$2.79 \cdot 10^{-2}$	$2.55 \cdot 10^{-2}$	$2.35 \cdot 10^{-3}$
	3	$4.42 \cdot 10^{-3}$	$3.26 \cdot 10^{-3}$	$1.15 \cdot 10^{-3}$
	4	$5.38 \cdot 10^{-3}$	$3.87 \cdot 10^{-3}$	$1.51 \cdot 10^{-3}$
3	1	$3.61 \cdot 10^{-3}$	$2.42 \cdot 10^{-3}$	$1.18 \cdot 10^{-3}$
	2	$2.91 \cdot 10^{-3}$	$2.11 \cdot 10^{-3}$	$7.98 \cdot 10^{-4}$
	3	$3.54 \cdot 10^{-3}$	$1.77 \cdot 10^{-3}$	$1.77 \cdot 10^{-3}$
	4	$2.83 \cdot 10^{-3}$	$1.9 \cdot 10^{-3}$	$9.25 \cdot 10^{-4}$
4	1	$5.58 \cdot 10^{-3}$	$2.95 \cdot 10^{-3}$	$2.63 \cdot 10^{-3}$
	2	$7.81 \cdot 10^{-3}$	$6.37 \cdot 10^{-3}$	$1.44 \cdot 10^{-3}$
	3	$4.62 \cdot 10^{-3}$	$3.15 \cdot 10^{-3}$	$1.47 \cdot 10^{-3}$
15	1	$6.31 \cdot 10^{-3}$	$3.85 \cdot 10^{-3}$	$2.45 \cdot 10^{-3}$
	2	$3.25 \cdot 10^{-3}$	$3.17 \cdot 10^{-3}$	$7.7 \cdot 10^{-5}$
16	1	$6.54 \cdot 10^{-3}$	$4.97 \cdot 10^{-3}$	$1.57 \cdot 10^{-3}$
	2	$3.7 \cdot 10^{-3}$	$1.53 \cdot 10^{-3}$	$2.17 \cdot 10^{-3}$
17	1	$4.49 \cdot 10^{-3}$	$2.84 \cdot 10^{-3}$	$1.65 \cdot 10^{-3}$
	2	$8.15 \cdot 10^{-3}$	$6.69 \cdot 10^{-3}$	$1.46 \cdot 10^{-3}$
18	1	$2.32 \cdot 10^{-3}$	$1.72 \cdot 10^{-3}$	$6 \cdot 10^{-4}$
	2	$2.63 \cdot 10^{-3}$	$9.3 \cdot 10^{-4}$	$1.7 \cdot 10^{-3}$
19	1	$1.73 \cdot 10^{-2}$	$1.59 \cdot 10^{-2}$	$1.37 \cdot 10^{-3}$
	2	$4.21 \cdot 10^{-3}$	$2.45 \cdot 10^{-3}$	$1.76 \cdot 10^{-3}$
20	1	$8.34 \cdot 10^{-3}$	$7.19 \cdot 10^{-3}$	$1.14 \cdot 10^{-3}$
	2	$3.41 \cdot 10^{-3}$	$1.87 \cdot 10^{-3}$	$1.54 \cdot 10^{-3}$
21	1	$3.94 \cdot 10^{-3}$	$1.79 \cdot 10^{-3}$	$2.15 \cdot 10^{-3}$
22	1	$4.64 \cdot 10^{-3}$	$3.07 \cdot 10^{-3}$	$1.57 \cdot 10^{-3}$
TOTAL		0.32	0.24	$8.69 \cdot 10^{-2}$

TABLE 4.3: Preliminary results summarizing the fitting monsterLM with triglycerides, with and without normalizing the interaction terms

Case	Just Genotype	Genotype + Interactions
Without QN	0.241004903	0.13899865
With QN	0.241004903	0.32281015

TABLE 4.4: Interaction R^2 Estimates

Biomarkers	Heritability	Total Variance	Interaction Variance
ApoB	0.25	0.7	0.45
Bilirubin	0.38	0.79	0.41
Cholesterol	0.17	0.28	0.11
CRP	0.26	0.22	$-3.75 \cdot 10^{-2}$
Glucose	$9.59 \cdot 10^{-2}$	0.27	0.17
HbA1c	0.26	0.45	0.19
HDL	0.3	0.88	0.58
LDL	0.17	0.25	$8.05 \cdot 10^{-2}$
Triglycerides	0.24	0.32	$8.69 \cdot 10^{-2}$

TABLE 4.5: Examining fraction of heritability estimates from analysis using multivariable p-values

Biomarkers	Significance	h^2_{SNPset}	Fraction of h^2_{tot}
ApoB	$1 \cdot 10^{-5}$	$2.24 \cdot 10^{-2}$	$9.08 \cdot 10^{-2}$
ApoB	$1 \cdot 10^{-4}$	$2.54 \cdot 10^{-2}$	0.1
ApoB	$1 \cdot 10^{-3}$	$4.47 \cdot 10^{-2}$	0.18
ApoB	$1 \cdot 10^{-2}$	$7.59 \cdot 10^{-2}$	0.31
Bili	$1 \cdot 10^{-5}$	0.25	0.65
Bili	$1 \cdot 10^{-4}$	0.25	0.66
Bili	$1 \cdot 10^{-3}$	0.25	0.67
Bili	$1 \cdot 10^{-2}$	0.27	0.73
Glucose	$1 \cdot 10^{-5}$	$2.99 \cdot 10^{-2}$	0.31
Glucose	$1 \cdot 10^{-4}$	$3.3 \cdot 10^{-2}$	0.34
Glucose	$1 \cdot 10^{-3}$	$3.83 \cdot 10^{-2}$	0.4
Glucose	$1 \cdot 10^{-2}$	$6.08 \cdot 10^{-2}$	0.63
Chol	$1 \cdot 10^{-5}$	$1.01 \cdot 10^{-3}$	$5.8 \cdot 10^{-3}$
Chol	$1 \cdot 10^{-4}$	$9.68 \cdot 10^{-3}$	$5.58 \cdot 10^{-2}$
Chol	$1 \cdot 10^{-3}$	$1.25 \cdot 10^{-2}$	$7.19 \cdot 10^{-2}$
Chol	$1 \cdot 10^{-2}$	$2.41 \cdot 10^{-2}$	0.14
HDL	$1 \cdot 10^{-5}$	$3.82 \cdot 10^{-2}$	0.13
HDL	$1 \cdot 10^{-4}$	$5.27 \cdot 10^{-2}$	0.18
HDL	$1 \cdot 10^{-3}$	$6.11 \cdot 10^{-2}$	0.21
HDL	$1 \cdot 10^{-2}$	0.1	0.34
HbA1c	$1 \cdot 10^{-5}$	$2.65 \cdot 10^{-2}$	0.1
HbA1c	$1 \cdot 10^{-4}$	$3.26 \cdot 10^{-2}$	0.13
HbA1c	$1 \cdot 10^{-3}$	$4.35 \cdot 10^{-2}$	0.17
HbA1c	$1 \cdot 10^{-2}$	$7.12 \cdot 10^{-2}$	0.28
LDL	$1 \cdot 10^{-5}$	$4.75 \cdot 10^{-2}$	0.28
LDL	$1 \cdot 10^{-4}$	$5.23 \cdot 10^{-2}$	0.31
LDL	$1 \cdot 10^{-3}$	$5.47 \cdot 10^{-2}$	0.32
LDL	$1 \cdot 10^{-2}$	$7.43 \cdot 10^{-2}$	0.43
TG	$1 \cdot 10^{-5}$	$3.21 \cdot 10^{-2}$	0.14
TG	$1 \cdot 10^{-4}$	$3.28 \cdot 10^{-2}$	0.14
TG	$1 \cdot 10^{-3}$	$4.43 \cdot 10^{-2}$	0.19
TG	$1 \cdot 10^{-2}$	$7.51 \cdot 10^{-2}$	0.32

TABLE 4.6: Comparing polygenic risk scores with and without interactions included at the significance levels displayed

Int Significance	Geno Significance	PRS R^2	PRS R^2 + Int	Percent Increase in R^2
0.00001	0.00001	0.1137	0.1146	0.79
0.00001	0.0001	0.1101	0.1164	5.41
0.00001	0.001	0.1011	0.10644	5.02
0.00001	0.01	0.0813	0.08204	0.9
0.0001	0.00001	0.1137	0.1147	0.87
0.0001	0.0001	0.1101	0.1163	5.33
0.0001	0.001	0.1011	0.1066	5.16
0.0001	0.01	0.0813	0.08268	1.67
0.001	0.00001	0.1137	0.1157	1.73
0.001	0.0001	0.1101	0.1185	7.09
0.001	0.001	0.1011	0.1075	5.95
0.001	0.01	0.0813	0.08328	2.38
0.01	0.00001	0.1137	0.1165	2.4
0.01	0.0001	0.1101	0.1197	8.02
0.01	0.001	0.1011	0.1086	6.91
0.01	0.01	0.0813	0.08268	1.67

Chapter 5

Discussion

5.1 Discussion of Primary Analysis: Interaction Results

Interactions between obesity-related environmental exposures and genes relating to lipids is documented (Walker and Jebb 2012). An adverse lipid profile of high triglycerides, total cholesterol and low-density lipoprotein cholesterol; and low HDL cholesterol has been observed to have genetic interactions in relation to cardiovascular disease Wilson et al. 1998. At a population level, diet is an important determinent of the lipid profile, but a great amount of inter-individual variation has been noted consistently, especially in the lipid responses to dietary changes. This variation has been attributed to differences in genetic background, in addition to age, gender and ethnicity. Reductions in total fat, saturated fatty acids and trans-fatty acids are all effective in reducing total cholesterol, LDL cholesterol and triglycerides (Keys and Parlin 1966). However, a reduction in total dietary fat intake achieved by replaced with carbohydrates has been associated with an unfavorable reduction in HDL cholesterol and increased triglycerides (Poppitt et al. 2002). Conversely, replacement of trans- or saturated fatty acids has been found to maintain HDL cholesterol. Increased dietary intake of n-3 polyunsaturated fatty acids has been associated with lower triglycerides. In addition to changes in dietary composition, a reduction in total energy leading to weight loss is associated with reductions in

cholesterol and triglycerides, and an increase in HDL cholesterol associated with weight loss (Dattilo and Kris-Etherton 1992, Poobalan et al. 2004). However, population-level effects of diets on lipid levels conceal significant heterogeneity at an individual level, which is often attributed in part to genetic variation. Heritability estimates for variation in lipid levels within a population range from 10 percent to 60 percent, which our heritability estimates are in line with.

Summarizing current genetic studies, the gene variants identified in apolipoproteins and genes involved in cholesterol synthesis and efflux, lipolysis, lipogenesis, triglycerides and cholesterol transfer, explaining some of the variation in lipid levels are $APOE, CETP$, and $APOA5$ (Corella and Ordovas 2005). It is quite interesting to note that in these studies, the genes examined are those that are already found associated with the trait through GWAS studies, where a single gene is examined at a time with the trait.

The literature described thus far explains some of the high interaction variance found in the lipoproteins and triglycerides from our primary results, however there is not as much literature available on the interactions of another biomarker: bilirubin with obesity-related exposure factors; however the literature does show the biomarker being associated to T2D, which increases risk to CVD by 5-fold. Another biomarker of interest is blood glucose levels as measured by glycated haemoglobin A1c, for its interaction levels with obesity-related exposure factors. A recent study examined some potential interactions between genetics and diet on haemoglobin A1c and type 2 diabetes risk (Eriksen et al. 2019). Statistical interactions were observed between gene risk scores for HbA1c and intake of wholegrains, with a p-value < 0.04 for the interaction coefficient with respect to risk of T2D. The effect of the diet interactions on HbA1c was greater in high-genetic risk individuals compared to lower genetic risk individuals. Furthermore significant interactions between gene-risk scores related to HbA1c and BMI categories were also observed. There was a greater effect of the gene-risk score on HbA1c among

obese candidates for interaction p-value < 0.03. A caveat of these results is that the analysis was only based on a single gene-risk score derived from 87 common SNPs associated with HbA1c; however from our primary analysis, there are many more SNPs that are associated with HbA1c that may interact specifically with an environmental exposure, and they may be identified through the use of $MonsterLM$ using a database with an extremely large cohort in the UK Biobank.

5.2 Interaction Variance Analysis Comparisons

Alternative methods that examine gene-environment interactions include the genetic-covariate interaction: genome-based restricted maximum likelihood method, denoted as:GCI-GREML which was developed(Robinson et al. 2017). This method defines GxE interaction in a very narrow sense as genotype–covariate interaction is detected if changes in genetic effects do not scale proportionally with changes in mean or phenotypic variance across ages as a time-dependent factor. Furthermore, the method computes multiple likelihood ratio tests (>250 tests), requiring a very high p-value threshold after Bonferroni correction, potentially rejecting many interactions with slightly weaker associations, but still having an impact. Their analysis focused specifically on age, smoking and diet exposures with respect to BMI variance, which is different than what our analysis focused on, making the primary results not completely comparable.

Another method that is very similar to $MonsterLM$ in computing GxE interaction effects was recently published as: GxE Mixed Model (Dahl et al. 2020. The method also characterizes the aggregate polygenic contributions of GxE. Also like $MonsterLM$; relative to single-variant GxE tests, GxEMM has higher power for polygenic traits but it differs in having lower resolution. Regardless, both methods are particularly useful for characterizing the biological relevance of environmental measures. GxEMM is limited in how many SNPs can be analyzed and the sample size cannot be very large, limiting the

use of asymptotic properties of estimators. This is a big limitation as large sample sizes are required in order to have well-powered analyses. When an analysis is underpowered, many results could be hidden as the number of samples is insufficient to accurately observe the overall variance explained by interactions, and other such quantities. Another limiting factor in this analysis, is the use of traditional OLS estimation, requiring the computationally intensive inverse of a large matrix to be computed, greatly increasing computation time as the number of SNPs and sample size increase to large numbers. Through the use of the conjugate gradient method and GPU systems, $MonsterLM$ can compute estimates faster than these methods for much larger datasets.

5.3 Heritability Analysis Comparisons

Furthermore, all heritability estimates are in similar range to previous literature on the heritability of these biomarkers (Bulik-Sullivan et al. 2015) in the same population, further validating the results of primary analysis, with the only notable exception in ApoB (Table 5.2). Another method quite similar to $MonsterLM$ in the context of estimating heritability is the generalized random effects model (Kangcheng et al. 2019) which doesn't prune for LD but uses a generalized inverse to take into account the LD structure has very similar results for traits like BMI and WHR, and identical results for the only biomarker examined Low-density cholesterol at 0.083. Table (5.1). However, their analysis also filtered out a great deal of SNPs from the UK Biobank, as our analysis computed more than 2.5 million SNPs, while theirs comprised of roughly 500K SNPs. Furthermore, their method doesn't examine the potential of computing interaction effects between exposures and SNPs, as the computational burden to compute twice as many features becomes a great deal higher. However; the purpose of describing the GRE method was to validate the method of $MonsterLM$ compared to traditional methods that rely on univariate regression or variance component computations.

LD score regression was another method used to estimate heritability in the comparison analysis; which differs quite a bit from $MonsterLM$, in that heritability is estimated from univariate regression from GWAS for each phenotype, and then the resulting heritability estimate for each SNP is corrected for each estimate by using the LD score for each SNP, which is the sum of Pearson correlations between all other SNPs associated with the current SNP being analyzed. This method uses summary statistics to estimate heritability instead of individual-level data, which provide less precise estimates. The purpose of comparing these methods is not to just state that $MonsterLM$ provides better estimates than LDSR, but more-so have a reference point to compare the heritability estimates for biomarkers in the UK Biobank cohort, where such estimates are currently limited.

Since $MonsterLM$ takes into account many more SNPs in analysis within a single block and relied on variance component methods due to the M much greater than N problem. Current methods also make many assumptions on genetic architecture such as polygenicity (the number of variants with effects larger than some small constant δ) and MAF/LD-dependence. Our model makes no such assumptions in the genetic data as our data is pruned for SNPs high in LD and have a MAF higher than 0.05. Under most models, there are additional assumptions on the distribution of β_i and the form of σ_i^2 that are unnecessary under $MonsterLM$.

TABLE 5.1: Comparing $MonsterLM$ with a similar model: generalized random effects model

Trait	$MonsterLM$	GRE
BMI	0.31	0.29
WHR	0.18	0.17
Cholesterol	0.083	0.082

TABLE 5.2: Comparing heritability estimates from the UK Biobank for the biomarkers

Biomarkers	MonsterLM h^2	Neale Lab h^2
ApoB	0.25	0.1
Bilirubin	0.38	0.44
Cholesterol	0.17	0.12
Glucose	$9 \cdot 10^{-2}$	$9 \cdot 10^{-2}$
HDL	0.3	0.33
HbA1c	0.26	0.2
LDL	0.17	0.1
TG	0.24	0.22

5.4 Discussion of Secondary Analysis

5.4.1 Univariate SNP-based Regression Discussion

From the results of Univariate SNP-based Regression (4.2), a few key observations are noted. The first observation is that R^2 increases consistently for heritability as more SNPs are included (as a result of increasing the significance threshold). However, the change in R^2 for interaction variance does not follow such a relationship, there are notable counterexamples in triglycerides and LDL where there is a decrease in R^2 for interaction variance after increasing the significance threshold from 0.0001 to 0.001.

The other major observation is that while a small subset of SNPs can capture between 50-80% of the total SNP-based heritability (h^2). The number of SNPs varied between 1000 at the 0.00001 significance level to 38 000 - 40 000 SNPs at the 0.01 significance level for the various biomarkers. Considering the original SNP set was of 1 031 025 SNPs, an enrichment of $\sim$0.1 to 4% of SNPs carrying the majority of variance is a well-documented but interesting observation. However, the same cannot be said for the interaction variance in these results. In fact, the interactions with the SNP set selected from univariate SNP-based p-values, captured very little of the interaction variance. At the 0.01 signficance level, only 10% of the interaction variance was captured by the

interactions between SNPs associated with the biomarker of interest and WHR. This result suggests that the SNPs that interact with the environmental exposure: WHR, are mostly different than the SNPs that are associated with the biomarker of interest. From this finding, we sought to investigate where the interaction variance could be attributed to in the next univariate regression analysis.

5.4.2 Interaction-based Univariate Regression Discussion

From these results (4.3, we see that, similar to the first univariate SNP-based analysis, the most significant interactions capture between 50 - 70 percent of the interaction variance observed in primary analysis at the 0.01 significance level which includes 3-5% of total SNPs in each block that have significant interactions. R^2 is consistently increasing at each significance level , however the rate of that increase is not as consistent as there are dramatic increases in R^2 for traits like LDL, triglycerides and total cholesterol. There are 2 main observations to take away from these results. The first is that as initially observed from the SNP-based analysis, the majority of interaction variance explained is due to SNPs that are either weakly associated or not associated with the biomarker of interest directly interacting with WHR to produce significant GxE interaction effects. This is because the univariate SNP-based analysis shown previously had low enrichment in interaction variance, which primarily examined the SNPs strongly associated with the biomarkers. Another key observation is that the this high percent of interaction variance is captured by a much smaller number of SNPs interacting with WHR. On a per block basis, there are approximately 25 000 SNPs in a block; however, the number of SNPs that have significant interactions with WHR at the 0.01 significance level ranges from 1000 to 1250. This means roughly 5 percent of SNPs with interactions are able to capture majority of the interaction variance. To exclude the observed enrichment being due to chance alone, we randomly selected the same number of SNPs irrespective of their interaction p-values. We tested ApoB due to its high interaction variance (40%)

and the observation that 72% was explained only by $\sim$ 50 000 SNPs (representing 5% of total SNPs; Table 5.3). A subset of results for this test for false positives is shown in Table (5.3). The total R^2_{int} for the randomly selected SNPs + interaction sets captured roughly 1 percent of the interaction variance for ApoB observed in primary analysis, while the SNPs + interaction sets passing the 0.01 significance threshold captured 70 percent of the interaction variance. This confirms the reproducibility of results and that the majority of interaction variance can be attributed to a small percentage of SNPs interacting with WHR that are either weakly or not associated directly to the biomarker of interest.

These findings are quite novel and not reported in most literature, in fact most analyses examine only SNPs associated with the trait of interest when estimating GxE interactions (Sulc et al. 2020), resulting in only minor estimates of GxE interactions. The unexplained variance in most interactions model being attributed to these weakly associated SNPs could provide quite novel information for predicting disease risk and personal medicine based on these newly identified interactions with SNPs and environmental exposures.

TABLE 5.3: Comparing interaction variance results between the set of SNPs passing an interaction p-value threshold of 0.01 and a set of SNPs randomly selected, where the R^2 was estimated per block analysed, displayed below are results for chromosomes 10-19

chr	block	h^2	ApoB R^2_{full}	R^2_{int}	Random R^2_{full}	Random R^2_{int}
10	1	$3.36 \cdot 10^{-3}$	$7.47 \cdot 10^{-3}$	$4.11 \cdot 10^{-3}$	$3.15 \cdot 10^{-3}$	$-2.16 \cdot 10^{-4}$
	2	$-1.6 \cdot 10^{-3}$	$3.66 \cdot 10^{-3}$	$5.25 \cdot 10^{-3}$	$-1.46 \cdot 10^{-3}$	$1.39 \cdot 10^{-4}$
	3	$3.96 \cdot 10^{-3}$	$7.91 \cdot 10^{-3}$	$3.95 \cdot 10^{-3}$	$4.26 \cdot 10^{-3}$	$2.95 \cdot 10^{-4}$
11	1	$1.86 \cdot 10^{-3}$	$5.04 \cdot 10^{-3}$	$3.17 \cdot 10^{-3}$	$2.16 \cdot 10^{-3}$	$3 \cdot 10^{-4}$
	2	$8.04 \cdot 10^{-3}$	$1.06 \cdot 10^{-2}$	$2.55 \cdot 10^{-3}$	$8.3 \cdot 10^{-3}$	$2.62 \cdot 10^{-4}$
	3	$9.1 \cdot 10^{-3}$	$1.34 \cdot 10^{-2}$	$4.32 \cdot 10^{-3}$	$9.52 \cdot 10^{-3}$	$4.28 \cdot 10^{-4}$
12	1	$-2.09 \cdot 10^{-3}$	$2.61 \cdot 10^{-3}$	$4.71 \cdot 10^{-3}$	$-2.9 \cdot 10^{-3}$	$-8.07 \cdot 10^{-4}$
	2	$-3.43 \cdot 10^{-3}$	$3.34 \cdot 10^{-3}$	$6.77 \cdot 10^{-3}$	$-2.16 \cdot 10^{-3}$	$1.27 \cdot 10^{-3}$
	3	$9.76 \cdot 10^{-3}$	$1.19 \cdot 10^{-2}$	$2.09 \cdot 10^{-3}$	$1.16 \cdot 10^{-2}$	$1.84 \cdot 10^{-3}$
13	1	$5.21 \cdot 10^{-4}$	$6.52 \cdot 10^{-3}$	$6 \cdot 10^{-3}$	$6.76 \cdot 10^{-4}$	$1.56 \cdot 10^{-4}$
	2	$-4.15 \cdot 10^{-4}$	$3.75 \cdot 10^{-3}$	$4.17 \cdot 10^{-3}$	$-2.81 \cdot 10^{-5}$	$3.87 \cdot 10^{-4}$
14	1	$1.28 \cdot 10^{-4}$	$5.77 \cdot 10^{-3}$	$5.64 \cdot 10^{-3}$	$2.09 \cdot 10^{-3}$	$1.96 \cdot 10^{-3}$
	2	$4.63 \cdot 10^{-3}$	$9.74 \cdot 10^{-3}$	$5.11 \cdot 10^{-3}$	$3.97 \cdot 10^{-3}$	$-6.63 \cdot 10^{-4}$
15	1	$1.75 \cdot 10^{-2}$	$2.18 \cdot 10^{-2}$	$4.29 \cdot 10^{-3}$	$1.85 \cdot 10^{-2}$	$1.01 \cdot 10^{-3}$
	2	$3.14 \cdot 10^{-3}$	$7.32 \cdot 10^{-3}$	$4.18 \cdot 10^{-3}$	$2.49 \cdot 10^{-3}$	$-6.46 \cdot 10^{-4}$
16	1	$5.99 \cdot 10^{-3}$	$9.34 \cdot 10^{-3}$	$3.36 \cdot 10^{-3}$	$6.88 \cdot 10^{-3}$	$8.87 \cdot 10^{-4}$
	2	$2.89 \cdot 10^{-2}$	$3.36 \cdot 10^{-2}$	$4.76 \cdot 10^{-3}$	$2.83 \cdot 10^{-2}$	$-5.68 \cdot 10^{-4}$
17	1	$5.48 \cdot 10^{-3}$	$8.7 \cdot 10^{-3}$	$3.22 \cdot 10^{-3}$	$6.7 \cdot 10^{-3}$	$1.22 \cdot 10^{-3}$
	2	$1.58 \cdot 10^{-2}$	$1.98 \cdot 10^{-2}$	$3.97 \cdot 10^{-3}$	$1.74 \cdot 10^{-2}$	$1.52 \cdot 10^{-3}$
18	1	$3.49 \cdot 10^{-4}$	$3.15 \cdot 10^{-3}$	$2.8 \cdot 10^{-3}$	$-4.83 \cdot 10^{-4}$	$-8.32 \cdot 10^{-4}$
	2	$1.28 \cdot 10^{-2}$	$1.72 \cdot 10^{-2}$	$4.33 \cdot 10^{-3}$	$1.44 \cdot 10^{-2}$	$1.56 \cdot 10^{-3}$
19	1	$4.98 \cdot 10^{-3}$	$8.14 \cdot 10^{-3}$	$3.16 \cdot 10^{-3}$	$5.05 \cdot 10^{-3}$	$6.89 \cdot 10^{-5}$
	2	$9.36 \cdot 10^{-3}$	$1.35 \cdot 10^{-2}$	$4.1 \cdot 10^{-3}$	$1.01 \cdot 10^{-2}$	$7.33 \cdot 10^{-4}$
TOTAL		0.28	0.53	0.25	0.3	$1.94 \cdot 10^{-2}$

5.5 Discussion of PRS Results

The PRS analysis shows that incorporation of interactions between genotype data and environmental exposures can improve polygenic risk scores. Although the effect may seem modest, power to identify specific interactions is low and thus the improvement can be expected to increase with larger sample size. Another notable insight is that if a p-value that is too liberal is used for either interaction or SNP-based univariate selection, then the predictiveness in the validation sample decreases. There is an optimal

zone of adding interactions at the 0.0001 significance level, where adding or removing interactions decreases the PRS R^2 in the validation set. Examining the literature, the incorporation of interactions into gene risk scores has been approached a variety of ways. A recent study (Sulc et al. 2020) examined the incorporation of interactions with environmental exposures, specifically the interaction between a single gene-risk score that aggregates variant data with a set of lifestyle factors that reflect obesity, with respect to BMI as their trait of interest. The first difference between our approaches in computing risk scores with interactions, is that while our method examines the interactions between each individual SNP interacting with the environmental exposure; while the study done by the Kutalic lab (Sulc et al. 2020) focused specifically on interactions with the risk-score itself and the exposure. The major difference is that this method is estimating the proportion of variance potentially explained by interactions of PRS with all environmental factors, estimating the increased predictiveness of PRS with a perfect knowledge of environment interactions with PRS. Both methods saw an increase in R^2 between 0.1 to 1%. While this other method may be more computationally efficient (as it only looks at the interaction with just the gene risk score), our method has the advantage of examining individual SNP interactions with the exposure, which has further applications in pathway analysis for specific gene-lists. Furthermore, as seen in previous secondary analysis, it was observed that majority of interaction variance is attributed to SNPs that were either weakly or not associated with the biomarker analysed. This result could explain why the increase in R^2 for PRS that included interactions for both of our analyses were quite small. Our results are a very important proof-of-concept that show the way for incorporation of GxE in PRS. Recent studies focus on the use of environmental interactions with the PRS itself, when in fact our approach shows that interactions come from incorporation of SNPs interacting with the environmental exposure.

Chapter 6

Conclusion and Future Directions for Research

6.1 Conclusion

In this work, we showcased the novel method we developed *MonsterLM* and its usefulness in understanding genome-wide interactions with environmental exposures. We first tested monsterlm with simulations using LD-pruned and MAF filtered real genotypes from the UK Biobank. During simulations we verified the properties of monsterlm to estimate the variance explained by interaction terms. We also considered multiple cases that could be encountered with real data regarding the dependency of the environmental exposure on SNPs.

Our results demonstrate the presence of Gene x Environment interactions between blood biochemistry biomarkers such as HbA1c, Triglycerides and ApoB with an environmental factor relating to obesity-related lifestyle factor: Waist-hip Ratio (WHR). The estimates for 8 of the 9 biomarkers analysed were significant, and prompted further analysis into these R^2 estimates ranging from 0.08 to 0.58. Furthermore, we verified the

heritability estimates from our method with other published methods with comparable and consistent results.

We further investigated these results through univariate p-value selection identify the distribution of interaction variance across SNPs. We observed high SNP enrichment within the genome 50-70% of interaction variance explained was through by 5% of total SNPs in the genome, and further observed that these SNPs were either weakly or not associated with the biomarker of interest. Lastly, we demonstrated the impact of interactions on improving polygenic risk scores for ApoB, as incorporation of interactions would improve PRS overall by 1% and a percent increase ranging from 0.8% to 8% with respect to baseline PRS.

As described by (**GxEPare)**, the collection of large amounts of genetic and phenotypic data has enabled investigators to examine their databases and explore the existence of modulations by environmental exposures for genetic association signals. Past methods involved using interaction meta-analyses on data from multiple cohorts. These methods may have been successful for genetic association studies, meta-analysis may not work well in the context of GxE interactions due to the diversity of measurements and data across cohorts, which degrades statistical power (Ahmad 2013). Through the use of $MonsterLM$ on a large cohort like the UK Biobank, we've followed a logical method to accurately measure interaction effects since this process is well-validated and uses standardized assessment methods.

Emphasis is frequently placed on translation when GxE is discussed, as identifying genetic markers that define patients who are substantially greater or lesser risk of disease than the general population given exposure to modifiable risk factors, or who will respond better or worse to treatments could optimize medical interventions. However, there are no translatable examples of GxE that are sufficiently convincing to guide medical interventions for T2D (Pare and Franks 2016). Fortunately, numerous examples from

Mendelian disorders and pharmacogenetics fuels hope that genetic data may eventually help tailor prevention or treatment strategies for complex diseases focused on lifestyle modifications.

6.1.1 Limitations

A limitation to $MonsterLM$ is the requirement to remove predictors that are perfectly or nearly perfectly correlated, which is done through LD pruning in our analysis. Another limitation of $MonsterLM$ is the assumption that the phenotypic trait and environmental exposure are both continuous variables. As many environmental exposure variables are categorical in nature, it is important to circumvent this issue, a potential solution is to use the liability scale to work with traits that are binary or categorical onto a continuous scale (Falconer 1965). The liability scale can convert the heritability of the observed binary trait and the more interpretable heritability of the continuous liability is computed as (Lee et al. 1965):

$$h^2_{liability} = h^2_{observed} \frac{K(1-K)(K(1-K)}{\psi(\Phi^{-1}[K]^2(P(1-P))}$$

where K is the frequency of the binary trait in the population, P is the frequency of the binary trait in the observed sample, and the denominator of the first fraction is the squared probability density function evaluated at the K quantile of the inverse cumulative density function of the standard normal distribution. Though this formula does not directly apply to the case of GxE interactions, and further derivations will need to conducted in order to solve this potential issue.

6.2 Future Analyses

The first analysis to conduct following this thesis study is to examine how interacting SNPs weakly or not associated with the trait would improve PRS R^2. Past literature also examined interactions only with SNPs that were significantly associated with the trait

of interest, and thus found small improvements in PRS R^2. Given that our secondary analysis showed strong enrichment in interaction variance explained by weakly associated SNPs, there is the potential that interactions with these SNPs can have stronger improvements in PRS R^2, which could have great clinical applications in predicting disease risks and further understanding these blood biochemistry biomarkers.

6.2.1 Pathway Analysis

After completing the remaining analysis with PRS scores, a further analysis to be done is pathway analysis; examining the genes to which the SNPs strongly interacting with WHR belong to and examining the potential protein pathways the genes are associated with. By examining the gene sets containing SNPs with strong interactions with WHR, we can potentially identify novel functions of proteins, which can further aid clinical applications and interventions for individuals who are identified as high risk for particular diseases.

Appendix A

Supplementary: Conjugate Gradient Algorithm

Suppose we want to solve the system of linear equations:

$$Ax = b \tag{A.1}$$

for the vector x, where the known $n \times n$ matrix A is symmetric, positive-definite and real, and b is known is well.

We say that two non-zero vectors u and v are conjugate with respect to A if:

$$u^T Av = 0 \tag{A.2}$$

Since A is symmetric and positive-definite, the left-hand side defines an inner product:

$$u^T Av = \langle u, v \rangle_A = \langle Au, v \rangle = \langle u, A^T v \rangle = \langle u, Av \rangle \tag{A.3}$$

Two vectors are conjugate if and only if they are orthogonal with respect to this

inner product. Being conjugate is a symmetric relation: if u is conjugate to v, then v is conjugate to u. Suppose then:

$$P = [p_1, ..., p_n]$$

is a set of n mutually conjugate vectors (with respect to A). Then P forms a basis for R^n, and we may express the solution x_* of $Ax = b$ in this basis (Gestenes and Stiefel 1952):

$$x_* = \sum_i^n \alpha_i p_i \tag{A.4}$$

Based on this expansion, we compute:

$$Ax_* = \sum_i^n \alpha_i A p_i \tag{A.5}$$

Left-multiplying by p_k^T:

$$p_k^T A x_* = \sum_i^n \alpha_i p_k^T A p_i \tag{A.6}$$

Substituting $Ax_* = b$ and $u^T A v = \langle u, v \rangle_A$:

$$p_k^T b = \sum_i^n \alpha_i \langle p_k, p_i \rangle_A \tag{A.7}$$

Then $u^T v = \langle u, v \rangle$ and for all i, k where $i \neq k$ $\langle u, v \rangle_A = 0$ yields:

$$\langle p_k, b \rangle = \alpha_k \langle p_k, p_k \rangle_A \tag{A.8}$$

which implies:

$$\alpha_k = \frac{\langle p_k, b \rangle}{\langle p_k, p_k \rangle_A}$$

Which gives a method to solve the linear system: find a sequence of n conjugate directions, and then compute the coefficients α_k

A1 The iterative method of CG Analysis

If we choose the conjugate vectors p_k carefully, then we don't need to compute all of them to obtain a good approximation to the solution. This also allows us to solve systems where n is so large that the direct method would take too much time (Barrett et al. 1994).

As state in chapter 3, we denote the initial guess for x_* now denoted as β_* by β_0. Starting with β_0 we search for the solution and in each iteration we need a memtric to tell us whether we are closer to the solution β_*. The metric comes from the unique minimizer of the modified normal equation:

$$f(\beta) = \frac{1}{2}\beta^T A\beta - \beta^T b, \beta \in R^n \tag{A.9}$$

The existence of a unique minimizer is confirmed as its second derivative is given by A, a symmetric positive-definite matrix, and that the minimizer solves the first derivative:

$$\nabla f(\beta) = A\beta - b. \tag{A.10}$$

This suggests taking the first basis vector p_0 to be the negative of the gradient of f at $\beta = \beta_0$. Starting with an initial guess β_0, this means we take $p_0 = b - A\beta_0$. The other vectors in the basis will be conjugate to the gradient. Note that p_0 is also the residual provided by the initial step of the algorithm described in chapter 3.

A2 GPU Acceleration

Due to the sparsity of the $X^T X$ matrix (noted as A), we can take advantage of this structure in order to apply the CG method in a very fast iterative manner. The sparse matrix–vector multiplication (SpMV) of the matrix A and the vector β is another time consuming operation in the CG algorithm. The implementation naturally depends on

the storage format of A. There are several popular storage formats such as the coordinate (COO) format, the compressed sparse row (CSR) format, the diagonal (DIA) format, the ellpack (ELL) format and some others. Due to symmetry of A, it is enough to store only the main diagonal and the two sub-diagonals, through manipulation of memory storage during the iterations of computations for the large A matrix, the computation of β becomes much much faster.

Bibliography

Ahmad, S. (2013). Gene x physical activity interactions in obesity: combined analysis of 111 421 individuals of European ancestry. *PLoS Genetics* 9, e1003607.

Ahmad, S., Varga, T., and Franks, P. (2013). Gene x environment interactions in obesity: the state of the evidence. *Hum Hered* 15(75), 106.

Alf Jr, E. and Graf, R. G. (1999). Asymptotic Confidence Limits for the Difference Between Two Squared Multiple Correlations: A Simplified Approach. *Psychological Methods* 4(1), 70–75.

Annicchiarico, P. and Mariani, G. (1996). Prediction of adaptability and yield stability of durum wheat genotypes from yield response in normal and artificially drought-stressed conditions. *Field Crop Res* 46(80), 71.

Azen, R., Budescu, D., and Reiser, B. (2001). Criticality of predictors in multiple regression. *British Journal of Mathematical and Statistical Psychology* 54, 201–225.

Barrett, R., Berry, M., Chan, T., Demmel, J., Donato, J., Dongarra, J., Eijkhout, V., Pozo, R., Romine, C., and Vorst, H. van der (1994). Iterative methods for sparse linear systems. *SIAM*, 13.

Bohacek, J., Kharicha, A., Ludwig, A., Wu, M., Holzmann, T., and Karimi-Sibaki, E. (2019). A GPU Solver for Symmetric Positive-Definite Matrices vs. Traditional Codes. *Computers Mathematics with Applications* 78(9), 2933–43.

Bolstad, B., Irizarry, R., Astrand, M., and Speed, T. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19(2), 185–193.

Bretscher, O. (1995). *Linear Algebra With Applications*. Vol. 3. Prentice Hall.

Budescu, D. (1993). Dominance analysis: a new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin* 114, 542–551.

Bulik-Sullivan, B., Loh, P., Finucane, H., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium Patterson, N., Daly, M., Price, A., and Neale, B. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *bioRxiv* 47(3), 291–295.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., and O'Connell, J. e. a. (2018). The UK biobank resource with deep phenotyping and genomic data. *Nature* 562, 203.

Cohen, J. and Cohen, P. (1983). Applied multiple regression/correlation analysis for the Behavioural Sciences. *Essays in probability and statistics*.

Consortium, I. S. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 460(7256), 748–752.

Corella, D. and Ordovas, J. (2005). Single nucleotide polymorphisms that influence lipid metabolism: interaction with dietary factors. *Annu Rev Nutr* 25, 341–90.

Dahl, A., Nguyen, K., Cai, N., Gandal, M., Flint, J., and Zaitlen, N. (2020). A Robust Method Uncovers Significant Context-Specific Heritability in Diverse Complex Traits. *The American Journal of Human Genetics* 106, 71–91.

Dattilo, A. and Kris-Etherton, P. (1992). Effects of weight reduction on blood lipids and lipoproteins: a meta-analysis. *Am J Clin Nutr* 56, 320–28.

Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLoS Genetics* 9(3).

Eriksen, R., Gibson, R., Aresu, M., Heard, A., Chan, Q., Evangelou, E., Gao, H., Elliott, P., and Frost, G. (2019). Gene-diet quality interactions on haemoglobin A1c and type 2 diabetes risk: The Airwave Health Monitoring Study. *Endocrinology, diabetes metabolism* 2(4), e00074.

Falconer, D. (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Ann. Hum. Genet* 29, 51–71.

Feingold, K. and Grunfeld, C. (2018). Introduction to Lipids and Lipoproteins. *MDText.com*.

Fisher, R. (1928). The general sampling distribution of the multiple correlation coefficient. *Proceedings of the Royal Society of London* 121, 654–673.

Franks, P. (2011). Gene x environment interactions in type 2 diabetes. *Curr Diab Rep* 11(61), 552.

Gestenes, M. and Stiefel, E. (1952). Methods of Conjugate Gradients for Solving Linear Systems. *Journal of Research of the National Bureau of Standards* 49(6), 409.

Hancock, A. e. a. (2011). Population genetic analysis of the uncoupling proteins supports a role for UCP3 in human cold resistance. *Mol Bio Evol* 28(14), 601.

Heid, I., Jackson, A., and Randall, J. e. a. (2010). Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nature Genetics* 42(11), 949–960.

Kangcheng, H., Kathryn, S., Arunabha, M., Huwenbo, S., Nicholas, M., Yue, W., Sriram, S., and Bogdan, P. (2019). Accurate estimation of SNP-heritability from biobank-scale data irrespective of genetic architecture. *Nature Genetics* 51(8), 1244–1251.

Keys, A. and Parlin, R. (1966). Serum Cholesterol Response to Changes in Dietary Lipids. *Am J Clin Nutr* 19, 175–81.

Krafka, J. (1920). The effect of temperature upon facet number in the bar-eyed mutant of Drosophila: part I. *J Gen Physiology* 2(32), 409.

Lee, S., Wray, N., Goddard, M., and Visscher, P. (1965). Estimating missing heritability for disease from genome-wide association studies. *American journal of human genetics* 88(3), 294–305.

Mørkedal, B., Romundstad, P., and Vatten, L. (2011). Informativeness of indices of blood pressure, obesity and serum lipids in relation to ischaemic heart disease mortality: the HUNT-II study. *European Journal of Epidemiology* 26(6), 457–461.

Olkin, I. and Finn, J. (1995). Asymptotic distribution of functions of a correlation matrix. *Psychological Bulletin* 118(1), 155–164.

Olkin, I. and Siotani, M. (1976). Asymptotic distribution of functions of a correlation matrix. *Essays in probability and statistics* 1(1), 235–251.

Pare, G. and Franks, P. (2016). Putting the Genome in Context: Gene-Environment Interactions in Type 2 Diabetes. *Curr Diab Rep* 16, 57.

Poobalan, A., Aucott, L., and Smith, W. e. a. (2004). Effects of weight loss in overweight/obese individuals and long-term lipid outcomes – a systematic review. *Obes Rev* 5, 43–50.

Poppitt, S., GF, K., and Prentice, A. e. a. (2002). Long-term effects of ad libitum low-fat, high-carbohydrate diets on body weight and serum lipids in overweight subjects with metabolic syndrome. *Am J Clin Nutr* 75, 11–20.

Pradhan, A. e. a. (2001). C-reactive protein, interleukin 6, and risk of developing type 2 diabetes mellitus. *JAMA* 286(3), 327–34.

Price, A., Patterson, N., Plenge, R., Weinblatt, M., Shadick, N., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* 38(8), 904.

Richard, Y. (2012). Residualization is not the answer: Rethinking how to address multicollinearity. *Social Science Research* 41(6), 1379–1386.

Robinson, M., English, G., and Moser, G. e. a. (2017). Genotype-covariate interaction effects and the heritability of adult body mass index. *Nature Genetics* 49(8), 1174–1181.

Shewchuk, J. e. a. (1994). An Introduction to the Conjugate Gradient Method without the Agonizing Pain. *Carnegie-Mellon University Internal Publication.*

Simonson, M., Wills, A., Keller, M., and McQueen, M. (2011). Recent methods for polygenic analysis of genome-wide data implicate an important effect of common variants on cardiovascular disease risk. *BMC Medical Genetics* 12(1), 146–155.

Singh, B. and Rabinovitch, A. (1993). Influence of microbial agents on the de- velopment and prevention of autoimmune diabetes. *Autoimmunity* 15(13), 209.

Southwood, K. (1978). Substantive Theory and Statistical Interaction: Five Models. *The American Journal of Sociology* 83(5), 1154–1203.

Stewart, S., Cutler, D., and Allison B. Rosen, A. (2009). Forecasting the Effects of Obesity and Smoking on U.S. Life Expectancy. *NEJM* 361, 2252–2260.

Sujit, R. e. a. (2017). Glycated haemoglobin A1c (HbA1c) for detection of diabetes mellitus and impaired fasting glucose in Malawi: a diagnostic accuracy study. *BMJ* 8(5).

Sulc, J., Mounier, N., and Günther, F. e. a. (2020). Quantification of the overall contribution of gene-environment interaction for obesity-related traits. *Nature Communications* 11, 1385.

Walker, C. and Jebb, S. (2012). Gene–Diet Interactions on Lipid Levels: Current Knowledge in the Era of Genome-Wide Association Studies. *Curr Nutr Rep* 1, 123–131.

Ware, E., Schmitz, L., Faul, J., Gard, A., Mitchell, C., Smith, J., Zhao, W., Weir, D., and Kardia, S. (2017). Common SNPs explain a large proportion of the heritability for human height. *bioRxiv*, 106062.

Wilson, P., D'Agostino, R., and Levy, D. e. a. (1998). Prediction of Coronary Heart Disease Using Risk Factor Categories. *Circulation* 97, 1837–47.

Wishart, J. (1931). The mean and second moment coefficient of the multiple correlation coefficient in samples from a normal population. *Biometrika* 22, 353–361.

Wright, S. (1932). *The roles of mutation, inbreeding, crossbreeding, and selection in evolution.* Vol. 1. Proceedings of the Sixth International Congress of Genetics, 356–366.

Yang, J., Benyamin, B., McEvoy, B., Gordon, S., Henders, A., Nyholt, D., Madden, P., Heath, A., Martin, N., and Montgomery, G. e. a. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* 42(7), 565.