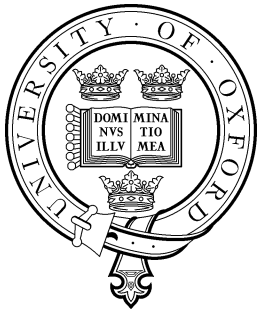
Interactions in Complex Traits



Alexander Young

Lincoln College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2016

Acknowledgements

I thank my supervisor, Peter Donnelly, for the opportunities and support he has given me, and Kate Distin-Harvey, his P.A., for her help and support. I thank Richard Durbin for discussions and comments on the material in Chapter 2. I thank Fabian Wauthier for discussions on the material in Chapters 3, 4, and 5. I thank Jonathan Marchini for comments on the presentation of scientific material. I owe a debt of gratitude to the Wellcome Trust for supporting my studies financially. Many thanks to my college, Lincoln, for the friends that I have made there and the unforgettable events I have shared with them.

Abstract

The availability of cheap genotyping technologies has enabled to collection of very large samples with both genetic and phenotypic information, enabling the interrogation of the genetic architecture of complex traits in humans and other organisms. The role of interactions between genetic variants and between genetic variants and environmental factors in complex traits is not well characterised, especially in humans. This is in part due to a lack of theory and methods designed for powerful investigation of interactions in complex traits in large-scale datasets. This thesis develops both theory and methods relating to interactions between genetic variants and between genetic and environmental factors, complemented by empirical analyses aimed at discovering the influence of interactions on complex traits. The effect of genetic variation on trait variation can be decomposed into components reflecting interactions involving different numbers of genetic variants. The first part of this thesis generalises classical theory on the decomposition of the genetic variance into components arising from different types of interaction to finite populations, where the influence of interactions is more easily detected. The theory is applied to determine the proportion of growth variance from pairwise and third and higher order interactions in a yeast cross. The subsequent parts of the thesis are more directly concerned with interactions between genetic variants and environmental factors. It is first demonstrated that multiple lifestyle factors modify the effect of variants in the *FTO* gene on body mass index (BMI). This motivates the development of the heteroskedastic linear mixed model (HLMM), which exploits changes in variability with genotype to aid discovery of genetic variants involved in interactions. An efficient algorithm for application of the HLMM to large scale datasets is developed and applied to discover genetic variants likely to be involved in interactions on BMI.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

The scientific study of inheritance took its first steps towards its modern statistical framework when Francis Galton started to formulate 'laws' of inheritance in the 1880s, culminating in the publication of *Natural Inheritance* in 1889[1]. His laws of inheritance derived from the observation that the variability of characters such as height in populations is approximately constant from generation to generation[2]. It must therefore be the case that the offspring of those that deviate from the mean do not tend to deviate more from the mean than their parents; in fact, they tend to 'regress' to the population mean. While Galton's intuitions about inheritance were groundbreaking, his mathematical formulations of the laws of inheritance were unsatisfactory and lacked a connection to a true model of biological inheritance.

The ideas of Mendelian inheritance were rediscovered in the early 20^{th} century. These laws of inheritance were not immediately synthesised with Galton's ideas about inheritance because the leading proponent of Galton's laws of inheritance, Karl Pearson, was a convinced Darwinist who believed the discrete nature of Mendel's laws to be incompatible with his 'biometrical' interpretation of evolution by gradual, contin-

uous change[2]. It was not until Fisher's seminal 1918 paper *The correlation between relatives on the supposition of Mendelian inheritance*[3] that Galton's intuitions about inheritance and the correlations between relatives could be understood to arise from a process of Mendelian inheritance.

Fisher derived the correlations between relatives due to the inheritance of Mendelian factors. Included in this analysis was the concept of 'Epistacy': which is the statistical phenomenon generated by interaction between Mendelian factors affecting a trait. The work of Fisher was independently generalised by Kempthorne and Cockerham in the 1950s to include more complex departures of the effects of multiple loci from the sum of their individual effects[4, 5, 6]. The first part of this thesis further generalises the results of Kempthorne and Cockerham to populations descending from a finite number of ancestors, where the influence of interactions between Mendelian factors is more readily apparent.

While Fisher's work was concerned with correlations between relatives induced by inheritance of Mendelian factors, correlations between relatives can also be generated by environmental effects. Furthermore, the effects of Mendelian factors and environmental factors may not be easily separable if the effects of one depend upon the other, i.e. they interact. While the first part of the thesis is concerned with the statistical implications of interactions between Mendelian factors, the subsequent parts are also concerned with interactions between Mendelian and environmental factors.

To aid comprehension of the subsequent chapters, we first discuss what is meant by a statistical interaction. We then discuss the components of a general quantitative trait model that includes gene-by-environment interactions. This gives the general framework within which all the analyses in this thesis can be understood.

1.1.1 Interactions

The common-sense notion of an interaction between two things implies that the two exert some causal influence on each other, resulting in an outcome that could only be predicted by considering both things together. While the statistical notion of an interaction has a connection to this common-sense notion, it is somewhat more subtle. Since statistics is interested in things that vary, it is also only interested in interactions that result in variation of outcome. Two proteins may physically interact in the body, but variations in these proteins may not result in any difference of outcome, leading to no statistically meaningful interaction.

The goal of interaction modelling in statistics is to discover which variables associated with an outcome need to be combined, and how they should be combined, in order to make meaningful inferences about the outcome or accurate predictions. This is the case when the association between one variable and the outcome depends upon the state of some other variables, which can result from a connection between the variables in the underlying causal structure of the outcome.

We follow D.R. Cox in our definition of no interaction[7, 8]: if Y is the response, then there is no interaction between X_1 and X_2 on Y if

$$\mathbb{E}[Y|X_1, X_2] = \eta_1(X_1) + \eta_2(X_2), \tag{1.1}$$

for some functions $\eta_1(X_1)$ and $\eta_2(X_2)$. If this is the case, then Y is said to follow a generalised additive model for its statistical relation to X_1 and X_2.

Deviations from generalised additivity do not necessarily imply a meaningful interaction and could simply reflect scale phenomena. This is another subtlety of the statistical notion of interaction. If the outcome, Y, is the area of a box, then clearly if X_1 is the length of the box and X_2 is the width of the box, X_1 and X_2 cannot be combined purely additively to perfectly model the area of the box. However, $\log(X_1)$

and $\log(X_2)$ could be combined additively to model the logarithm of the area of the box. In this case, a scale transformation results in an additive model, and this is the case for a whole class of interactions, which we discuss later. Whether an additive model on a transformed scale should be preferred to an interaction model on another, potentially more meaningful scale, is a more difficult question that does not have a universal answer.

It may be necessary to combine more than two variables to model the expectation of a response, leading to 'higher order interaction' terms involving three or more variables. The definition of no pairwise interaction can be extended to no k^{th} order interaction: there is no interaction of order k between variables $X_1, \ldots, X_k$ if

$$\mathbb{E}[Y|X_1, \ldots, X_k] = \sum_{S \subseteq \{X_1, \ldots, X_k\};\ |S|=k-1} \eta_S(S), \tag{1.2}$$

i.e. if the expectation can be written as the sum of functions involving only $k-1$ variables.

Looking at only one variable, a necessary condition for that variable to be involved in an interaction is failure of simple additivity of the response given changes in one of the interacting variables[7, 8]. Formally, this means that, if $F(y|X_1 = x) = \mathbb{P}(Y \leq y|X_1 = x)$ and there is no interaction involving X_1, then $\forall\, x, a\ \exists\, c$:

$$F(y|X_1 = x + a) = F(y - c|X_1 = x). \tag{1.3}$$

In other words, if there is no interaction involving X_1, the distribution function of the response given a translation of X_1 is a translation of the original distribution function. Note that deviations from simple additivity of response do not imply a meaningful interaction and may simply reflect distributional or scale properties — non-linear transformations can remove or ameliorate these effects in many cases. For approximately normal distributions, which are characterised by means and variances,

this implies that there will be a dependence between the variance of the response and the state of a variable involved in an interaction.

There are many possible ways for the expectation function to deviate from generalised additivity. For many forms of non-additivity, an interaction effect can be removed by non-linear transformation. This can be useful but is not always desirable: an interaction effect on a meaningful scale should be preferred to an additive effect on a non-meaningful scale[7, 8]. It is most useful when there is a physical argument for why variables may be related in a multiplicative way even when their effects can be considered as independent, such as variables determining the dimensions of a cuboid, or variables that each proportionally change an outcome independently of one another.

There are interactions that cannot be removed by non-linear transformation: interactions where the rank order of the response with respect to X_1 changes with X_2, such as when the direction of association with X_1 is reversed by a change in X_2. These interactions are called ordinal interactions[8]. For non-ordinal interactions, there is usually no 'true' scale on which an interaction can be objectively declared. Throughout this work, we adopt scales of measurement that are meaningful and interpretable in terms of the problem at hand but that remove, as much as possible, the influence of uninteresting, physical non-linearity.

1.1.2 General trait model

In this thesis, we consider only quantitative traits, although some of the methods can be extended to discrete traits. The trait (or equivalently phenotype), Y, is modelled as the sum of genetic, G, and environmental, E, components, along with their interaction, δ_{GE}:

$$Y = G + E + \delta_{\mathrm{GE}}. \tag{1.4}$$

Assuming that $\text{Cov}(\delta_{\text{GE}}, G) = \text{Cov}(\delta_{\text{GE}}, E) = 0$, this gives the variance of Y as

$$\text{Var}(Y) = \text{Var}(G) + \text{Var}(E) + 2\text{Cov}(G, E) + \text{Var}(\delta_{\text{GE}}). \tag{1.5}$$

Often without justification, gene-environment interaction and gene-environment covariance are ignored. This gives a simplified variance decomposition:

$$\text{Var}(Y) = \text{Var}(G) + \text{Var}(E). \tag{1.6}$$

The broad-sense heritability, H^2, measures the proportion of phenotypic variance that is due to variation in genomes[9]. In the case of no gene-by-environment interaction, it is:

$$H^2 = \frac{\text{Var}(G)}{\text{Var}(Y)}. \tag{1.7}$$

When gene-by-environment interactions are present, it is not clear whether $\text{Var}(\delta_{\text{GE}})$ should be included in the broad sense heritability.

1.1.3 Additive genetic effects

We define a genetic variant to be any varying part of the population of genomes under consideration. A genome is comprised of all of the genetic material of an organism, typically formed of DNA (deoxyribonucleic acid). In all of the applications considered here, we consider variations at the single-nucleotide level called single nucleotide polymorphisms (SNPs).

The SNPs considered here generally have two variations (alleles) of non-negligible frequency in the population, and hence are termed bi-allelic. This enables a binary encoding of the SNP as having 0 or 1 copies of either allele on a chromosome, and therefore an encoding of 0, 1, or 2 copies of either allele in an organism possessing two copies of the chromosome, such as is the case for all autosomal chromosomes in

diploid organisms such as humans. The frequency of an allele is the proportion of chromosomes that carry that allele in the population of genomes. We build up the constituents of quantitative genetic models in terms of bi-allelic SNPs encoded as above.

The additive effect of an allele substitution is the average linear change in the trait with the number of copies of the allele. If we denote the encoding of the number of copies of an allele at a SNP as X_1, then, given that X_1 is uncorrelated with other causal SNPs, the additive effect is the coefficient of X_1 in the linear model that minimises the sum of squared deviations, $(Y - \mu_1 - X_1\beta_1)^2$, in the population. If we imagine a set of L genetic variants $X_1, \ldots, X_L$, then the additive effects of the genetic variants are the coefficients from the joint linear regression of Y on $X = [X_1, \ldots, X_L]$: they are the β that minimises the sum of squared deviations, $(Y - \mu - X\beta)^2$, in the population. Then the narrow-sense heritability, which measures the proportion of phenotypic variance that is due to additive effects, is

$$h^2 = \frac{\mathrm{Var}(X\beta)}{\mathrm{Var}(Y)}. \tag{1.8}$$

We note that this is still well defined when gene-by-environment interactions are present, simply reflecting the amount of variance that could be explained by a linear model of allele substitution effects.

The first part of this thesis is concerned with the component of H^2 that is not captured by additive effects: $H^2 - h^2$, which includes the statistical effect of interactions between genetic variants.

1.1.4 Interactions in genetics

1.1.4.1 Dominance

Dominance can be thought of as an interaction between alleles of a single genetic variant that leads to a departure from additivity of allelic substitution effects (Figure 1.1: C and D). In Fisher's 1918 paper[3], he defines dominance as a deviation from additivity in the relationship between the phenotypic mean, $\mathbb{E}[Y]$, and the number of copies of a Mendelian factor, X_1. The dominance deviation is the residual component of the effect of allelic substitutions at one locus that remains after fitting a linear model by least squares. When the relationship is non-monotonic in the number of alleles at a locus (Figure 1.1C), the additive effect can be small relative to the dominance effect; otherwise, the additive effect will generally explain more variance than the dominance effect.

1.1.4.2 Interactions between genetic variants

Interactions between genetic variants are generally considered examples of 'epistasis'. The term 'epistasis' was first used by Bateson in 1909 to describe the phenomenon whereby the effect on a phenotype of an allele at one locus is masked by an allele at another locus[10, 11]. This was seen as an extension of the concept of 'dominance' at a single locus, introduced by Gregor Mendel, whereby the effect of an allele at a single locus is masked by the presence of an alternative allele at the same locus. The classical definition of epistasis is not a statistical definition, although classical epistasis will generate statistical interactions between loci on a certain scale given appropriate population genetic parameters.

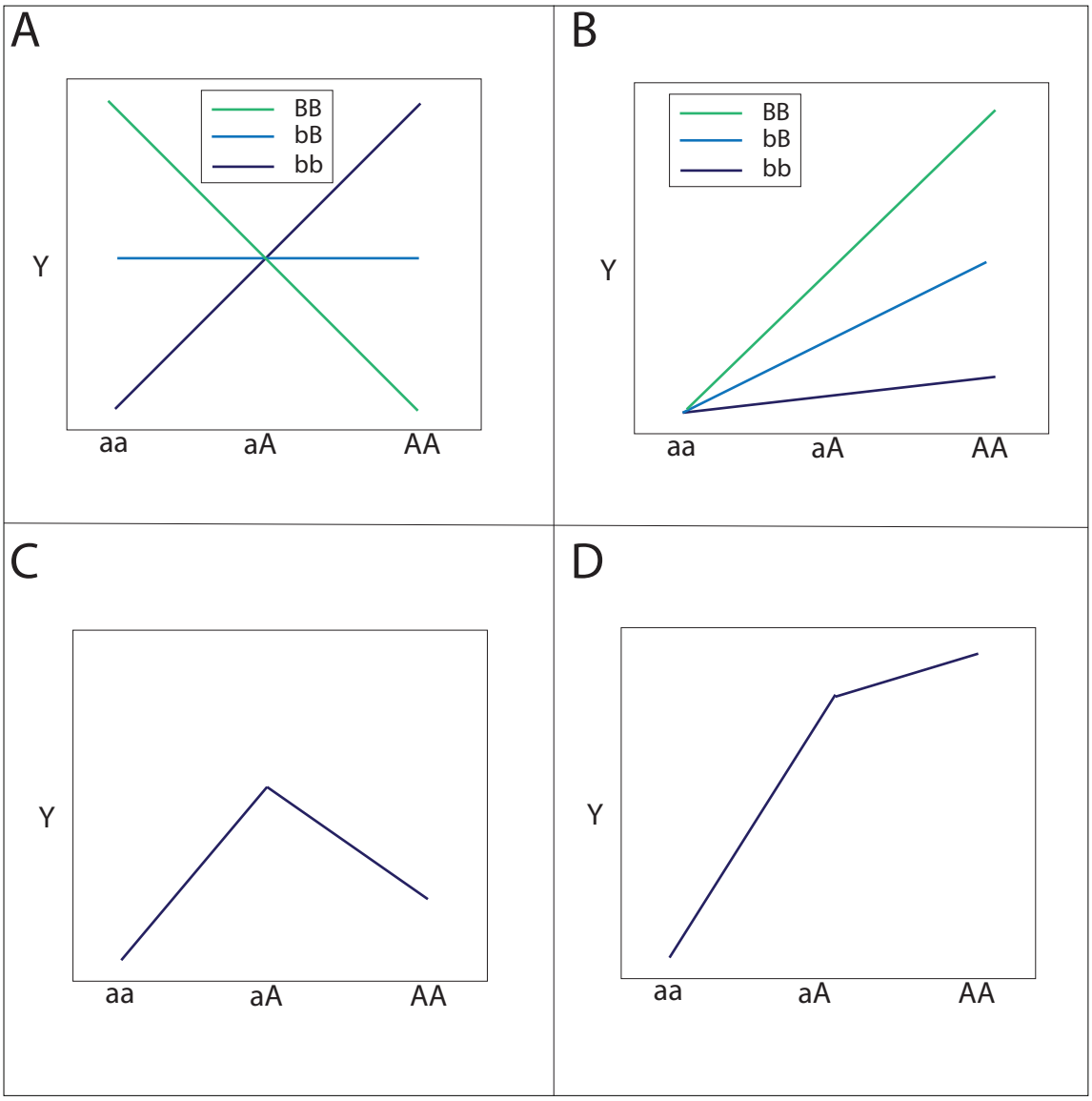


Figure 1.1: **Illustration of epistasis and dominance.** A) Ordinal epistasis: the effect of additional copies of the A allele at one locus changes sign depending on the state of the other locus (bb, bB, BB). B) Non-ordinal epistasis: the effect of additional copies of the A allele at one locus is enhanced by additional copies of the B allele at another locus. C) Dominance leading to a non-monotonic relationship between the number of copies of the A allele at a locus and the phenotype: the heterozygote, aA, is larger to both homozygotes. D) Dominance leading to monotonic non-linearity in the relationship between the number of copies of the A allele at a locus and the phenotype.

Statistical epistasis between two loci is what remains after removing the best fitting model involving only sums of the effects of the two loci. Physical interaction between genetic variants does not correspond simply to statistical epistasis, which

depends on the scale of measurement of the trait and population genetic parameters. For example, if $\mathbb{E}[Y|X_1 = 2, X_2 = 2] = 1$ but $\mathbb{E}[Y|X_1, X_2] = 0$ otherwise, this would imply a strong physical dependence between the effects of the two loci. However, if $\mathbb{P}(X_2 = 2) \approx 1$ but $\mathbb{P}(X_1 = 2) = 0.25$, then the interaction is statistically irrelevant and all that matters statistically is whether $X_1 = 2$.

Ordinal epistasis, for example when the sign of effect of allele substitutions at one locus changes depending on the state of another locus (Figure 1.1A), can result in little or no additive effect when there is symmetry between the two opposing effects. Additive effects tend to explain most of the variance due to a pair of loci exhibiting non-ordinal epistasis (Figure 1.1B).

Epistasis can be extended to multiple loci, which then allows for the possibility of higher-order interactions. There is no interaction of order k between $L \geq k$ loci if the effect of variation at the loci can be modelled involving interactions between $k - 1$ or fewer loci. Analogous to successive approximations to a non-linear function with polynomials of higher and higher degree, the maximal amount of variance is explained using interactions of order $k - 1$ or less, leaving at most the residual variance for interactions of order k or higher.

There has been a controversy and confusion about the role of epistasis in human and non-human genetics since the invention of the concept, part of which derives from the terminology being used differently by different subsets of researchers and part of which derives from the difficulty of establishing epistatic interactions between alleles. How much statistical epistasis there is affects many questions, including: the correlations between relatives for heritable traits[5], the evolutionary dynamics of a trait[12, 13], and the complexity of models necessary for optimal prediction of traits from genetic information. Although statistical epistasis does not correspond simply to physical epistasis, finding statistical epistasis between genetic variants is one way of discovering physical epistasis, which may lead to discovery of causal mechanisms.

1.1.4.3 Interactions between genetic variants and environment

Interactions between genetic variants and environmental variables give rise to situations where the effect of a genetic variant depends upon the state of some environment variables or, equivalently, the effect of an environmental variable depends upon some genetic variables. R.A. Fisher was the first person to define statistical epistasis, and he was the first person to describe what is now termed a gene-by-environment interaction[14, 15], as he found in an analysis of potato crops that '*the yields of different varieties under different manurial treatments are better fitted by a product formula than by a sum formula*'. Again, we see that a genotype-by-environment interaction is indicated by a deviation from an additive model — this time an additive model of genetic and environmental variables rather than two genetic variables: for a genetic variant X_1 and an environmental variable E, the expectation as a function of X_1 and E cannot be written as

$$\mathbb{E}[Y|X_1, E] = \eta_1(X_1) + \eta_2(E), \tag{1.9}$$

for some functions η_1 and η_2.

The statistical interaction of genetic and environmental variation is important because it results in the breakdown of additivity, and therefore easy separability, of genetic and environmental influences on a trait. While there are abundant examples in model organisms of genetic variants whose effects change depending on experimental environment[16, 17, 18, 19, 20], there are few uncontroversial examples in humans outside pharmacogenetics[21], which has found many examples of genetic variation affecting drug response and toxicity. Many potential gene-by-environment interactions affecting body mass index (BMI) and obesity have been proposed, which we discuss in more detail below and in Chapters 3 and 5.

It is possible that gene-by-environment interactions play an important role in

evolution: exposure to novel environments may change the effects of genetic variants on a trait related to fitness, leading to novel selection pressures on extant genetic variation[16]. This phenomenon may be especially important in modern humans, who have experienced dramatic environmental changes since the dawn of agricultural and industrial economies, in addition to rapidly changing cultural practices that are not relevant to other organisms.

1.1.5 Existing methods for interaction discovery

1.1.5.1 Interactions between genetic variants

This thesis is not primarily concerned with developing new methods for detecting interactions between particular genetic variants, so we give only a brief summary of pre-existing approaches.

The number of possible pairwise interactions grows in proportion to the square of the number of variables. If the variables are genome-wide SNPs from a genotyping array, then this number can be on the order of 10^{12}. This imposes both computational and statistical challenges for any method for detecting interactions between genetic variants, with the problem becoming orders of magnitude worse for higher order interactions.

The optimal method for detecting an interaction between genetic variants affecting a trait will depend upon which interaction model is most appropriate. If variants involved in interactions have reasonably strong additive effects (as is likely to be the case for the kind of epistasis in Figure 1.1B), then it may be prudent to reduce the search space by only considering interactions where at least one of the variants passes a threshold of evidence for an additive effect. However, if the interaction is ordinal (Figure 1.1) and results in little or no additive effect, then methods that filter based on additive effects will not perform well.

Most of the methods that do not make assumptions about additive effects at-

tempt to screen all possible pairwise interactions from a large set of genome-wide SNPs, which can be computationally feasible[22]. Various computational improvements have been proposed, mainly for binary traits, making this a viable approach for large sample sizes and hundreds of thousands of SNPs[23, 24]. The advantage of exhaustive pairwise search methods is that they do not make assumptions about additive effects, so are able to detect interactions that result in little additive variance. However, they test a very large number of interactions, each with a very low prior probability of truly affecting a trait. The correlations between the different pairs of loci are complex, which makes adjustment for multiple testing complicated without incurring an overly conservative multiple-comparison correction penalty. An interesting alternative approach for detecting epistasis in case-control traits is BEAM[25]. It uses a Bayesian approach to partition SNPs into categories: those unlinked to the trait, those only individually linked to the trait, and those linked to the trait in combination with other SNPs. While this type of method can use prior information to search through potentially complicated interaction models, the Markov Chain Monte-Carlo computations required make it computationally too demanding to apply to very large sample sizes with very large numbers of SNPs.

Other methods screen variables based on evidence for additive effects to reduce the interaction model search space to pairs of loci more likely to be involved in interactions than random pairs. Two such recent methods assume that at least one of the interacting SNPs has a detectable additive effect[26, 27]. The method from Li et al. uses a correlation screening procedure[28] to prune the model before forming interaction effects, which are then further pruned by correlation screening[27]. The method of Bien et al. potentially puts less emphasis on strong additive effects by considering additive and interaction effects jointly through a penalised and constrained regression framework[26]. Other methods screen based on a change in variance with genotype, which does not make any assumption about the marginal effects of interacting SNPs.

We discuss these methods further in Chapters 4 and 5.

1.1.5.2 Interactions between genetic variants and environmental variables

The search for gene-by-environment interactions in humans has been limited by availability of suitably sized datasets where both genome-wide genetic information and environmental information is available. In contrast, in model organisms experimental designs with controlled environmental exposures have led to many discoveries of gene-by-environment interactions[16, 17, 18, 19, 20]. Many potential gene-by-environment interactions were identified by candidate gene studies in psychiatry, but this approach has been criticised as underpowered and liable to generate false positives[29]. To gain enough power to robustly discover gene-by-environment interactions, meta-analyses have been used to test for interactions between known, common causal loci and measured lifestyle and environmental variables[30]. Outside of pharmacogenetics, there has been limited success in using randomised trials to test whether different treatments or interventions result in different outcomes depending on genotype[31], partially because the large sample sizes needed to have sufficient power are very expensive to collect.

Rather than simply discovering interactions between genetic and environmental factors, it has been proposed that known environmental factors can be used to increase power to discover causal loci when they are involved in interactions with that environmental factor[32]. While this approach has been successful for studying lung function[33], where smoking is a known strong lifestyle effect, it would be harder to adapt to traits where environmental and lifestyle effects are less understood or well-measured. Very large sample sizes may make it possible to investigate interactions between genome-wide genetic variants and multiple, measured environmental factors in a more agnostic manner[34].

As for finding epistatic interactions, methods utilising evidence for a change in

variance with genotype can be used to filter variants before testing interaction models. We discuss these methods further in Chapters 4 and 5.

1.2 Variance component models

In this section, we give a brief overview of the aspects of variance component models, including linear mixed models, most relevant to this thesis. For this, we rely heavily on *Variance Components* by Searle, Casella, and McCulloch[35].

1.2.1 Origins

Variance component models aim to partition the total variance of some variable of interest into components arising from different sources. For example, we might consider measuring how many hours a set of people slept over a series of nights. There are two apparent sources of variation in these observations: the variation in mean sleep duration between individuals, and the variation within an individual between different nights. A variance component model applied to these data may aim to partition the total variation in the observations into these two components. The variance component model would then give information about the relative importance of the different sources of variation: it could answer, for example, whether within-individual or between-individual variation in sleep duration contributes more to overall sleep variation.

Variance component models stretch back to modelling of astronomical observations in the middle of the 19^{th} Century[35]. However, these models do not seem to have had much direct impact upon subsequent developments in this area. The further development of variance component models was motivated by both genetics and analysis of experimental data. Fisher first used variance component models in his partitioning of genetic variance into additive and non-additive components in his

seminal 1918 paper[3]. Fisher would then go on to lay the foundations of a method for estimating variance components, now called 'analysis of variance' (ANOVA).

The ANOVA method was first developed for analysis of randomised experimental designs. These models introduced what we now term 'random effects'. Random effects have been defined in different ways by different authors. One definition is that they are effects that differ in a random way between different levels of a factor according to some distribution. Another definition is that the levels of the factor are randomly sampled from some population of levels, whose effects are distributed according to some distribution in that population. Random effects can also be interpreted in a Bayesian way. While there are philosophical differences and differences in interpretation between different definitions of random effects, the practical properties of random effects derive from the assumption that the effects of different levels of a factor differ according to some distribution, possibly with unknown parameters. The simplest model for this kind of effect involves only one factor (A): observation j of the response Y for level i of the factor can be modelled as

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \tag{1.10}$$

where α_i is the deviation from the mean caused by level i of the factor, and ϵ_{ij} is individual deviation from that expected due to the factor being at level i. The total sum of the squared deviations from the mean can be written as a linear function of the within group (a level of the factor) sum of squared deviations from the group mean, and the between group sum of squared deviations from the overall mean: let $\bar{y}$ be the overall mean and $\bar{y}_i$ be the sample mean of the observations with the factor at level i, then

$$\sum_i \sum_j (y_{ij} - \bar{y})^2 = \sum_i \sum_j (\bar{y}_i - \bar{y})^2 + \sum_i \sum_j (y_{ij} - \bar{y}_i)^2. \tag{1.11}$$

The simplest ANOVA methods relate the components of the partition of the sample sum-of-squares (1.11) to their expectations. The expectations of the sum of squares can be derived simply under the assumption that α_i and ϵ_{ij} are independent for all observations and that the ϵ_{ij} are independent and identically distributed across observations. In this case, the between-group sum of squares and within group sum of squares are simple linear functions of the variance component due to the systematic effects of the factor and the residual component.

Experiments can be designed so that there are the same number of observations of each level of the factor, known as a balanced 1-way classification random model when there is only one systematic varying factor. This enables ANOVA estimates of the variance components to be written in closed form. However, for most applications in genetics, it is not possible to design balanced experiments, so closed form ANOVA estimators are not available. Furthermore, while ANOVA estimators can be shown to be the unbiased quadratic estimators with minimal variance when the data is balanced[35, 36], this is not the case for unbalanced data. We therefore focus on methods for unbalanced data in the subsequent subsections.

1.2.2 Linear mixed models

Linear mixed models incorporate both 'fixed effects' and 'random effects' to model systematic sources of variation in a set of observations. While there can be many reasons for choosing random or fixed models for effects, in general modelling something as a fixed effect focuses on the location effect of the variable, its shift in the mean, whereas modelling something as a random effect focuses inference on how much variation in the observations is explained by variation in the variable.

Linear mixed models can be used for balanced data, where ANOVA estimators may be preferred. However, in genetics, balanced data is not usually available or appropriate, so we do not consider balanced models and their inference methods.

Following Chapter 4 of *Variance Components*[35], the mixed model is defined by the model matrix for the fixed effects, X, which is $[N \times c]$, where N is the overall sample size, and c is the number of fixed effects in the model, and the model matrix for the random effects, Z, which is $[N \times l]$, where l is the number of random effects in the model. The N-vector observations, Y, is modelled as

$$Y = X\beta + ZU + \epsilon, \tag{1.12}$$

where β is a c-vector of fixed effects, U is an l-vector of random effects, and ϵ is the residual vector.

The residual and random effects vectors are usually taken to have zero mean, implying that $\mathbb{E}[Y] = X\beta$ and $\mathbb{E}[Y|U = u] = X\beta + Zu$.

The residual vector is usually taken to have covariance matrix proportional to the identity matrix, $\mathrm{Cov}(\epsilon) = \sigma_\epsilon^2 I_N$. We relax this assumption in Chapters 4 and 5. Furthermore, the residual vector and the random effects vector are taken to be uncorrelated.

If the random effects vector is partitioned so that $U^T = [U_1^T U_2^T \dots U_k^T]$, where U_i has length q_i, then a common model assumes all random effects are uncorrelated and that variances are the same within each group of the partition, giving:

$$H = \mathrm{Cov}(U) = \begin{bmatrix} \sigma_1^2 I_{q_1} & & & \\ & \sigma_2^2 I_{q_2} & & \\ & & \ddots & \\ & & & \sigma_k^2 I_{q_k} \end{bmatrix}. \tag{1.13}$$

If Z is partitioned to correspond to the partition in U: $Z = [Z_1 Z_2 \dots Z_k]$, then

$$Y = X\beta + \sum_{i=1}^{k} Z_i U_i + \epsilon, \tag{1.14}$$

giving

$$\Sigma = \mathrm{Var}(Y) = ZHZ^T + \sigma^2_\epsilon I = \sum_{i=1}^{k} \sigma^2_i Z_i Z_i^T + \sigma^2_\epsilon I. \tag{1.15}$$

We consider an alternative way to parameterise heteroskedasticity in the random effects in Chapter 4.

It is often necessary to assume a particular distributional form for the random effects and residual vector — for example, for estimation of sampling errors or for maximum likelihood inference. The most common distribution form assumed is the multivariate normal distribution, giving

$$Y \sim \mathcal{N}(X\beta, ZHZ^T + \sigma^2_\epsilon I). \tag{1.16}$$

1.2.3 Animal breeding

Linear mixed models have a long history in genetics, mostly in the animal and plant breeding literature. The models in the animal and plant breeding literature rely on the fact that the genetic covariance between individuals due to the additive effects of genetic variants can be shown to be $V_g R$, where R is the additive relatedness matrix, and V_g is the additive genetic variance. The additive relatedness between two individuals is twice the probability that two alleles, one randomly selected from one individual and the other randomly selected from the other individual, were inherited from the same recent common ancestor (Chapter 2). The additive relatedness matrix is usually estimated from a known pedigree in animal breeding. However, estimation of relatedness from observed genetic information is also possible.

The derivation of this form of the genetic covariance relies upon considering the effects of genetic variants as fixed and causal genotypes as random and unobserved, given certain population genetic assumptions (Chapter 2). If only additive genetic effects are considered and environmental effects are identically distributed and uncor-

related between individuals, then the phenotypic covariance matrix is

$$\text{Cov}(Y) = V_g R + \sigma_\epsilon^2 I. \tag{1.17}$$

This covariance matrix could also be derived from a random effects model with a single random effect for each individual that represents the additive genetic 'value' of that individual, the expected phenotypic value averaged over environments, commonly termed the 'breeding value' of the individual in animal breeding literature[37]. If B is the vector of breeding values, then $\text{Cov}(B) = V_g R$. Fixed effects, such as age and sex, are commonly fitted, and normality assumed, leading to linear mixed models of the form

$$Y \sim \mathcal{N}(X\beta, V_g R + \sigma_\epsilon^2 I). \tag{1.18}$$

Estimates of the random effects representing the breeding values are used to predict the economic value of livestock for breeding. Inference on the variance components, V_g and σ_ϵ^2, can be used to estimate heritability: $h^2 = V_g/(V_g+\sigma_\epsilon^2)$, although the estimates of heritability may be confounded with environmental variance. These models can be extended to include components of variation due to non-additive genetic effects (Chapter 2).

1.2.4 Twin and family studies

Correlations between relatives can be caused by both genetic and environmental sources of variation as well as their interaction. Twin and family studies aim to separate genetic and environment components of variation by examining similarities between different relative classes.

The most common twin studies model, the additive-common-environment (ACE) model, assumes the variance can be partitioned into additive (V_g), common-environment

(c), and unique environmental, (σ^2_ϵ), components:

$$\text{Var}(Y) = V_g + c + \sigma^2_\epsilon. \tag{1.19}$$

The common environment is assumed to be shared between twin pairs, leading to the same covariance between monozygotic and dizygotic twins[38]. For studies including only monozygotic and dizygotic twins, a simple moment based estimator can be used to estimate heritability:

$$\hat{h}^2 = 2(\hat{r}_{\text{MZ}} - \hat{r}_{\text{DZ}}). \tag{1.20}$$

This is analogous to the ANOVA estimators of variance components in random effects models and does not require any distributional assumptions for unbiasedness.

It is generally more difficult to derive closed-form moments based estimators for variance components in more complicated models involving multiple relative classes. In these cases, more complicated variance component models may be used allowing for different correlations between relatives due to environment and non-additive genetic effects. These lead to variance component models that can be estimated by maximum likelihood by assuming a (usually normal) distributional form.

1.2.5 Genome-wide association studies

Linear mixed models have proven useful in cases where thousands of genome-wide SNPs have been observed. The number of SNPs genotyped often exceeds the sample size, preventing standard linear regression models from being fitted. One way to fit a model including effects for all genotyped SNPs is to model the SNP effects as random effects coming from some distribution, usually normal, with unknown variance. If Z is the $[N \times l]$ (usually normalised) genotype matrix, then the model, including fixed

effects also, would be

$$Y = X\beta + ZU + \epsilon,\ lU \sim \mathcal{N}(0, V_Z). \tag{1.21}$$

$$\Rightarrow Y \sim \mathcal{N}(X\beta, V_Z R_Z + \sigma_\epsilon^2 I),\ R_Z = l^{-1} Z Z^T. \tag{1.22}$$

The matrix R_Z is a measure of additive relatedness at the SNPs in Z. The variance component V_Z is the amount of variance in the phenotype explained by an additive model of the SNPs in Z.

1.2.5.1 Heritability

A measure of heritability can be extracted by estimating the ratio: $V_Z/(V_Z + \sigma_\epsilon^2)$[39]. Assuming the model is correct, this can only be a lower-bound on the heritability, as the genotyped SNPs capture only a subset of the full genetic variation in the population. Extensions have been suggested to deal with data with near complete observations of all genetic variants above a very low frequency[40]. However, the most important problem with these models is model misspecification: they assume that there are no environmental correlations between the individuals in the sample. If environmental correlation is itself correlated with genetic relatedness in the sample, then these methods will overestimate heritability.

1.2.5.2 Power

Genome-wide association studies test association models at a large set of genome-wide SNPs. To do this in a mixed model, the association model for a particular locus is encoded in the fixed effects: $X = [X_c\, X_{\text{SNP}}]$, where X_c is a matrix of covariates, and X_{SNP} models the effect of the SNP on the mean of the trait. To prevent a loss of power from modelling the SNP effect as both a random and fixed effect, it is advised that SNPs that are correlated with the test SNP are removed from the random effect

matrix[41].

The advantages of association testing within a linear mixed model depend upon which SNPs other than the test SNP are modelled as having random effects. It is possible to increase power to detect true associations by including SNPs that have true associations with the trait[41]. This is because, by modelling true associations that are not related to the test SNP, one reduces the unexplained variation in the model that can mask a true association signal, effectively improving the signal-to-noise ratio for the association test.

1.2.5.3 Population structure

Population structure occurs when there are systematic differences in genetic ancestry between parts of a population[42]. It can be compared to a random-mating population, where there would be no systematic differences in ancestry between different parts of the population, and hence no population structure. Population structure can arise through geographic effects: people who are physically far from each other mating less frequently than those close to each other, for example. It can also arise through migration, mixing of populations, and social effects. It creates problems for statistical modelling of genetic effects when the structure of the population tracks differences in environment, an example of gene-environment covariance, $\mathrm{Cov}(G, E)$ from (1.5).

By modelling the effects of genome-wide SNPs, linear mixed models effectively condition on ancestry in a precise way, removing much of the spurious association signal that can occur in structured populations. It may, however, be possible to achieve similar results using only a small number of the top eigenvectors of the relatedness matrix as fixed effects[41], as they tend to capture the major geographic structure in the genetic data that is confounded with environmental variation[43].

1.2.5.4 Relatedness

When close family relatedness is present within samples, this implies that some of the phenotypic observations are likely to be highly correlated, due to both genetic and environmental causes. If the correlations are not modelled, then the amount of information in the observations will be overestimated, leading to inflated test statistics for association. The relatedness matrix estimated from genome-wide SNPs captures the additive, genetic component of this correlation between relatives, which is often correlated with the environmental component of the correlation between relatives. Therefore, association testing in a linear mixed model can reduce the inflation of test statistics due to correlation between relatives in the sample. This effect can only be achieved when enough genome-wide SNPs are included in the random effect to estimate relatedness between relatives accurately[41]. Including all genome-wide SNPs can, however, reduce the power relative to including a smaller number of SNPs with predictive ability[41]. There is therefore a trade-off between power and reduction of test-statistic inflation in choosing the SNPs for which random effects are modelled.

1.2.6 Maximum likelihood estimation

Maximum likelihood estimation has become the preferred method of estimation for variance component models in genetics. While it has the disadvantage of requiring an assumed distributional form, usually normal, it has the advantage of being able to handle complex, unbalanced designs and to give estimates of parameter uncertainty. Furthermore, when distributional assumptions are accurate, maximum likelihood estimators of the variance components have known favourable statistical properties[44].

One disadvantage of maximum likelihood estimation of linear mixed models is that the models are generally restricted to assuming that the effects of the SNPs are normally distributed, which is not realistic. Bayesian methods have been developed that allow for more sophisticated modelling of the effect distribution of genome-wide

SNPs, which can increase power to detect true associations and improve prediction[45, 46]. We do not consider such models in this thesis even though the method developed in Chapter 4 could be extended to allow for different effect size distributions.

We do not give a comprehensive exposition of maximum likelihood inference for variance component models here, focusing on the aspects that are relevant to application to contemporary problems in genetics. For a more complete exposition, see Chapter 6 of *Variance Components*[35].

The parameter space is constrained so that all of the variance components are non-negative and the residual error variance is positive. However, the maximum of the unconstrained likelihood may lie outside of the parameter space. This creates a problem for unconstrained optimisation methods for finding the maximum likelihood variance component estimates. One solution to this problem is to set components that are negative at the maximum of the unconstrained likelihood to zero and to re-estimated the model without that variance component. However, this process will introduce bias into the estimator of the variance components.

1.2.6.1 Likelihood

The log-likelihood of the linear mixed model (1.16) given an observed phenotype vector y is

$$l(\beta, \sigma_1^2, \ldots, \sigma_k^2 | y, X, Z) = -\frac{1}{2} N \log(2\pi) - \frac{1}{2}|\Sigma| - \frac{1}{2}(y - X\beta)^T \Sigma^{-1}(y - X\beta), \quad (1.23)$$

where $\Sigma = \sum_{i=1}^{k} \sigma_i^2 Z_i Z_i^T + \sigma_\epsilon^2 I$ is the covariance matrix.

1.2.6.2 Derivatives

The derivative with respect to the fixed effects, β, is

$$\frac{\partial l}{\partial \beta} = X^T \Sigma^{-1} y - X^T \Sigma^{-1} X \beta, \tag{1.24}$$

which gives a linear system that β must satisfy for the maximum likelihood to be attained.

The derivative with respect to a variance component σ_i^2 is

$$\frac{\partial l}{\partial \sigma_i^2} = -\frac{1}{2}\mathrm{tr}(\Sigma^{-1} Z_i Z_i^T) + \frac{1}{2}(y - X\beta)^T \Sigma^{-1} Z_i Z_i^T \Sigma^{-1} (y - X\beta). \tag{1.25}$$

1.2.6.3 Restricted maximum likelihood

Maximum likelihood estimation methods do not give unbiased estimates of the variance components in mixed models because they do not take into account the loss of degrees-of-freedom from fitting the fixed effects. Restricted maximum likelihood (REML) gives unbiased estimates of the variance component by maximising the 'restricted likelihood', the likelihood of the model that remains after marginalising out the fixed effects part. When the sample size is very large compared to the number of fixed effects, the difference between REML and maximum likelihood estimators will be small.

The REML equations are derived by transforming the model into the space orthogonal to the fixed effects. This is achieved by multiplying by a matrix K such that $KX = 0$. In this case, $\mathbb{E}[KY] = KX\beta = 0$, and therefore

$$KY \sim \mathcal{N}(0, K\Sigma K^T). \tag{1.26}$$

The restricted likelihood equations can be derived from this and are invariant to the

choice of K[35].

1.2.6.4 Computation

Closed form maximum likelihood estimators cannot be obtained except in certain balanced designs[35]. In practical applications in genetics, iterative procedures are used to fit variance component models. Most algorithms for general inference of linear-mixed models are either Newton or quasi-Newton based, requiring analytical expressions for the Hessian of the log-likelihood. A popular method for estimating variance components by restricted maximum likelihood is the Average Information algorithm, which uses the average of the observed and expected second derivative of the log-likelihood instead of the observed second derivative to speed up computation[47].

To compute the likelihood in its standard form (1.23) requires inversion of Σ, which is cubic in the sample size, and can become prohibitively expensive for very large sample sizes such as in the UK Biobank, especially when testing is performed at thousands of genome-wide SNPs. The storage of an $N \times N$ covariance matrix can also become problematic for very large sample sizes.

Assuming there is only one variance component, so $\Sigma = \sigma^2 ZZ^T + \sigma_\epsilon^2 I$ the model can be diagonalised by first computing the eigendecomposition of $ZZ^T = UDU^T$, where $UU^T = U^T U = I$ and D is diagonal. Multiplying by U^T gives

$$U^T Y \sim \mathcal{N}(U^T X\beta, D + \sigma^2 I). \tag{1.27}$$

This allows the likelihood and its derivatives to be computed in $\mathrm{O}(N)$ operations after the eigendecomposition and transformation have been performed, which is cubic in N. This was the approach taken by EMMA[48] to make it feasible to apply linear-mixed models to perform association tests at thousands of genome-wide SNPs.

If a smaller number of SNPs than N is used in the random effect, then the singular-

value decomposition (SVD) of Z provides a faster way to diagonalise the model than the eigendecomposition of ZZ^T. The SVD of Z is UDV^T, which can be computed in $O(Nl^2)$ operations, which is linear in N for a fixed number of random effects. This approach was used by Fast-LMM[49], and a method for selecting a small set of SNPs based on their predictive ability was developed[50]. Selecting a small set of predictive SNPs can have advantages for dealing with certain types of confounding[51] and can increase power to detect true associations[41].

To deal most effectively with inflation of test statistics due to relatedness in a sample, random effects should be modelled for all observed genome-wide SNPs. To do this for very large sample sizes using an algorithm that scales cubicly in time is not practical. Methods have been developed that use Monte Carlo samples to solve the mixed model equations while avoiding expensive matrix operations[52]. These methods have been further developed into Bayesian mixed-model association testing software that has approximate O($N^{1.5}l$) time-scaling, making it practical to apply to very large samples with large numbers of observed SNPs[46].

1.3 Missing heritability

The invention of cheap genotyping arrays allowed for the collection of samples of thousands of individuals with both genome-wide genotype data and phenotype data, enabling an unprecedented rate of discovery of genetic variants associated with human traits and diseases[53]. The amount of variance explained by the additive effects of discovered loci is currently only a small fraction of the estimated heritability for most traits[53], leading to the so-called 'problem of missing heritability'[54]. There are multiple, overlapping explanations for the gap between estimated heritability and variance explained by discovered variants, including: overestimation of heritability, lack of power to detect variants, importance of rare variants, importance of struc-

tural variants and other complex genomic variations, parent-of-origin effects, gene-by-environment interactions, and epistasis.

We do not attempt to give a review of all possible explanations of missing heritability here, restricting our comments to the aspects most pertinent to the current study.

Mixed-model methodology has been used to estimate the variance explained by the additive effects of all the genetic variation tagged by a genotyping array[39], not just the statistically significant variants. This has given evidence that a considerable fraction of the heritability estimated from twin and family studies can be thus 'explained'. However, these methods make unproven assumptions about the relationship between genetic relatedness and environmental similarity that make their direct comparison to twin and family studies difficult.

It is likely that twin and family studies have overestimated heritability, possibly partially as a result of interactions between genetic variants[55] and/or interactions between genetic and environmental factors[56]. It is also likely that epistasis and gene-by-environment interactions make discovery of genetic variants harder when using standard additive modelling. It is therefore possible that both epistasis and gene-by-environment interactions have made discovery more difficult and led to overestimation of heritability, making both phenomena plausible candidates for explaining some of the 'missing heritability'.

1.4 UK Biobank data

The UK Biobank is a large, population-based prospective cohort with extensive and reliable measurement of a wide range of phenotype, lifestyle and environmental factors[57]. The UK Biobank contains data on 503,323 participants aged 40-69 years old and living in the United Kingdom at the time of recruitment. While only the

interim release of the genetic data was available at the time of this study ($\sim 152,000$ individuals), the full data release will include genetic data on all participants.

The phenotype and lifestyle and environment data is a mixture of self-reported data and physical measurement. Samples taken included blood, urine, and saliva; physical measurements were made; and a verbal interview and touchscreen questionnaire was given to record lifestyle, exposure, health and environmental information[57]. The self-reported nature of the lifestyle and environmental data is likely to reduce power to detect gene-by-environment interactions and could possibly introduce bias. However, the joint measurement of genetic information along with many lifestyle and environmental factors on such a large sample gives us an unprecedented opportunity to discover gene-by-environment interactions.

1.4.1 Genetic data

The analyses in Chapters 3 and 5 use the interim release of genetic data on $\sim 152,000$ individuals. We describe the aspects of the data relevant to the present work and the processing performed prior to the analyses in Chapters 3 and 5.

The genetic data in the UK Biobank was collected using two closely related genotyping arrays: the UK Biobank Axiom array and the UK BiLEVE array[58]. Because those genotyped on the UK BiLEVE were recruited for a study of lung function, the two samples genotyped on the two different arrays differ strongly on many traits. After quality control (described in genotyping quality control document[58]), 152,736 samples ($\sim$ 99.9% of total samples) and 806,466 SNPs ($>$ 99% of array content) remained. We excluded 480 individuals from the analysis that had been flagged as problematic by the Biobank quality control.

1.4.1.1 Population structure

The UK Biobank sample exhibits population structure on many scales. While over 90% of the sample is of British and Irish descent, the remaining individuals' ethnic backgrounds span the rest of Europe, Asia and Africa. Even the British Samples exhibit fine-scale population structure that could cause confounding in genetic association studies[59].

To aid quality control and analysis, UK Biobank identified the largest cluster of individuals on the first two genetic principal components of worldwide genetic variation, which should correspond roughly to those with British ancestry[58]. This major cluster could be used for analyses that are sensitive to strong population structure.

Family relatedness can also cause problems for genetic association testing. The correlations between relatives, if not properly accounted for, inflate association test statistics[42], and shared environmental factors between relatives can lead to confounding of genetic and environmental effects. UK Biobank determined genetic relatedness, and labeled pairs as related if their kinship indicated third degree or closer rrelatedness. UK Biobank pruned the British individuals of relative pairs, with one of each relative pair moved out of the British group. While this procedure will have removed the influence of close relatedness in the 'British' subsample, individuals in the 'British' sample will still have close relatives in the complementary sample, decreasing the independence of these two samples. For our analyses, it is useful if the two subsamples are approximately independent. To achieve this, we further modified the UK Biobank groupings by moving individuals who had any genotyped third degree or closer relatives from the British Sample into the complementary sample, which we label the 'Diverse Sample'. That is, while UK Biobank had removed one member of each pair of close relatives from the British sample, we additionally removed the other individual in each pair, so that all pairs of related (third degree or closer) individuals fall in the Diverse Sample in our analysis. The final British Sample had 112,338 indi-

viduals, and the final Diverse Sample had 39,897 individuals. The British and Diverse Samples' positions on the first two principal components of genetic variation in the overall UK Biobank sample are compared in Figure 1.2. The self-declared ethnicities of the Diverse Sample are recorded in Table 1.1.

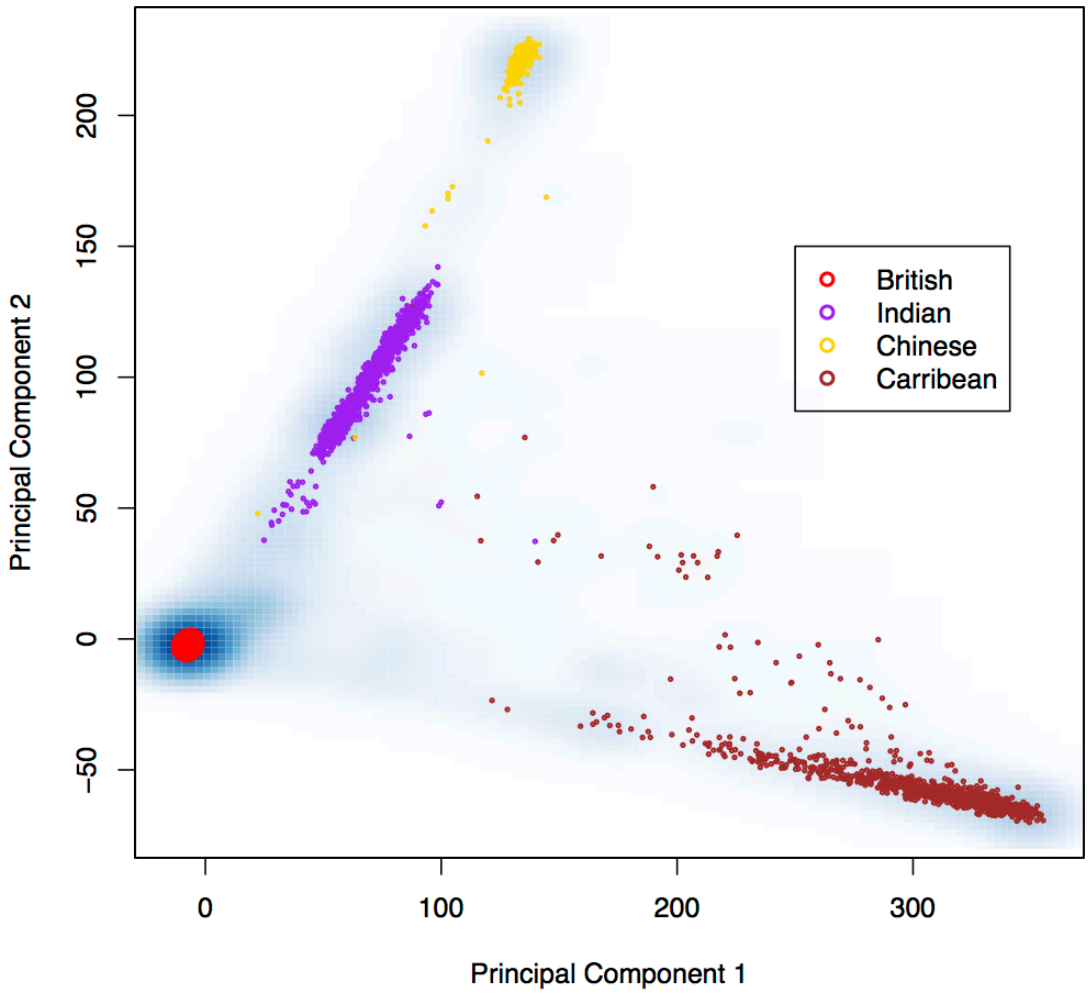


Figure 1.2: **Comparison of sub-samples on the first two principal components of genetic variation**. The British Sample is plotted with red points. The sub-samples of the Diverse Sample with self-declared Indian, Chinese, and Carribean ancestry are highlighted with different coloured points. The smoothed density of the Diverse Sample is shown in blue shading.

H

Ethnicity	%
British	61
Other white background	12
Irish	10
Indian	3.3
Carribean	2.6
Other	2.5
African	1.9
Any other asian background	1.0
Pakistani	1.0
Prefer not to answer	1.0

Table 1.1: **Composition of the Diverse Sample by self-declared ethnicity**. Groups with 1% or greater representation are shown.

1.5 Genetics of body mass index

In this section, we give a concise review of the parts of the literature on the genetics of body mass index most pertinent to the analyses of body mass index in Chapters 3 and 5.

The obesity epidemic is causing a growing burden on public health[60]. Body mass index (BMI), defined as weight divided by height squared, is the most commonly used measure of adiposity, with individuals exceeding a certain BMI threshold classed as obese[61]. BMI, like obesity, is positively correlated with metabolic abnormalities, many common diseases, and all-cause mortality[62]. While the principal causes of the obesity epidemic may be environmental, studies have shown that genetic differences underlie much of the variation in BMI between individuals[63].

Obesity is commonly seen as the simple result of an excess of sloth or gluttony. While willpower undoubtedly plays a role, genetic and environmental factors out of people's control influence the functioning of their metabolism. There is considerable evidence that the body defends a particular fat mass 'set point', which could be under partial genetic control, making it harder for some people to gain or lose weight than

others[64, 65].

Genetic studies have played an important role in identifying key pathways related to energy homeostasis and appetite regulation. Leptin is a hormone produced primarily by white adipose tissue, levels of which are correlated with satiety, and its role was discovered as a result of mapping the gene that caused obesity in a mouse model[66]. Subsequently, 25 human individuals with homozygous mutations in the leptin gene have been identified, with severe obesity with hyperphagia a common feature of leptin deficient individuals[66]. Leptin acts to suppress appetite partly through the leptin-melanocortin signalling pathway in the hypothalamus. Both common and rare variants affecting the melanocortin 4 (MC4R) receptor have been found to increase the risk of obesity[66].

1.5.1 Variants identified by genome-wide association studies

To date, most variants robustly associated with body mass index have been identified through large meta-analyses[67, 68]. In the most recent meta-analysis[68], the 97 loci they found passing genome-wide significance ($p < 5 \times 10^{-8}$) explained around 2.7% of the variation in BMI, a small fraction of the estimated heritability, suggesting a very large number of genetic variants affect body mass index. The common variants with the strongest effect on body mass index are located in the fat mass and obesity related gene (*FTO*), and the potential interactions with the environment of these variants are the subject of Chapter 3.

The associated loci are enriched for nervous system function, and many genes implicated by association results have central nervous system functions, with glutamate signalling particularly strongly implicated. Other genes implicated by genetic associations are involved in insulin secretion and action, energy metabolism, lipid biology, and/or adipogenesis[68]. Overall, genetic association studies support the hypothesis for an important role of the central nervous system, especially the hypothalamus,

in regulating body mass, with contributions from metabolic systems more directly related to energy processing and fat storage.

1.5.2 Gene-by-environment interactions

The fact that body mass index has risen over time in western nations[60] raises the possibility of gene-by-environment interactions over time: that genetic variants predisposing people to gain weight had smaller effects in the recent past. It is therefore possible that the present variations in lifestyle and environment, mirroring differences between the present at the past, also alter the effects of genetic variants on body mass index. While the prior plausibility of gene-by-environment interactions affecting body mass index is high, the small effect sizes of common variants affecting BMI implies that the power to detect gene-by-environment interactions will be low without very large sample sizes. Heterogeneity of measurement of the environment also makes replication and meta-analysis difficult[30]. It is therefore unsurprising that there are no reliably replicated gene-by-environment interactions affecting obesity apart from interactions between *FTO* variants and physical activity (see Chapter 3), although there have been many small studies reporting significant interactions[69, 31]. Overall, we believe the strong prior plausibility of gene-by-environment interactions affecting body mass index make it an ideal test case for methods for discovery of gene-by-environment interactions.

If gene-by-environment interactions are important in the regulation of body mass index, it is likely that twin studies estimates of the heritability of body mass index are inflated[56]. The oft-quoted 40-70% estimate of the heritability of body mass index[63] is quite likely to be an overestimate. If true, this implies that, while the proportion of the narrow-sense heritability explained by known variants is higher than estimated, the nature of that heritability is different and environmentally dependent.

Chapter 2

Variance components in finite populations

2.1 Introduction

Genome-wide association studies (GWAS) have renewed interest in methods for estimating the narrow sense heritability, which is the maximum proportion of the phenotypic variance that the additive effects found by GWAS could explain. The variance explained by the known associations for a trait is typically only a fraction of the estimated narrow sense heritability, with the remaining heritability often labelled 'missing' [53, 54].

Most estimates of narrow sense heritability come from twin and family studies[38], which can be upwardly biased in the presence of genetic interactions[55]. Genetic interactions introduce this bias by introducing convex non-linearity to the relationship between phenotypic correlation and kinship (Figure 2.1). The most common twin studies estimator is the Additive-Common-Environment (ACE) estimator that fits a model for additive genetic, shared (common) environmental, and unshared environmental effects by comparing monozygotic (identical) twin correlations to dizygotic

(non-identical) twin correlations[38]. Both the ACE estimator and a method that exploits the variation in kinship between siblings[70], assume a linear relationship between phenotypic correlation and kinship. Assuming no environmental confounding, the gradient at the mean level of kinship in the population is the true narrow sense heritability[55], with the gradient increasing as kinship increases above the mean due to the influence of genetic interactions. The degree to which this has biased twin and family study estimates depends on the amount of epistatic variance, which is not known for most complex traits.

How much variance there is from interaction effects reflects the genetic architecture of a trait and the statistical complexity of the relationship between genotype and phenotype, and is therefore of interest beyond the debate about 'missing heritability'. If we knew in advance which traits exhibited considerable variance from interactions, it would help focus resources on searching for interactions in those traits.

Current examples of interactions between common variants in humans explain only a small amount of the phenotypic variance[71, 72]. By analogy with the problem of missing heritability for additive effects, it is unlikely that we will have the power to detect all of the interactions influencing a trait, so the only way to assess the statistical importance of interaction effects will be by methods which measure the variance they contribute in aggregate.

The evolutionary model of a phenotype depends on the partition of its genetic variance into additive and non-additive components. It has been argued that natural populations have evolved suppressing epistatic interactions as 'canalizing' mechanisms, which is a possible explanation for why we do not observe many common, large marginal effect alleles [73]. It has been suggested that epistasis explains why there is more genetic variation than expected for traits under selection, a phenomenon called 'stasis' [13] . Some authors have suggested that some of the associations found by GWAS only look additive because of incomplete linkage disequilibrium between

genotyped variants and interacting causal variants[13, 74]. These hypotheses could be tested by accurately estimating the variance from pairwise interactions.

Classical quantitative genetic theory generally assumes an infinite, outbred, random-mating population – see [75] for a list of typical assumptions. In a classic paper, R.A. Fisher showed how pairwise genetic interactions influence the covariance between relatives under these population assumptions[3]. Thirty-six years later, Fisher's theory was generalised to include all orders of genetic interaction by Kempthorne and Cockerham independently of each other[4, 5, 6].

It has proven very difficult to estimate the variance from pairwise interactions in the outbred populations for which the theory was derived. This is because there is almost no contribution from interaction effects to the covariance between pairs of unrelated individuals in outbred populations. Samples of unrelated individuals therefore contain very little information about the contribution of interaction effects to phenotypic variation. Samples of closely related individuals do, but the information is often confounded with shared environmental and dominance effects.

The difficulty of estimating the variance from pairwise interactions can be likened to the difficulty of fitting the quadratic curve shown for the epistatic trait in Figure 2.1. Fitting a quadratic requires information about a wider range of points than fitting a line. Therefore, estimating the variance from interactions requires information about phenotypic correlation over a wider range of kinship than estimating the narrow sense heritability does. Founder populations, which have gone through recent population bottlenecks, are characterised by increased kinship variation and mean kinship compared to outbred populations[76, 77]. A sample from a founder population therefore contains more information about the non-linear change in phenotypic correlation with kinship, thereby bringing the variance from genetic interactions into statistical reach.

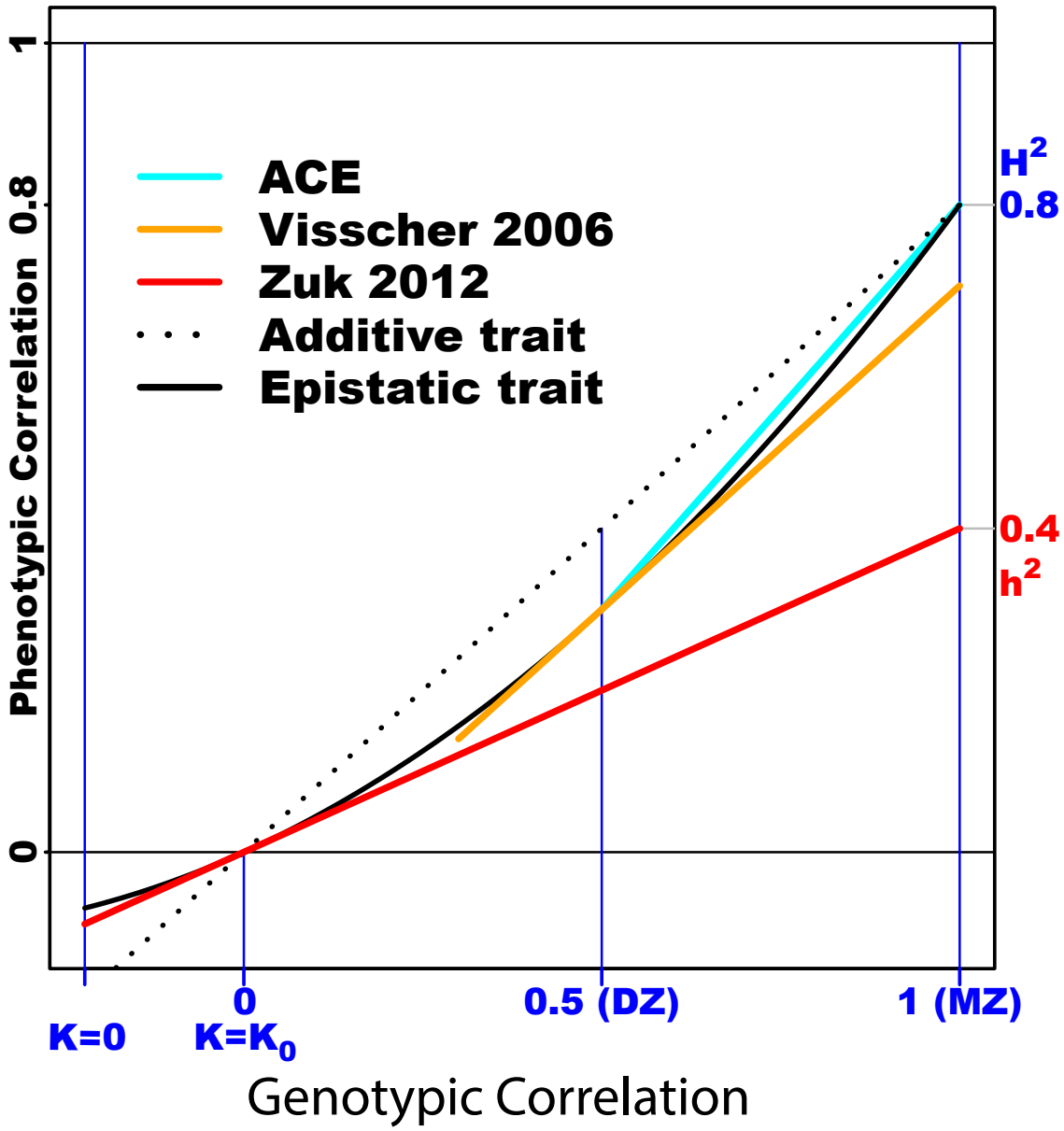


Figure 2.1: **Epistatic trait with heritability estimators**. Phenotypic correlation as a function of genotypic correlation plotted for an epistatic trait (solid black curve) with narrow sense heritability $h^2 = 0.4$ and broad sense heritability $H^2 = 0.8$, and for an additive trait (dotted black line) with narrow-sense heritability $h^2 = 0.8$. The genotypic correlation is a function of the kinship coefficient, K. When $K = K_0$, the mean kinship coefficient in the population, the genotypic correlation is zero. When $K = 0 < K_0$, the genotypic correlation is negative. These points are marked on the x-axis. The genotypic correlations for dizygotic twins (DZ), 0.5, and monozygotic twins (MZ), 1, are indicated on the x-axis. The ACE estimator is twice the difference between the monozygotic and dizygotic phenotypic correlation, the gradient of the blue line, which is 1 here. The gradient of the orange line (0.8) is the rate of change of phenotypic correlation around the mean genotypic correlation for siblings[70]. The gradient of the red line, which is equal to the narrow sense heritability (0.4) is the rate of change of phenotypic correlation around the mean kinship in the population[55].

Theoretical work in finite populations was previously restricted to the contribution of the variance from pairwise interactions to the covariance between half and full siblings[78, 79]. The approach taken was to adjust the covariance between siblings for the background relatedness introduced by a population bottleneck[79] . We instead adjust the kinship for recent finite population size and use this expression to derive the covariance between relatives in a finite population as a function of their kinship, the mean kinship in the population, and the variance from genetic interactions of different order. This gives the estimator proposed by Zuk et al.[55] (see Figure 2.1) as a simple corollary.

The theory we develop applies to certain laboratory crosses as well as to natural founder populations. Laboratory crosses often start with a small number of founding individuals, such as the 'outbred' rat and mice populations[80]. The small number of founders gives a larger range of kinship than occurs in natural human founder populations, giving more power to estimate interaction variance. We exploit this greater power to perform the first estimates of the variance from third and higher order interactions in a yeast cross.

2.2 Theory

The theory is derived for a population recently founded by a finite number of ancestors carrying a total of A haplotypes, with random mating after founding.

2.2.1 Genotypic covariance

We consider an allele at a locus i on haplotypes t and u. We calculate the covariance of the allelic states of the haplotypes by conditioning on whether or not the alleles were inherited from the same founder: whether or not the alleles are identical-by-descent (IBD). The allelic state is coded as a binary variable: g_{ti} for haplotype t and

g_{ui} for haplotype u.

If there were A ancestral haplotypes at the time of the bottleneck, and c_i of these haplotypes carried the allele, then the probability that the two haplotypes carry the allele given that they are IBD at the locus is

$$\mathbb{P}(g_{ui} = g_{ti} = 1 | \text{IBD}^i_{u,t}) = \frac{c_i}{A}, \tag{2.1}$$

where $\text{IBD}^i_{u,t}$ indicates that haplotypes u and t are IBD at locus i.

Conversely, given that the two haplotypes are not IBD at the locus, then the probability they both carry the allele is

$$\mathbb{P}(g_{ui} = g_{ti} = 1 | \neg\text{IBD}^i_{u,t}) = \frac{c_i(c_i - 1)}{A(A-1)}. \tag{2.2}$$

This is because, given the alleles are not IBD, if one haplotype inherits the allele from one of the c_i ancestral haplotypes carrying the allele, the other haplotype can only inherit the allele from one of the $(c_i - 1)$ other ancestral haplotypes carrying the allele – sampling from the ancestral haplotypes without replacement.

We define the kinship coefficient between two individuals to be the probability that a randomly selected allele from one individual is IBD with a randomly selected individual from the other individual. While the probabilities used by other authors are commonly derived from pedigree relations, which gives the expected kinship coefficient between individuals, we use the realised kinship coefficient that is calculated from observations of the genetic material inherited by the pair of individuals. The probability that a pair of haplotypes sampled without replacement is IBD at a locus is the mean kinship coefficient, defined to be K_0, which is A^{-1} for a random mating population. If we define the expected allele frequency to be $f_i = c_i/A$, (2.2) can

therefore be expressed as

$$\mathbb{P}(g_{ui} = g_{ti} = 1 | \neg \text{IBD}^i_{u,t}) = \frac{f_i(f_i - K_0)}{1 - K_0}. \tag{2.3}$$

Note that because $f_i \geq K_0 \geq 0$, $f_i^2 \geq \mathbb{P}(g_{ui} = g_{ti} = 1 | \neg \text{IBD}^i_{u,t}) \geq 0$. Therefore, if $\kappa_{t,u}$ is the probability that haplotypes t and u are IBD at locus i, the probability that both haplotypes carry the allele is

$$\mathbb{P}(g_{ti} = g_{ui} = 1) = \kappa_{t,u} f_i + (1 - \kappa_{t,u}) \frac{f_i(f_i - K_0)}{1 - K_0}. \tag{2.4}$$

Because $\mathbb{E}[g_{ui}] = \mathbb{E}[g_{ti}] = f_i$, the covariance between g_{ti} and g_{ui} is therefore

$$\begin{aligned} \text{Cov}(g_{ti}, g_{ui}) &= \mathbb{E}[g_{ti} g_{ui}] - \mathbb{E}[g_{ui}]\mathbb{E}[g_{ti}] && (2.5) \\ &= \mathbb{P}(g_{ti} = g_{ui} = 1) - f_i^2 && (2.6) \\ &= f_i(1 - f_i) \frac{\kappa_{t,u} - K_0}{1 - K_0}. && (2.7) \end{aligned}$$

K_0 parameterises the adjustment of the genotypic covariance for finite population history. When $K_0 = 0$, as in an infinite, random-mating population, alleles are independent when not shared IBD.

2.2.2 Covariance between relatives

We derive the result first for the simple case of two interacting bi-allelic loci that are inherited independently, and are therefore in linkage equilibrium, with the generalisation following the same logic.

The phenotype of a diploid individual t is modelled as

$$Y_t = \mu + \beta_1 x_{t1} + \beta_2 x_{t2} + \beta_{1,2} x_{t1} x_{t2} + \epsilon_t. \tag{2.8}$$

The deviation in minor allele copy number from the mean at locus i for individual t is x_{ti}:

$$x_{ti} = g_{ti}^m + g_{ti}^p - 2f_i, \tag{2.9}$$

where g_{ti}^m and g_{ti}^p are indicator variables for whether the maternal and paternal haplotypes of individual t carry the minor allele at locus i. The frequency of the minor allele at locus i is f_i, and therefore $\mathbb{E}[x_{ti}] = 0$. The phenotypic mean is μ, and ϵ_t is the residual error, with mean zero and variance σ_ϵ^2, which includes both environmental influences and random noise. For the purposes of this theorem, we assume that there is no gene-environment covariance or gene-environment interaction.

Expressing the genetic contribution to the phenotypic value in this way gives an orthogonal partition of the phenotypic variance. Because of linkage equilibrium, $\mathrm{Cov}(x_{t1}, x_{t2}) = 0$. $\mathrm{Cov}(x_{t1}, x_{t1}x_{t2})$ also equals zero because

$$\begin{aligned} \mathrm{Cov}(x_{t1}, x_{t1}x_{t2}) &= \mathbb{E}[x_{t1}^2 x_{t2}] \\ &= \mathbb{E}[x_{t1}^2]\mathbb{E}[x_{t2}] = 0, \end{aligned} \tag{2.10}$$

where we have again relied on the fact that the loci are in linkage equilibrium. This implies that β_1 is the regression coefficient of the genotype at locus 1 on the phenotype. $\beta_{1,2}x_{t1}x_{t2}$ is the residual effect of the interaction between loci 1 and 2 after accounting for the marginal effects of the loci.

The covariance between the phenotypes of two individuals t and u relies upon the covariance of their genotypes, which is a function of the IBD sharing between their haplotypes:

$$\begin{aligned} \mathrm{Cov}(x_{t1}, x_{u1}) &= \sum_{i=m,p}\sum_{j=m,p} \mathrm{Cov}(g_{t1}^i, g_{u1}^j) \\ &= f_1(1-f_1)\sum_{i=m,p}\sum_{j=m,p} \frac{\kappa_{t,u}^{i,j} - K_0}{1 - K_0} \end{aligned} \tag{2.11}$$

by (2.7), and where $\kappa_{t,u}^{m,p}$ is the proportion of the maternal haplotype of individual t that is IBD with the paternal haplotype of individual u. This can be expressed in terms of the kinship coefficient between t and u, defined to be $K_{t,u}$. This is the probability that an allele drawn at random from each individual is IBD. Therefore,

$$K_{t,u} = \frac{1}{4} \sum_{i=m,p} \sum_{j=m,p} \kappa_{t,u}^{i,j}. \tag{2.12}$$

The covariance can thereby be expressed as

$$\mathrm{Cov}(x_{t1}, x_{u1}) = 2\frac{K_{t,u} - K_0}{1 - K_0}\mathrm{Var}(x_1), \tag{2.13}$$

where K_0 is the mean kinship coefficient in the population.

The covariance of the interaction effects is

$$\mathrm{Cov}(\beta_{1,2}x_{t1}x_{t2}, \beta_{1,2}x_{u1}x_{u2}) = \beta_{1,2}^2\mathbb{E}[x_{t1}x_{t2}x_{u1}x_{u2}]. \tag{2.14}$$

This can be evaluated by further assuming, in addition to linkage equilibrium, that IBD sharing at locus 1 is independent of IBD sharing at locus 2, i.e. that there is no identity disequilibrium[79] between the loci. The covariance of the interaction effects is then (Appendix A)

$$\mathrm{Cov}(\beta_{1,2}x_{t1}x_{t2}, \beta_{1,2}x_{u1}x_{u2}) = \beta_{1,2}^2\mathbb{E}[x_{t1}x_{u1}]\mathbb{E}[x_{t2}x_{u2}]. \tag{2.15}$$

This is equivalent to

$$\mathrm{Cov}(\beta_{1,2}x_{t1}x_{t2}, \beta_{1,2}x_{u1}x_{u2}) = 4\left(\frac{K_{t,u} - K_0}{1 - K_0}\right)^2 \beta_{1,2}^2\mathrm{Var}(x_1)\mathrm{Var}(x_2). \tag{2.16}$$

Therefore, the phenotypic covariance is

$$\mathrm{Cov}(Y_t, Y_u) = 2\left(\frac{K_{t,u} - K_0}{1 - K_0}\right) v_1 + 4\left(\frac{K_{t,u} - K_0}{1 - K_0}\right)^2 v_2 + \mathrm{Cov}(\epsilon_t, \epsilon_u), \tag{2.17}$$

where

$$v_1 = (\beta_1^2 \mathrm{Var}(x_1) + \beta_2^2 \mathrm{Var}(x_2)) \tag{2.18}$$

is the additive variance, and

$$v_2 = \beta_{1,2}^2 \mathrm{Var}(x_1)\mathrm{Var}(x_2) \tag{2.19}$$

is the pairwise interaction variance.

The phenotypic variance of individual t is a function of their inbreeding coefficient, F_t. Setting $t = u$ in (2.17) and using the fact that $K_{t,t} = (1 + F_t)/2$,

$$\begin{aligned} \mathrm{Var}(Y_t) = &\left(1 + \frac{F_t - K_0}{1 - K_0}\right) v_1 + \\ &\left(1 + \frac{F_t - K_0}{1 - K_0}\right)^2 v_2 + \sigma_\epsilon^2, \end{aligned} \tag{2.20}$$

In Appendix A, we extend the two locus model to include dominance effects at the loci. In a founder population, inbreeding induces a correlation between the additive and dominance effects at a locus. The change in the mean due to inbreeding – inbreeding depression – introduces a further variance component. The individual level variance is

$$\begin{aligned} \mathrm{Var}(Y_t) = &\sum_{\tau=1}^{2} \left(1 + \frac{F_t - K_0}{1 - K_0}\right)^\tau v_\tau + (1 - F_t) v_\delta + \\ &4\frac{F_t - K_0}{1 - K_0} C_{a,d} + F_t v_h + \frac{F_t(1 - F_t)}{(1 - K_0)^2} SS_{\mu_h} + \sigma_\epsilon^2; \end{aligned} \tag{2.21}$$

where v_δ is approximately equal to the dominance variance as defined in an outbred population; $C_{a,d}$ is the covariance between additive and dominance effects; v_h

is the dominance variance in a homozygous population; and SS_{μ_h} is the sum of the squared inbreeding depressions at the loci, which is approximately equal to v_δ when K_0 is small. The components, apart from v_δ, are as defined in [76]; however, their coefficients are different.

The variance of the phenotype in the population is found by applying the law of total variance, $\mathrm{Var}(Y) = \mathbb{E}_t[\mathrm{Var}(Y_t)] + \mathrm{Var}_t(\mathbb{E}[Y_t])$, to equation 2.21. Because the mean inbreeding coefficient is equal to the mean kinship coefficient in a random-mating population,

$$\begin{aligned}\mathrm{Var}(Y) = v_1 + \left(1 + \frac{\mathrm{Var}(F)}{(1-K_0)^2}\right) v_2 + (1-K_0)v_\delta + \\ K_0 v_h + \frac{K_0}{(1-K_0)} SS_{\mu_h} + \frac{\mathrm{Var}(F)}{(1-K_0)^2}\left(\mu_h^2 - SS_{\mu_h}\right) + \sigma_\epsilon^2.\end{aligned} \tag{2.22}$$

The narrow sense heritability in the population is $h^2 = v_1/\mathrm{Var}(Y)$. For an outbred population, the variance in inbreeding coefficient, $\mathrm{Var}(F)$, is zero, so the proportion of phenotypic variance explained by the interaction is $v_2/\mathrm{Var}(Y)$. However, for strongly bottlenecked populations, variation in inbreeding coefficient increases the contribution of the interaction to $\mathrm{Var}(Y)$. The dominance variance components arising from inbreeding, v_h, SS_{μ_h} and μ_h^2, do not contribute much to population variation except in the most strongly bottlenecked populations.

We now give the generalisation of (2.17) to arbitrary epistasis between a set of causal loci each with any number of alternative alleles – for the detailed derivation, see Appendix A. The phenotypic covariance, for a set of causal loci N, is

$$\mathrm{Cov}(Y_t, Y_u) = \sum_{\tau=1}^{|N|} 2^\tau \left(\frac{K_{t,u} - K_0}{1 - K_0}\right)^\tau v_\tau + \mathrm{Cov}(\epsilon_t, \epsilon_u), \tag{2.23}$$

where v_τ is the variance from interactions involving τ loci.

If we take the limit of (2.23) as $K_0 \to 0$, we get

$$\mathrm{Cov}(Y_t, Y_u) = \sum_{\tau=1}^{|N|} (2K_{t,u})^{\tau} v_{\tau} + \mathrm{Cov}(\epsilon_t, \epsilon_u), \tag{2.24}$$

which is equivalent to the classic result of Kempthorne without dominance effects[5].

Under more restrictive assumptions, Zuk et al. derived that, for haploids, the gradient of the phenotypic correlation at the mean IBD sharing is the narrow sense heritability[55] – see Figure 2.1 for a visualisation of this. The diploid version of the theorem is a corollary of (2.23) given by

$$v_1 = \frac{(1 - K_0)}{2} \frac{\partial \mathrm{Cov}(Y_t, Y_u)}{\partial K_{t,u}} |_{K_{t,u}=K_0}. \tag{2.25}$$

This shows that (2.23) unifies the estimator proposed by [55] with the classic result of Kempthorne[5] .

The regression method proposed by [55] to estimate v_1 does not take into account the dependencies between pairs of phenotype observations. It is therefore preferable to fit variance components by maximum likelihood or restricted maximum likelihood, as in [76, 81, 82]. The off-diagonal elements of the phenotypic covariance matrix are given by (2.23), with the diagonal elements given by

$$\mathrm{Var}(Y_t) = \sum_{\tau=1}^{|N|} \left(1 + \frac{F_t - K_0}{1 - K_0}\right)^{\tau} v_{\tau} + \sigma_{\epsilon}^2. \tag{2.26}$$

2.2.3 Haploid case

The theory simplifies in the haploid case due to absence of inbreeding or dominance effects. The kinship between two haploids i and j, $K_{i,j}$, is simply the proportion of

the genome shared IBD, and the phenotypic covariance matrix is

$$\mathrm{Cov}(Y) = \sum_{\tau=1}^{|N|} v_\tau K_\tau + \mathrm{Cov}(\epsilon), \tag{2.27}$$

where K_τ is a symmetric matrix with 1s on the diagonal and off diagonal elements

$$[K_\tau]_{ij} = \left(\frac{K_{i,j} - K_0}{1 - K_0} \right)^\tau, \tag{2.28}$$

and $\mathrm{Cov}(\epsilon)$ is the covariance matrix of the environmental effects.

2.3 Methods

2.3.1 Simulations for variance component inference

2.3.1.1 Pairwise interaction variance

To investigate the precision with which the variance from pairwise interactions could be estimated in different populations, we simulated founder populations with different mean kinship by varying the number of founding haplotypes.

The allele frequencies of the variants in the ancestral population were generated by randomly sampling from a distribution with density proportional to 1/f, where f is the allele frequency. We simulated 100 variants in this way.

Each chromosome in the sample was made as a mosaic of independently inherited segments: the length of each segment was drawn from an exponential distribution with a mean of ten, and the genotypes in the segment were copied from a random ancestral haplotype. The expected number of independently inherited segments for each haplotype in the sample was therefore ten. The ancestor from whom each segment was inherited was recorded.

To calculate the diploid kinship coefficient for a pair of individuals, the total num-

ber of variants descending from the same ancestor for each of the four maternal/paternal-maternal/paternal haplotype pairs, one from each individual, was calculated; the sum total sharing across the four haplotype pairs divided by four times the number of variants gives the diploid kinship coefficient between the two individuals. The mean kinship coefficient, K_0, was taken to be the inverse of the number of ancestral haplotypes, which is its expectation. There will be negligible deviation of the sample K_0, calculated over all pairs, from its expectation.

Following the theoretical results, we calculated the component of the covariance matrix due to additive effects, defined to be R_1, by calculating element s, t of R_1 as $2(K_{s,t} - K_0)/(1 - K_0)$, where $K_{s,t}$ is the diploid kinship coefficient of the pair of individuals s and t. The component of the covariance matrix due to pairwise interaction effects, R_2, was calculated as the Hadamard square of R_1.

The kinship coefficients were calculated using all 100 variants, corresponding to calculating kinship from genome wide IBD sharing. However, only a small proportion of the genome is likely to affect a particular trait. To simulate the sparsity of causal variants, the traits were simulated by randomly choosing 10 variants to be causal.

The variants were independently chosen for each simulated trait, covering a range of different frequency distributions of causal variants. Each variant was given an additive effect, and each pair of variants was given an interaction effect. Effects were drawn from normal distributions scaled so that $v_1 = 0.4$ and $v_2 = 0.2$; Gaussian error was added with variance 0.4. The variance components were inferred by fitting the covariance matrix as

$$v_1 R_1 + v_2 R_2 + \sigma_\epsilon^2 I. \tag{2.29}$$

Variance components were estimated by restricted maximum likelihood using the average information algorithm in *GCTA* [83].

We simulated four populations with mean kinship ranging from 1/240 to 1/30, covering a broad range of human founder populations. We simulated 500 phenotypes

with the same variance components for each population.

To investigate how epistasis might bias inference of additive variance, we fitted the covariance matrix as $v_1 R_1 + \sigma_\epsilon^2 I$, ignoring any epistasis, across the four simulated populations for the phenotypes with $v_1 = 0.4$ and $v_2 = 0.2$. To measure the effect of the amount of epistasis on the bias, we simulated further phenotypes varying v_2 from 0.1 to 0.4 for the population with mean kinship equal to 1/240.

2.3.1.2 Third order interaction variance

To investigate the limits of our ability to fit epistatic variance components, for each of the four populations we simulated 200 additional phenotypes with $v_1 = 0.4$, $v_2 = 0.2$, and $v_3 = 0.2$; Gaussian error was added with variance 0.2. The phenotypes were simulated as above except every combination of three causal variants was given a third order interaction effect, scaled so that the total variance from third order interactions was 0.2. The variance components were inferred by fitting the covariance matrix as

$$v_1 R_1 + v_2 R_2 + v_3 R_3 + \sigma_\epsilon^2 I, \tag{2.30}$$

where R_3 is the Hadamard cube of R_1.

2.3.2 Yeast Cross

Bloom et al. presented data from a cross of a lab strain and a wine strain of yeast[84] . They sequenced the founder strains and 1008 genetically distinct haploid descendants (segregants) of the cross of the two strains. This allowed them to infer from which founder each allele had been inherited. We inferred IBD sharing proportions for each pair of haploid segregants by calculating the probability that a randomly chosen variant was inherited from the same founder. The phenotype data is final colony size for each segregant on 46 different growth media. Bloom et al. estimated broad sense

heritabilities, H^2, by analysing biological replicates[84].

2.3.2.1 Inference of heritability components

We fitted the following model to each phenotype Y,

$$Y \sim N(\mu, v_1 K_1 + v_2 K_2 + \sigma^2 I), \tag{2.31}$$

where K_1 and K_2 are as defined in Equation 2.28 and are calculated from IBD sharing between segregants. We used the average information algorithm [47] as implemented in *GCTA* [83] to find the restricted maximum likelihood estimates of the narrow sense heritability, $h^2 = v_1/\mathrm{Var}(Y)$, and the proportion of phenotypic variance from pairwise interactions, $h_2^2 = v_2/\mathrm{Var}(Y)$. By using the estimates of the broad sense heritability[84], we were able to estimate the variance from third and higher order interactions by application of:

$$\sum_{\tau=3}^{n} \frac{v_\tau}{\mathrm{Var}(Y)} = H^2 - (h^2 + h_2^2). \tag{2.32}$$

to estimate $\sum_{\tau=3}^{n} v_\tau/\mathrm{Var}(Y)$, which we define to be $h_>^2$. This is the component of the broad sense heritability which originates exclusively in interactions involving three or more loci.

The standard error of the estimate of $h_>^2$ is estimated as

$$\sqrt{s.e.(\hat{h^2})^2 + s.e.(\hat{h_2^2})^2 + s.e.(\hat{H^2})^2} \tag{2.33}$$

where $\hat{h^2}$ and $\hat{h_2^2}$ are our maximum likelihood estimates, and $\hat{H^2}$ is from [84].

2.3.2.2 Simulation of epistatic traits from yeast data

To test inference of h^2, h_2^2, and $h_>^2$, we simulated 500 phenotypes from the genotypes. For each phenotype, 50 causal variants were sampled independently and at random from across the genome. All 50 causal variants were given additive effects, and all pairs were given interaction effects; 10 of the 50 causal variants were chosen at random to have third order interactions with each other; and 8 of these were chosen at random to have fourth order interactions with each other. The effects were drawn from normal distributions scaled so that $h^2 = 0.4$, $h_2^2 = 0.3$, and $h_>^2 = 0.2$, with the higher order variance equally divided between third and fourth order interactions. Gaussian error was added so that $H^2 = 0.9$.

2.4 Results

2.4.1 Simulations

2.4.1.1 Pairwise interaction variance

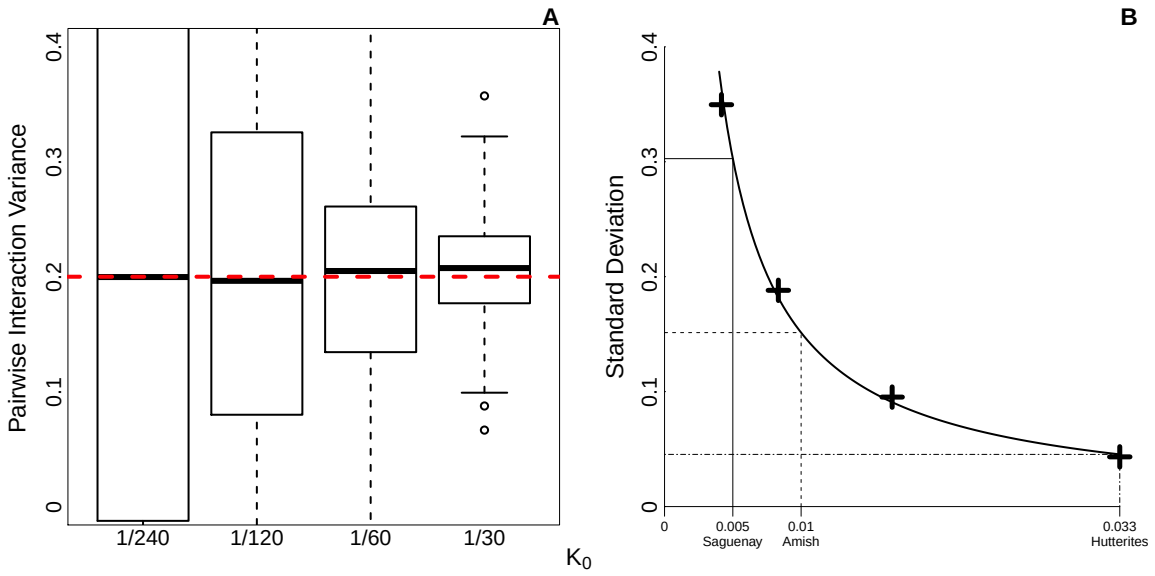


Figure 2.2: **Simulation results for the estimation of the variance from pairwise interactions**. Phenotypes were simulated 500 times for four simulated populations with different mean kinship, each comprised of 5000 individuals. A) shows boxplots of the simulation estimates of the variance from pairwise interactions for the four populations. The dotted red line indicates the true variance from pairwise interactions, 0.2. B) shows the standard deviation of the simulation estimates of the variance from pairwise interactions plotted against the mean kinship of the sample. The points marked on the x-axis correspond to estimates of the mean kinship in Saguenay [85], the Amish [86], and the Hutterites [76]. The curve drawn is proportional to $1/K_0$.

Figure 2.2A shows that the mean estimate is close to the true value for each population; the mean estimate across the four populations was 0.204, indicating the estimation was unbiased. Figure 2.2B shows that the standard devitations of the simulation estimates scale in proportion to $1/K_0$. It may therefore be preferable to use a smaller sample from a more strongly bottlenecked population than a larger sample from a less strongly bottlenecked population.

2.4.1.2 Third order interaction variance

The estimation of the variance from third order interactions was unbiased, with a mean estimate equal to 0.204 across the populations with $K_0 = 1/120, 1/60, 1/30$. The information matrix was not invertible for the model in the population with $K_0 = 1/240$.

The standard deviation for the third order interaction variance was at least twice as large as the standard deviation for the pairwise interaction variance across the populations (Figure 2.3). Even for the most strongly bottlenecked population, the standard error for the third order interaction variance is nearly 15% of the phenotypic variance, comparable to the size of the variance component.

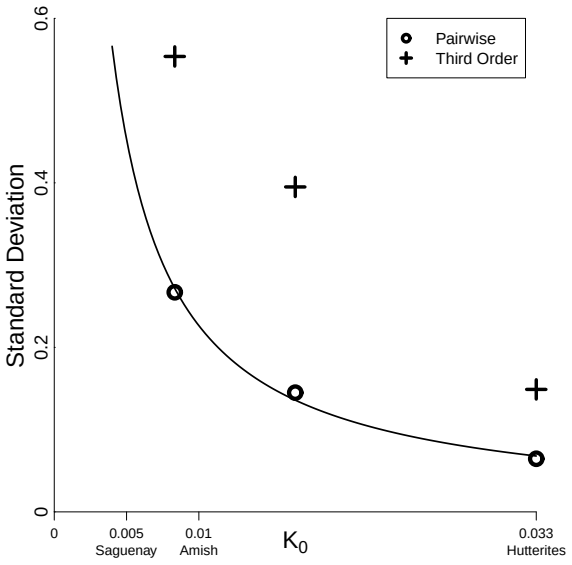


Figure 2.3: **Precision of estimates of variance from third order interactions.** Standard deviations for pairwise and third order variance component estimates in a simulation that includes third order interactions, plotted as in Figure 2.2B.

2.4.1.3 Ignoring epistasis biases additive variance estimates

To investigate possible bias in the estimation of narrow sense heritability that may arise from ignoring epistasis, we fitted models with only additive variance components to the simulation data used in Figure 2.2. We found that the bias did not depend on the mean kinship of the population (Table A.1).

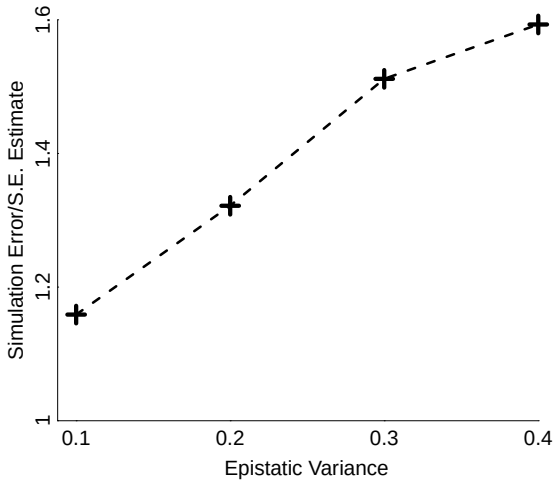


Figure 2.4: **The effect of ignoring epistasis on standard error estimates**. Additive only models were fitted to phenotypes with $v_1 = 0.4$ and v_2 ranging from 0.1 to 0.4. The ratio of the standard deviations of the estimates of the additive variance, denoted simulation error, to the standard error estimates from *GCTA* are plotted on the y-axis.

We simulated additional phenotypes with varying amounts of epistasis ($v_2 = 0.1, 0.2, 0.3, 0.4$) for the population with $K_0 = 1/240$. Table 2.1 shows the proportion of epistatic variance that is detected as additive variance is approximately constant across traits with differing variance from pairwise interactions. The amount of bias is therefore approximately proportional to the amount of epistatic variance; the additive variance estimates were inflated by around 6% of v_2.

v_2	0.1	0.2	0.3	0.4
$100(\hat{v_1} - v_1)/v_2$	6.2	5.9	5.3	5.8

Table 2.1: **The bias in the additive variance estimate for an epistatic trait**. This shows the bias in $\hat{v_1}$, when fitting an additive only model, as a percentage of the pairwise epistatic variance, v_2.

Ignoring epistasis resulted in inaccurate standard error estimates. Figure 2.4 shows that even when only 10% of the phenotypic variance is epistatic, the standard error of the additive variance estimated from simulations is over 15% larger than the standard

error estimated by *GCTA*.

2.4.2 Approximate analytic standard error

The amount of information a sample contains about the epistatic variance of a trait depends on the distribution of kinship in that sample. To better understand how the standard error of the estimator of v_2 depends on the moments of the kinship distribution, we extend the analogy of fitting a quadratic from Figure 2.1 to derive an approximate analytic expression for the standard error.

If we define $R_{s,t} = 2(K_{s,t} - K_0)/(1 - K_0)$, then the process of fitting the covariance matrix implied by (2.23) can be likened to fitting a polynomial in R. Fitting the additive variance, v_1, and the variance from pairwise interactions, v_2, is similar to fitting to all pairs s, t the regression model

$$\Sigma_{s,t} \sim N(v_1 R_{s,t} + v_2 R_{s,t}^2, \sigma^2), \tag{2.34}$$

where $\Sigma_{s,t} = (Y_s - \mathbb{E}[Y])(Y_t - \mathbb{E}[Y])$ is the observed similarity between s and t. If s and t are independent and Y has variance 1, then $\sigma^2 = 1$.

In Appendix A, we derive the asymptotic variance of the maximum likelihood estimator of v_2 in this model. If μ_c is the c^{th} central moment of the distribution of R, then

$$\sqrt{\mathrm{Var}(\hat{v_2})} \geq \eta^{-\frac{1}{2}} \left(\mu_4 - \frac{\mu_3^2}{\mathrm{Var}(R)} \right)^{-\frac{1}{2}} \approx n^{-1} \left(\mu_4 - \frac{\mu_3^2}{\mathrm{Var}(R)} \right)^{-\frac{1}{2}}, \tag{2.35}$$

where η is the number of pairs, and where we take σ to be 1. This implies the standard error scales in proportion to the inverse of the sample size, as has been noted by others for inference of the additive variance[87].

Testing this using the simulation results, we find the standard deviation of the simulation estimates for $K_0 = 1/30, 1/60$ to be very close to (2.35) calculated from

the sample kinship statistics, with the error increasing above (2.35) for smaller K_0 – see Figure 2.5.

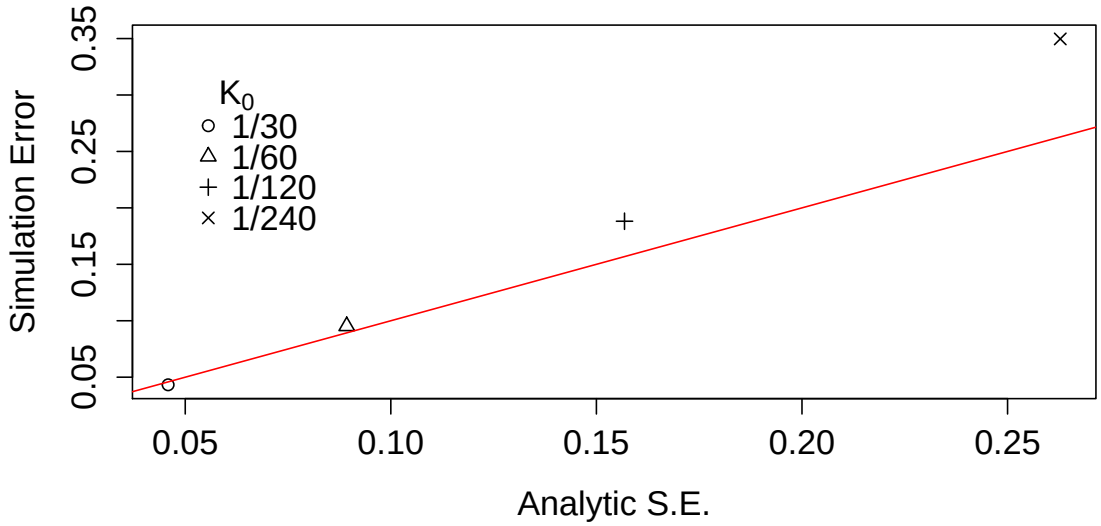


Figure 2.5: **Comparison of approximate analytic standard error to simulation results.** For the four simulated populations, the approximate analytic standard error of the variance from pairwise interactions is compared to the simulation error, the standard deviation of the estimates across simulations. The red line is the line of equality.

The information about the epistatic variance in a sample increases with the fourth central moment of the kinship distribution, which is the unnormalised kurtosis. The samples with heavy tails in their kinship distributions are therefore likely to have the largest kurtosis and most information about epistatic variance.

2.4.3 Yeast cross

We analysed data from a cross of a lab strain (BY) and a wine strain (RM) of yeast [84]. The data included 46 growth phenotypes measured for 1008 haploids dissected from tetrads produced by crossing the two founder strains.

2.4.3.1 Variance components

To establish that our methods worked for these data, we first simulated epistatic traits from the genetic data. Table 2.2 shows the variance component estimates from the simulated traits were unbiased and the standard error estimates were accurate.

	Mean (SD)	Mean $\hat{SE}$
h^2	0.40 (0.05)	0.06
h^2_2	0.30 (0.07)	0.07
$h^2_>$	0.20 (0.09)	0.10

Table 2.2: **Results of simulations of variance component inference using yeast cross data.** The columns are, from left to right, the sample mean (standard deviation) of the estimates, as well as the mean of the standard error estimates, from 500 simulated phenotypes. True values are $h^2 = 0.4$, $h^2_2 = 0.3$, $h^2_> = 0.2$, with the variance from higher order interactions divided equally between third and forth order interactions.

Next we applied our approach to partition the phenotypic variance of the 46 growth phenotypes into additive, pairwise, and higher order genetic components, plus a residual. Figure 2.6 visualises this partitioning. The numerical results of the analysis are in Table A.2. The mean proportion of phenotypic variance explained by pairwise interactions (h^2_2) and higher order interactions ($h^2_>$) was 0.10 and 0.14, respectively.

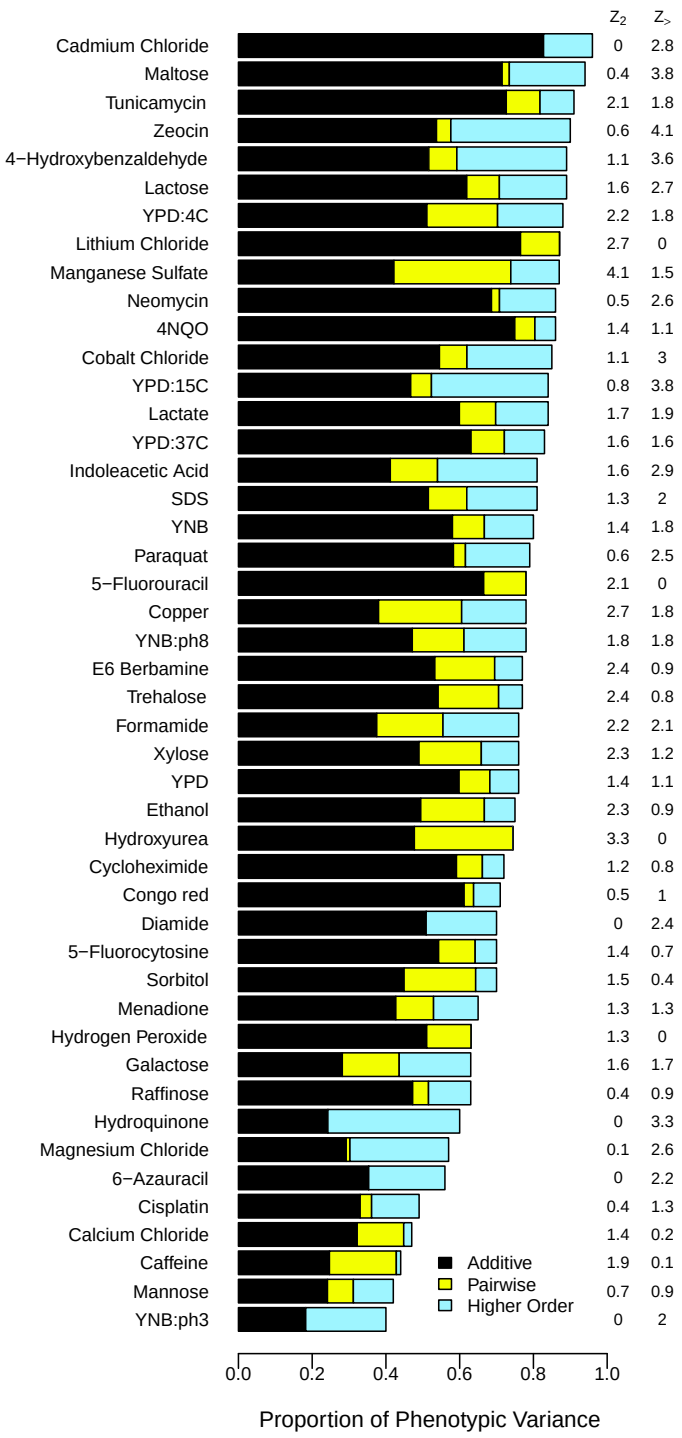


Figure 2.6: **Variance components inferred for 46 different growth traits in the yeast cross**. The length of the bars give the estimated proportions of phenotypic variance explained by the components: additive (black), pairwise interactions (yellow), and interactions of order higher than pairwise (blue). Z_2 gives the estimate of the variance from pairwise interactions divided by the estimated standard error for the trait. $Z_>$ gives the estimate of the variance from third order interactions divided by the estimated standard error.

2.5 Discussion

2.5.1 Theory

We used an approach based on allelic indicator variables to calculate the covariance between individuals in a founder population as a function of their kinship, the mean kinship in the population, and the variance components of the phenotype. This extends the classic result for infinite populations [5] to finite populations. Equations 2.23 and 2.26 together determine the phenotypic covariance matrix, the parameters of which can be estimated by (restricted) maximum likelihood by assuming the phenotype follows a particular distribution. These parameters, along with the central moments of the inbreeding distribution, determine the proportion of population variance explained by different orders of genetic interaction in a founder population. The relationship between (2.23) and Figure 2.1 can be seen by writing the phenotypic covariance as a polynomial function of $R = 2(K - K_0)/(1 - K_0)$, the x-axis of Figure 2.1. The correlation for the epistatic trait in Figure 2.1 as a function of R is

$$v_1 R + v_2 R^2, \tag{2.36}$$

where v_1 is the additive variance in the population and is equal to 0.4.

The model only applies exactly to populations that have been randomly mating since being founded; however, the allelic indicator variable approach could be extended to non-random-mating populations by considering models for non-random inheritance of alleles. Extending the method to explicitly include linkage disequilibrium would be possible but would rely on knowing the linkage disequilibrium between unknown causal alleles.

We have derived the individual and population level variance decomposition for two interacting loci with dominance in a founder population. Finite population his-

tory induces dependence between allelic states, both within and between individuals. This prohibits a simple and exact expression for the covariance between relatives for the additional variance components introduced by dominance. However, except for the most strongly bottlenecked populations, using the identity states implemented by [76] will probably give a good approximation.

We note that an alternative theoretical approach would be to extend the frequency weighted identity-by-state (IBS) estimator employed by [39] to epistatic variance components. The IBS based method of estimating the additive variance was compared to an IBD based approach by [82] using Icelandic data. They found the IBS based approach underestimated the additive variance relative to the IBD based approach by a considerable amount. The same underestimation would be expected to apply to an IBS based epistatic variance estimator, because it originates from incomplete linkage disequilibrium between markers and causal variants. The underestimation could be even more severe for epistatic variance components because, for the variance from an interaction to be properly detected, all of the loci involved in the interaction would have to be in strong linkage with the markers. We therefore argue for the IBD based approach, which takes advantage of the long shared segments present in a founder population to reduce the bias in the estimates.

2.5.2 Simulations and sampling

The simulations (Figure 2.2) suggest that with a sample of $\sim 5,000$ Hutterites, it would be possible to estimate the variance from pairwise interactions with a standard error of less than 5% of the phenotypic variance. The standard error scales in proportion to the inverse of the mean kinship. This explains why one cannot simply use a very large, random sample from an outbred population to fit the epistatic variance.

Including close relatives would increase the precision of the estimator. However, the phenotypic similarity between close relatives could be due to shared environment

as well as shared genetics. The confounding with shared environment could be ameliorated by fitting additional variance components for different relative classes. However, if shared environmental effects extend beyond first degree relatives, the model may become excessively complex. Dominance could also cause phenotypic similarity between siblings above what is expected by additive effects, leading to overestimation of the epistatic variance. For traits that are known to have little dominance variance or shared environmental effects, including close relatives would increase precision without causing bias. Otherwise, very large samples of close relatives would be required to disentangle epistasis, dominance, and shared environment.

Power calculations can be aided by the approximate analytic formula for the standard error of the variance from pairwise interactions. This acts like a lower bound for the standard error in the simulated data – see Figure 2.4. The moments of the kinship distribution can be calculated from a small sample, and, from these, an estimate of the standard error can be calculated for different sample sizes. If this is too high to give a useful estimate, then the sample is probably not appropriate for estimating the variance from pairwise interactions.

Direct estimation of the variance from third order interactions may be beyond the limits of possibility for current human samples. Even with 5,000 Hutterites, the standard error is likely to be at least 15% of the phenotypic variance (Figure 2.3). Unless this component is a large part of the phenotypic variance, it is unlikely that any current samples of human founder populations would provide the power to detect that the component is nonzero.

Founder populations have been used to estimate the narrow sense heritability [82, 81]. We found in simulations that ignoring epistasis leads to a slight overestimation of the additive variance in proportion to the amount of epistatic variance (Tables 2.1 and A.1), as well as underestimation of the standard error (Figure 2.4). This could cause improper calibration of statistical tests. It is possible that these problems could

be reduced by restricting to a smaller range of relatedness, but this would increase the standard error.

2.5.3 Yeast cross

Bloom et al. found evidence for epistatic variance in the difference between H^2 and h^2[84] . In the yeast cross analysis (Figure 2.6 and Table A.2), we have gone further by partitioning the epistatic variance into components arising from pairwise interactions, and from third and higher order interactions. While the individual estimates of the higher order components are not very precise, we provide strong evidence that the variance from pairwise interactions does not in general explain all of the difference between H^2 and h^2. Subsequent work by others has followed a similar approach and found similar results on the pairwise interaction variance, while also taking advantage of a larger sample size to estimate that the variance from third order interactions is small and does not explain all of the remaining estimated broad sense heritability[88].

It is impossible to draw precise conclusions about the relative size of h_2^2 and $h_>^2$ for individual traits, because the method of estimation results in a negative correlation between the estimates. A larger sample from a similarly designed experiment could overcome some of these difficulties and enable direct estimation of the variance from third order interactions.

The statistical importance of pairwise and higher order interactions in the yeast cross cannot be readily generalised to natural populations. For some interaction models, the proportion of the variance which is epistatic rather than additive is greatest for interacting alleles at intermediate frequencies[89, 73]. Therefore, if interactions occur between rarer alleles in natural populations, the proportion of the variance which is epistatic could be reduced.

The large amount of epistatic variance in the cross could be explained by the breakdown of co-adapted variant combinations. The cross is between a lab strain and

a wine strain of yeast, which have diverged under different selection pressures[90]. Given that hybrid incompatibilities were observed between experimentally evolved strains[91], it is plausible that these strains have accumulated them.

The large amount of epistatic variance arising from third and higher order interactions in particular could be explained by the build up of hybrid incompatibilities. Incompatibilities between three or more loci are theoretically expected to be more common than incompatibilities between two loci, because a greater proportion of evolutionary paths to higher order incompatibilities do not pass through a less fit intermediary[92].

2.5.4 Conclusion

These methods can be used to investigate the role of pairwise and higher order epistasis in model organisms by applying them to appropriate crosses. In particular, by measuring the variance that higher order interactions contribute to crosses between diverged populations, these methods could be used to investigate the role of higher order interactions in hybrid incompatibilities.

We anticipate that it will be possible to apply these methods to precisely estimate the variance from pairwise interactions in human founder populations. It may be possible to apply this method to very large samples from outbred populations, such as the full UK Biobank sample of $\sim 500,000$; however, doing so would be computationally challenging due to needing to consider over 100-billion pairs of individuals. These estimates, combined with estimates of the additive and dominant components of the variance, will help in answering where the 'missing' heritability is, in searching for causal loci, in building prediction models, and in testing evolutionary models of traits.

Chapter 3

Gene-by-environment interactions modify the effect of *FTO* variants on body mass index

3.1 Introduction

The genetic variants with the largest effect on BMI variation between individuals are located in an intron of the fat mass and obesity associated (*FTO*) gene[93]. In Europeans, each additional copy of the risk allele at SNP rs1558902, one of the cluster of associated SNPs, increases average BMI by between 0.35 kgm^{-2} and 0.43 kgm^{-2}, explaining around 0.34% of the variation in BMI[93, 67]. The fact that the variants are located in an intron of the *FTO* gene does not establish that they act through that gene. Nevertheless, for convenience, we will follow other authors and refer to the locus and to the associated variants as *FTO*, with specific single nucleotide polymorphisms referenced when relevant.

Progress has been made in understanding the causal mechanisms through which the *FTO* risk alleles increase BMI, implicating regulation of expression of other

genes[94, 95, 96]. A recent study showed that the rs1421085 T-to-C SNP, one of the SNPs with the strongest association with BMI in the *FTO* region, disrupts a conserved motif for the *ARID5B* repressor, leading to a doubling of *IRX3* and *IRX5* expression during early adipocyte differentiation[94]. This was shown to lead to a fivefold reduction in mitochondrial thermogenesis due to a shift in the composition of adipocytes from energy-dissipating beige adipocytes to energy-storing white adipocytes. While this study has uncovered a mechanism through which *FTO* variation can influence adiposity, it is not clear if this is the only mechanism underlying the association with body mass index. It has been shown that the body mass increasing *FTO* variant increases protein intake[97]. This may be the result of compensation for reduced thermogenesis by exploiting the fact that higher protein diets appear to be associated with increased thermogenesis[98].

There have been many small studies examining interactions between the *FTO* locus and various environmental and lifestyle variables. Results have often been inconsistent, especially when comparing studies across different cultures or ethnicities. One possible cause of the inconsistencies may be the difficulty of measuring environmental variables consistently across studies, but low power to detect interactions may also be a contributing factor. Nevertheless, large meta-analyses have found a reduction of the effect of *FTO* on BMI of around 30% in physically active people[99, 31, 100]. *FTO* has also been linked to interactions with diet, especially fried and fatty foods[101, 102, 103], but did not appear to interact with macronutrient intake or dietary energy in a meta-analysis[104].

Meta-analyses typically involve some level of data aggregation within studies before combining across studies. For example, meta-analyses have usually proceeded by dichotomizing continuous or ordinal variables related to physical activity and diet so as to reduce between study heterogeneity. This leads to a loss of power compared to a similar sized study in which it is feasible to use the original measurements[31, 105].

Dichotomizing variables can also reduce the specificity and interpretability of results, which can reduce their utility for public health. The heterogeneity between different studied cohorts, in both measurement of the environment and genetic and cultural heterogeneity, can also reduce power compared to a similarly sized single cohort study[31].

A major challenge in studying environmental risk factors is that many of these are highly correlated with each other. It can then be unclear whether an observed interaction between *FTO* and a particular environmental variable might actually be driven by its correlation with other environmental variables. When they are simultaneously measured on the same individuals, fitting multiple interaction effects jointly can help determine if the environmental variables interact with *FTO* independently of their correlations with each other. Power considerations, and/or lack of the appropriate data, have precluded the fitting of multiple interaction effects simultaneously in smaller studies, and large meta-analyses have typically analysed only one interaction effect at a time[99, 31, 104]. It is therefore unclear whether many of the interactions reported in the literature are truly independent of each other and of other variables.

The UK Biobank is a large prospective study of 500,000 individuals, aged between 40-69 years at recruitment between 2006 and 2010. Extensive measurements and questionnaire responses, including rich lifestyle and environmental information, were gathered from individuals at baseline, and biological samples taken to allow additional assays[106, 57, 58]. The recent interim genetic data release includes genotype data on $\sim 152,000$ of these individuals[58, 107]. The UK Biobank therefore offers a unique opportunity to examine interactions between *FTO* and various lifestyle and environmental variables simultaneously in a large and relatively homogeneous sample.

By joint modelling, we investigated interactions between *FTO* (specifically SNP rs1421085) and physical activity, frequency of alcohol consumption, dietary variation, sleep duration, smoking, TV watching, and socioeconomic-status. We focussed

on rs1421085 following a recent study suggesting this is the causal variant[94]. We note that rs1421085 is highly correlated with the main previously-studied SNPs, in particular rs9939609 ($R^2 = 0.89$). This facilitates comparison with previous interaction studies[99], and we note that the results of our analyses are little changed if rs9939609 is used for the genetic effect.

We found evidence for novel interactions between *FTO* and frequency of alcohol consumption and deviations from mean sleep duration, with the effect of *FTO* diminishing with frequency of alcohol consumption and increasing with deviations from average sleep duration. We estimated that dietary variation has the strongest interaction with *FTO* and made a novel observation that the effect of *FTO* on BMI is enhanced in those who add salt to food more frequently. Our joint modelling increases confidence that the interaction between *FTO* and physical activity is not due to confounding with other lifestyle variables in our model.

These findings increase our understanding of how lifestyle modifies the effect of *FTO* on BMI, which may be indicative of more general mechanisms of relevance to the management of obesity genetic risk.

3.2 Methods

3.2.1 Overview

The UK Biobank data is described in the Introduction. While genetic and environmental data was available for a subset ($\sim 152,000$) of the full sample, only environmental data was available for the remaining individuals ($\sim 350,000$), which we refer to as the 'non-genotyped' sample. We used cross-validation in the 'non-genotyped sample' to learn about the relationship between lifestyle factors and BMI in a joint model including many lifestyle variables. This allowed us to collapse categories of variables into single summary scores. The scores weight the different variables in a category

(e.g., physical activity) by the strength and direction of their association with BMI, while accounting for their correlations with other predictors of BMI. In brief, in the non-genotyped sample, we regressed log-BMI on all the variables in the categories together with other variables associated with BMI. We then used cross-validation to remove variables without any apparent predictive ability[108], and refitted the model on the remaining variables. The fitted coefficients from this model for the measurements in a particular category were then applied to the measured variables in that category for each individual with genotype data to calculate the category score for that individual. The category score can be thought of as the best single predictor of BMI based on the variables in that category. Note that to avoid possible over-fitting, we estimated the coefficients used to calculate the score in a distinct set of individuals from those in which we actually calculated scores. The variables used to construct each score are listed in Table 3.1 under the 'BMI' model.

We added *FTO* and interactions between *FTO* and the lifestyle factors and 'activity' and 'diet' scores from the model selected by cross-validation in the 'non-genotyped sample' to test for interactions between *FTO* and lifestyle factors (the 'Scores' model in Table 3.1). To investigate whether there was evidence for interactions between *FTO* and particular activity or dietary variables, we fitted models that, for each of the 'activity' and 'diet' scores, replaced the score variable with its constituent variables (the 'Activity' and 'Diet' models in Table 3.1).

We fitted models separately in the British and Diverse Samples and tested for heterogeneity between the samples. We combined estimates in a fixed effects meta-analysis if the p-value for the heterogeneity test was above 0.05. We performed 25 interaction tests in total and considered an interaction significant if its p-value was less than the (conservative) Bonferroni-corrected significance threshold of $0.05/25 = 0.002$. The same numbers of tests for main effects were performed, so we used the same significance threshold for these. We report the uncorrected p-values.

Model	BMI	Scores	Activity	Diet
Age and sex (sex, age × sex, age^2, age^2×sex, age^3, age^3×sex)	✓	✓	✓	✓
East co-ordinate	✓	-	-	-
East co-ordinate × age	✓	-	-	-
East co-ordinate × sex	✓	-	-	-
FTO (rs1421085)	-	✓	✓	✓
Activity variables ('Number of days/week walk 10+ mins', 'Number of days/week moderate physical activity 10+ mins', 'Number of days/week vigorous physical activity 10+ mins', and their interactions with age and sex.)	✓	-	✓	-
Activity variables × *FTO*	-	-	✓	-
Activity score	-	✓	-	✓
Activity score × *FTO*	-	✓	-	✓
Diet Variables ('cooked vegetable intake', 'non-oily fish intake', 'oily fish intake', 'processed meat intake', 'poultry intake', 'beef intake', 'lamb/mutton intake', 'pork intake', 'cheese intake', 'bread intake', 'tea intake', 'frequency of added salt')	✓	-	-	✓
Diet Variables × *FTO*	-	-	-	✓
Diet Score	-	✓	✓	-
Diet Score × *FTO*	-	✓	✓	-
Other Variables ('age','alcohol intake frequency', 'sleep duration','sleep duration^2' 'current regular smoker (yes/no)','Townsend deprivation index' , and 'Hours watch tv', and their interactions with age and sex)	✓	✓	✓	✓
Other Variables × *FTO*	✓	✓	✓	✓
Genotyping Array	-	✓	✓	✓
Genotyping Array × *FTO*	-	✓	✓	✓

Table 3.1: **Summary of the variables used as predictors of BMI in each of the models**. An '×' between two variables indicates an interaction effect. The 'BMI' model is the model chosen by the cross validation procedure in the non-genotyped sample (Methods), and the 'Scores' model uses the coefficients fitted in the 'BMI' model to construct the activity and diet scores. The 'Activity', and 'Diet' models each have their relevant score variable replaced with the constituent variables of the score: 'Activity score' replaced with 'Activity variables', etc. Note that to adjust for population structure in the models fitted in the genotyped samples, we added principal components in the British Sample, and we added random effects in a mixed model in the Diverse Sample (Methods).

3.2.2 Measurement of BMI

Height was measured at assessment centres by a Seca 240cm height measure, and weight was measured using a Tanita BC418MA body composition analyser[109].

3.2.3 Selection of lifestyle variables

The variable names here are taken verbatim from the UK Biobank release.

3.2.3.1 Diet

Out of 17 continuous and ordinal dietary intake variables, we selected those without a large amount of missing data in the genotyped sample, where we count those who chose not to answer the question or did not the know the answer as missing. This left 12 variables, 9 of which were ordinal ('Oily fish intake', 'Non-oily fish intake', 'Processed meat intake', 'Poultry intake', 'Beef intake', 'Lamb/mutton intake', 'Pork intake', 'Cheese intake', 'Salt Added to Food') and 3 of which were continuous ('Cooked vegetable intake', 'Bread intake', 'Tea intake'). All of the ordinal variables apart from 'Salt Added to Food' were encoded as: 0, never; 1, less than once a week; 2, once a week; 3, 2-4 times a week; 4, 5-6 times a week; 5, once or more daily. Salt Added to Food uses the encoding: 1, never/rarely; 2, sometimes; 3, usually; 4, always. We found no strong evidence for anything beyond a linear association between log-BMI and the encoding of 'Salt added to food'. We refer to 'Salt added to food' as 'added salt' for convenience. For the remaining continuous intake variables, we excluded individuals who had values above the 99^{th} percentile of the distribution in each sample to prevent being overly influenced by outliers.

3.2.3.2 Physical activity

We chose physical activity variables with near complete observations: 'Number of days/week walk 10 minutes or more', 'Number of days/week of moderate physical

activity 10+ minutes', 'Number of days/week of vigorous physical activity 10+ minutes'.

3.2.3.3 Alcohol

To prevent loss of power, we selected the one alcohol intake variable with near complete data, 'alcohol intake frequency', which is encoded as: 0, never; 1, special occasions only; 2, one to three times a month; 3, once or twice a week; 4, three or more times a week; 5, daily or almost daily. We compared the model fit of the regression of log-BMI on the original encoding of alcohol to the model fit when alcohol is encoded using the midpoint of implied monthly drinking sessions in each category. We found the model with the original encoding fitted better, so we retained the original encoding.

3.2.3.4 Sleep duration

We used individuals answers to the question 'About how much sleep do you get in every 24 hours? (please include naps)', and we excluded individuals in the bottom and top percentiles of the distribution to prevent being overly influenced by outliers.

3.2.3.5 Townsend Deprivation Index

Townsend Deprivation Index[110] was calculated immediately prior to participants joining UK Biobank based on the preceding national census output areas. Each participant was assigned a score corresponding to the output area in which their postcode is located.

3.2.3.6 Smoking

Individuals current smoking status was summarised by UK Biobank as 'Never', 'Previous', and 'Current'. For simplicity, we created a binary variable reflecting whether

they answered 'Current' or not.

3.2.3.7 TV Watching

We took individuals answers to the question 'In a typical day, how many hours do you spend watching TV? (Put 0 if you do not spend any time doing it)', and we excluded those in the upper percentile of the distribution to avoid being overly influenced by outliers. There was an option to put 'Less than an hour a day', which for definiteness we encoded as 0.5 hours a day.

3.2.3.8 Birth co-ordinates

We used north and east co-ordinates (latitude and longitude) of place of birth in the UK as covariates to control for population stratification in the non-genotyped sample. In the genotyped sample, these were dropped in favour of genetic measures, namely principal components or a mixed model (see below).

3.2.4 Modelling

We log-transformed BMI. This can be supported by the fact that BMI is restricted to be positive and fits a log-normal distribution better than a normal distribution (Figure 3.1). Additionally, the effect of *FTO* on BMI can be modelled slightly better on the log scale as can be seen by fitting models for BMI and log-BMI in the British Sample with *FTO*, age, sex, age^2, the confounding variables, and the top 20 principal components as covariates: the variance explained is 6.8% for BMI, whereas for log-BMI it is 7.2%. In addition, the residuals of the fitted model are closer to being normally distributed for log-BMI than when fitting BMI (Figure 3.1). The standard errors of the regression coefficients from models on the log scale should therefore be better calibrated than those from models on the original scale.

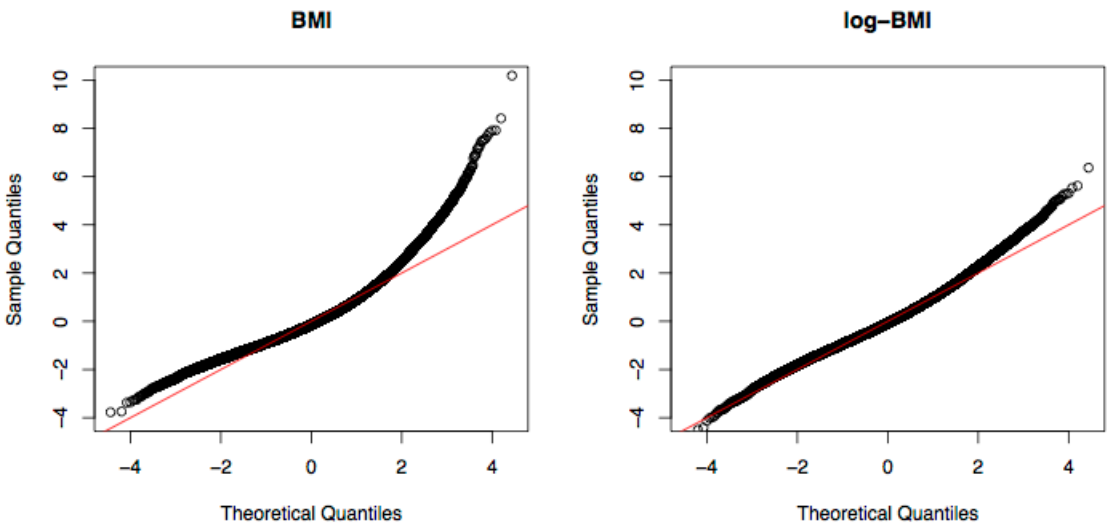


Figure 3.1: **Effect on normality of BMI of log-transform.** Normal quantile-quantile plots for the residuals of the regression of BMI and log-BMI on the model variables, excluding interactions with *FTO*.

If a trait follows a log-normal distribution better than a normal distribution, this may indicate that influences on the trait act multiplicatively rather than additively. This could lead to inflation of test statistics for multiplicative interaction models when fitting them on the original rather than log-transformed scale. Our analysis is therefore likely to be more conservative than those on the original, untransformed scale.

We fit models with pairwise interactions between the number of copies of the *FTO* risk allele and multiple environmental variables. We illustrate this with a simple model where *FTO* interacts with a single environmental variable, x, on the log scale. Formally the model is

$$\log(\text{BMI}) = m + a \times FTO + b \times x + c \times FTO \times x + \epsilon, \tag{3.1}$$

where m is the mean of log-BMI, FTO is the mean-centred number of copies of the *FTO* risk allele, x is the mean centred environmental variable, and ϵ is an independent

normal error term. Transformed back onto the original BMI scale this is:

$$\text{BMI} = e^{m} e^{bx} e^{(a+cx)FTO} e^{\epsilon}. \tag{3.2}$$

The effect of an interaction with an environmental variable x is to either enhance or dampen the proportional change in BMI expected with each additional copy of the *FTO* risk allele. For example, if x is a physical activity variable with a negative interaction with *FTO*, then for those with above average levels of physical activity, each copy of *FTO* may raise BMI by less than 1% on average, whereas for inactive people it may raise BMI by over 1% on average. In the analysis, we generalise this to *FTO* interacting with multiple environmental factors simultaneously, which gives the model the ability to determine if specific environmental factors are interacting with *FTO* independently of their correlations with the other modelled environmental variables. We give an example of the model for two lifestyle factors x and y that have main effects b and c and interaction effects b_{fto} and c_{fto} with *FTO* on BMI:

$$\log(\text{BMI}) = m + b \times x + c \times y + (a + b_{\text{fto}} \times x + c_{\text{fto}} \times y) \times FTO + \epsilon. \tag{3.3}$$

3.2.5 Model selection and score construction

We constructed diet and activity scores from individuals in the UK currently without genotype data. We selected 231,906 of these individuals who had self-declared British ancestry, complete data on the relevant variables, and were known to be born in the UK or Ireland.

We fitted a joint model for log-BMI with age, sex, age^2, age^3, the interactions of sex with age and its square and cube, the north and east co-ordinates of birthplace, and all of the variables listed above. We include interactions with age and sex for: the activity variables, frequency of alcohol consumption, TV watching, sleep duration,

current smoking, and Townsend Deprivation Index. We did not include interactions with age and sex for dietary variables for simplicity.

We calculated the t-statistic of the marginal regression coefficients in R, and we used cross-validation to select a t-statistic magnitude threshold for inclusion of the variables in the model. We used ten-fold cross validation to select the model with the highest estimated out-of-sample R^2, which was 13.01% and corresponded to a t-score threshold of 1.15 and a model with 42 variables. We call this model the 'BMI' model and list its variables in Table 3.1.

To estimate the coefficients of the scores, we refitted the BMI model on the sample comprised of all ten folds combined, in which it had an R^2 of 13.03%. We then used the coefficients from this model fit to calculate diet and activity scores for both the British and Diverse Samples. We used the calculated diet and activity scores along with the other variables kept by cross-validation to assess evidence for interactions with *FTO* in the Scores, Activity and Diet models (Table 3.1).

3.2.6 Genotype data

We used the sample split into 'British' and 'Diverse' subsamples as described in the Introduction.

We obtained genotypes for SNP rs1421085. At this SNP, 0.20% of calls are missing in the British Sample, and 0.24% of calls are missing in the Diverse Sample. We excluded samples with missing calls. The frequency of the minor allele is 40.03% in the European sample and 39.93% in the Diverse Sample.

The final British Sample has 89,552 individuals with no close relatives genotyped, and the final Diverse Sample has 29,580 individuals, containing close relatives.

Individuals in the UK Biobank interim data release were genotyped on one of two very similar genotyping chips, called the Axiom UKBiLEVE or Axiom UKBiobank array. As recommended in the UK Biobank QC documentation[58], we include the

array on which the individual was genotyped as a covariate in all analyses, and we also include the interaction of *FTO* with genotyping array as a variable in all the models (Table 3.1).

3.2.7 Control of population structure

For the British Sample, we calculated principal components from the sample determined to be genetically British by UK Biobank. We LD-pruned SNPs using PLINK in a sliding window of size 1000 to ensure that no pair of SNPs within the window had an R^2 of more than 0.1. We filtered out SNPs with minor allele frequency less than 0.05, missingness greater than 1%, and Hardy-Weinberg exact test p-value less than 10^{-6}. This left $\sim 104,000$ SNPs across the genome. We used EIGENSOFT[111] with fastmode[112] on to calculate the top 20 principal components. We fitted the Scores, Activity, Alcohol, and Diet models (Table 3.1) with the top 20 principal components added.

For the Diverse Sample, we used a mixed model to ameliorate confounding due to family relatedness and population structure not captured by principal components[42, 45, **?**, 49]. We filtered out SNPs with minor allele frequency less than 0.01, with more than 1% missing calls, and Hardy-Weinberg equilibrium exact test p-value less than 10^{-10}. We used a stronger threshold for the Hardy-Weinberg equilibrium exact test for the Diverse Sample because, while we wanted problematic SNPs with gross violations of equilibrium to be removed, Hardy Weinberg equilibrium is not expected to hold exactly in ethnically mixed samples. To fit the models in the Diverse Sample, we used a mixed model with two random effects: one from the SNPs on chromosomes other than 16, and one from the SNPs on chromosome 16 more than 2 centi-Morgans away from rs1421085, where genetic distance was determined using the genetic map provided by UK Biobank. We calculated genetic relatedness matrices using GCTA[83], and fitted the models using the Average Information algorithm in GCTA. These cor-

respond to the maximum likelihood estimates of the fixed effects given the variance components that maximise the restricted likelihood.

3.2.7.1 Efficacy of population structure control

If population structure has been controlled effectively and there are no true causal loci, then the association test statistics at independent SNPs across the genome should be sampled from the null distribution. A common measure of effectiveness of control of population structure is the inflation factor[113]: this estimates the ratio of the median test statistic across the genotyped variants to the median that would be expected from the null distribution of test statistics. A weakness of this measure is that if a trait has many causal variants, which BMI is known to have[114, 67], then the inflation factor should be greater than 1 even if population structure has been controlled for perfectly[115]. In the following, we calculate inflation factors for SNPs across the genome to measure how effective our control of population structure is in both samples. To test whether a mixed model could control for the kind of structure in the Diverse Sample, we used BOLT-LMM[46], with the LMM-Inf setting, to calculate association statistics between log-BMI and the SNPs on the chromosomes other than 16, which contains the *FTO* locus. We used the BMI model (Table 3.1) variables as fixed effects, excluding any interactions with *FTO*. We used BOLT-LMM instead of GCTA because of the greater computational efficiency. (Note that for this analysis we undertake association analyses at SNPs genome-wide, whereas our primary analyses are focused on a single *FTO* SNP.) The results should be comparable because BOLT-LMM with the LMM-inf setting fits the same infinitesimal mixed model as GCTA. The inflation factor over the tested chromosomes was 1.07, which is lower than 1.09 reported for a BMI meta-analysis[115].

We measured how effective adjusting for the top 20 principal components in the British Sample was at controlling population structure by computing association

statistics for a sample of SNPs across the genome. To ensure the association test statistics were comparable to our *FTO* analysis, we used the same code and model within *R* as for the primary analysis. However, this imposed computational constraints, preventing a genome-wide analysis. We therefore selected 100 SNPs from each chromosome, leaving a gap of 100 genotyped SNPs between each selected SNP. We kept those with minor allele frequency $> 5\%$ and missingness $< 1\%$, leaving 872 SNPs. We used the Scores model (Table 3.1) with all of the *FTO* variables removed and replaced with the test SNP. The inflation factor was 1.12. While the inflation factor is higher than in the Diverse Sample, it is close to the inflation factor of 1.09 reported for a BMI meta-analysis[115]. The consistency of our estimates of the effects of *FTO* across the two samples (Figure 3.2A) also argues against our analysis being overly contaminated by population structure. Many other GWAS studies of samples taken from UK Caucasians have also shown that population structure is not a major factor[116, 117].

3.2.8 Nutrient analysis

We used the nutrient estimates calculated by UK Biobank from the 24 hour dietary recall questionnaire[118] to investigate which nutritional variables were associated with BMI and dietary variables. The variables are continuous and non-negative, with many extreme upper outliers. We therefore set to missing observations from individuals who had values in the top percentile of any of the nutrient variables. We used the 'Scores' model (Table 3.1) variables without the *FTO* variables or the diet score as covariates in linear models for log-BMI, the diet score, cooked vegetable intake and frequency of added salt. There were 12,747 in the British Sample and 4,413 individuals in the Diverse Sample with complete observations of these variables. If effects were consistent between the British and Diverse Samples, they were combined in a fixed effects meta-analysis.

3.3 Results

3.3.1 Baseline characteristics

	British	**Diverse**
Sample size	89,552	29,580
BMI	27.4 (4.69)	27.4 (4.82)
Age (years)	56.8 (7.93)	55.6 (8.21)
% Male	48.2%	46.1%
Copies of rs1421085 risk allele	0.801 (0.69)	0.799 (0.694)
Townsend deprivation index	-1.64 (2.9)	-0.881 (3.26)
Sleep duration (hours per night)	7.18 (1.05)	7.13 (1.11)
% Regular Tobacco Smoker	9%	10.1%
Hours watch TV per day	2.78 (1.57)	2.71 (1.68)
Alcohol intake frequency (0-5)	3.2 (1.47)	2.96 (1.56)
Number of days/week walk 10+ mins	5.37 (1.96)	5.38 (1.97)
Number of days/week moderate physical activity 10+ minutes	3.59 (2.33)	3.62 (2.35)
Number of days/week vigorous physical activity 10+ mins	1.84 (1.94)	1.91 (2)
Cooked vegetable intake (heaped teaspoons per day)	2.69 (1.57)	2.85 (1.95)
Oily fish intake*	1.64 (0.912)	1.66 (0.942)
Non-oily fish intake*	1.81 (0.765)	1.77 (0.8)
Processed meat intake*	1.92 (1.04)	1.82 (1.09)
Poultry intake*	2.32 (0.859)	2.3 (0.911)
Beef intake*	1.48 (0.821)	1.42 (0.869)
Lamb/Mutton intake*	1.11 (0.694)	1.12 (0.749)
Pork intake*	1.16 (0.698)	1.11 (0.772)
Cheese intake*	2.56 (1.06)	2.51 (1.1)
Bread intake (slices per week)	12.7 (8.35)	12.3 (8.49)
Tea intake (cups per day)	3.54 (2.77)	3.4 (2.76)
Frequency of added salt (1-4)	1.64 (0.855)	1.71 (0.899)

Table 3.2: **Baseline characteristics of the samples**. The mean and standard deviation (in brackets) are shown. A * indicates that the variable is encoded as: 0, never; 1, less than once a week; 2, once a week; 3, 2-4 times a week; 4, 5-6 times a week; 5, once or more daily. Alcohol intake frequency is encoded as: 0, never; 1, special occasions only; 2, one to three times a month; 3, once or twice a week; 4, three or four times a week; 5, daily or almost daily. For frequency of added salt, the categories are: 1, never/rarely; 2, sometimes; 3, usually; 4, always.

The baseline characteristics of the samples are recorded in Table 3.2. The most striking difference between the two groups is in the variance of Townsend deprivation

index. On further examination, the greater variance in the Diverse Sample is due in part to the presence of subsamples with higher mean and higher variance in Townsend deprivation index than in the British Sample: those with 'Mixed', 'Asian or Asian British', 'Black or black British', or 'Other' self-declared ethnicities.

3.3.2 Main effects and interactions with *FTO*

Unless otherwise stated, we modelled BMI on the log-scale, and we report estimated effects transformed back onto the original scale, where we express them as a percentage change in BMI per standard deviation (S.D.) of the relevant predictor. We give the 95% confidence intervals in square brackets.

Variable	Estimate	95% C.I.	p-value
Activity Score	-0.19	(-0.34,-0.05)	1.0e-02
	-0.35	(-0.6,-0.1)	5.8e-03
	-0.23	(-0.36,-0.11)	3.1e-04
Alcohol Freq.	-0.28	(-0.43,-0.13)	2.8e-04
	-0.12	(-0.38,0.14)	3.6e-01
	-0.24	(-0.37,-0.11)	3.0e-04
Diet Score	0.25	(0.11,0.4)	7.0e-04
	0.43	(0.17,0.69)	1.1e-03
	0.30	(0.17,0.43)	5.0e-06
Sleep Squared	0.13	(0.04,0.22)	4.6e-03
	0.14	(-0.01,0.3)	7.3e-02
	0.13	(0.06,0.21)	8.0e-04
Added Salt	0.23	(0.08,0.38)	2.9e-03
	0.17	(-0.09,0.44)	1.9e-01
	0.21	(0.08,0.34)	1.2e-03

Table 3.3: **Summary of the variables with evidence for interactions with *FTO*.** The table shows the estimated interaction effect with *FTO* expressed as the % change in BMI per copy of *FTO* and per S.D. of the variable. The first line for each variable gives the estimate in the British Sample, the second line gives the estimate in the Diverse Sample, and the third line gives the combined estimate. 'Sleep Squared' refers to squared deviations from mean sleep duration. 'Added salt' refers to the frequency of adding salt to food.

Our primary analyses involved fitting the Scores Model (Table 3.1) that jointly models all main effects of BMI associated variables, including the activity and diet

scores, and their interactions with *FTO*. We report effects from fitting this model (Figure 3.2), and additionally we report main and interaction effects for the components of the activity (Figure 3.3) and diet (Figure 3.5) scores from the Activity and Diet models (Table 3.1). Where there is no strong evidence for heterogeneity of effects between samples, we report only the combined estimate in the main text. Table 3.3 gives a statistical summary of the estimated effects.

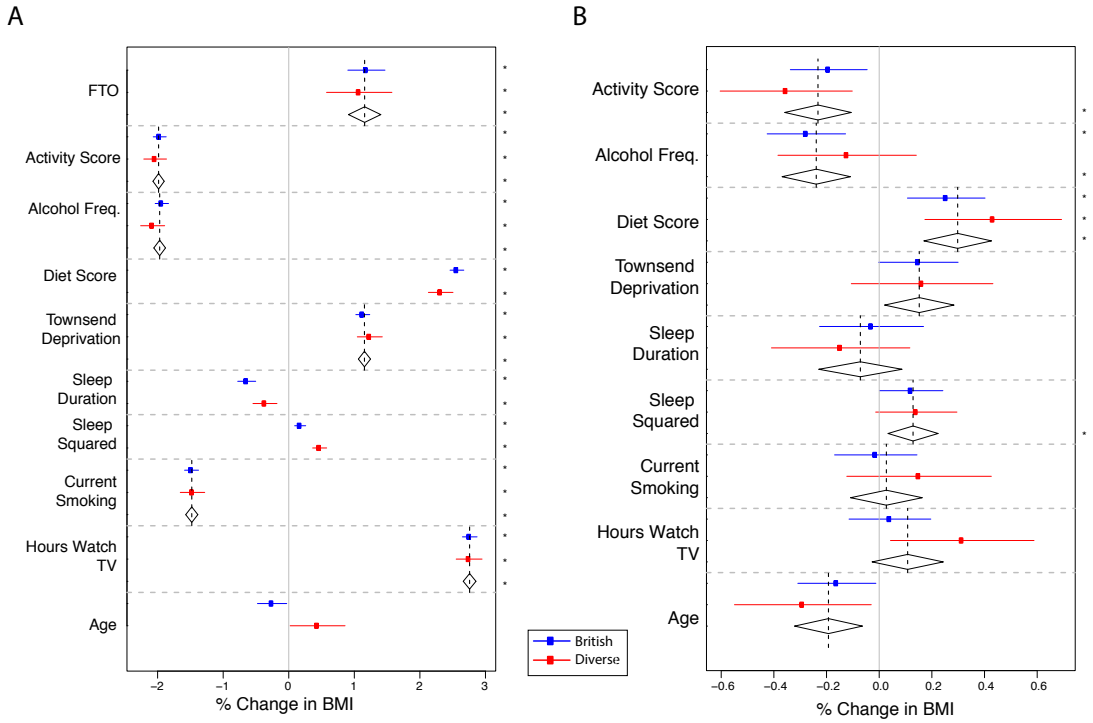


Figure 3.2: **Main effects and interactions with *FTO***. Estimated A) main effects on BMI (% change in BMI per risk allele for *FTO*, per decade for age, and per standard deviation for other variables) B) interaction effects with *FTO* on BMI (% change in BMI per *FTO* risk allele per decade for age, and % change in BMI per *FTO* risk allele per standard deviation for other variables). All main and interaction effects were fitted jointly in the 'Scores' model (Table 3.1) in both the British ($n \sim 90,000$) and diverse ($n \sim 30,000$) samples. The estimated effects are shown along with their 95% confidence intervals in both the British (blue) and diverse (red) samples along with the combined estimate from a fixed effects meta-analysis when no significant heterogeneity between samples was observed (diamonds). 'Sleep Squared' refers to squared deviations from mean sleep duration. A star on the right indicates a p-value below the Bonferroni corrected significance threshold of 0.05/25=0.002

3.3.2.1 *FTO*

In our primary analyses of log-BMI in the Scores Model (Table 3.1), we estimate that each additional copy of the rs1421085 risk allele is associated with a BMI increase of 1.17% in the British Sample ([0.90%,1.44%], $p = 1.0 \times 10^{-17}$), and 1.07% in the Diverse Sample ([0.58%,1.57%], $p = 2.2 \times 10^{-5}$). There is no evidence for heterogeneity (p=0.73), with a combined estimate of 1.15% ([0.91%,1.38%], $p = 1.22 \times 10^{-21}$) (Figure 3.2A). To check comparability with earlier studies, we also fitted the Scores Model (Table 3.1) on untransformed BMI in the British Sample, giving an estimated additive effect of 0.34kgm^{-2}, with a 95% confidence interval of [0.26 kgm^{-2}, 0.41 kgm^{-2}]. This is in agreement with a previous meta-analysis (total $N \sim 250,000$) estimate of 0.39kgm^{-2} for rs1558902, another *FTO* SNP in strong linkage disequilibrium with rs14210855[67].

3.3.2.2 Physical activity

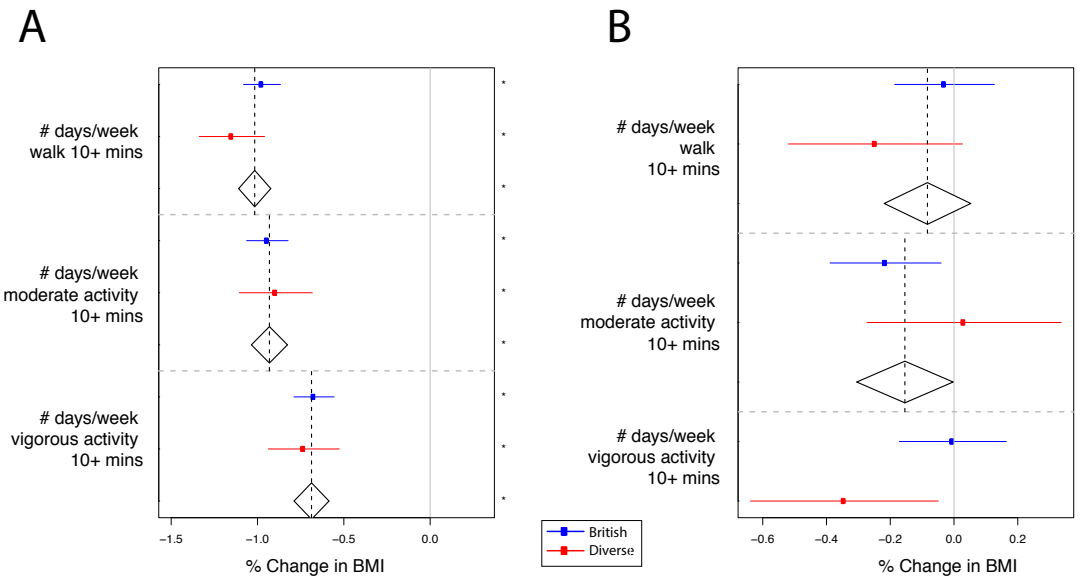


Figure 3.3: **Main effects and interactions with *FTO* of activity variables**. For the components of the activity score, estimated A) main effects on BMI (% change in BMI per standard deviation) B) interaction effects with *FTO* on BMI (% change in BMI per *FTO* risk allele per standard deviation). All main and interaction effects were fitted jointly in the 'Activity' model (Table 3.1) in both the British ($n \sim 90,000$) and diverse ($n \sim 30,000$) samples. The estimated effects are shown along with their 95% confidence intervals in both the British (blue) and diverse (red) samples along with the combined estimate from a fixed effects meta-analysis when no significant heterogeneity between samples was observed (diamonds). A star on the right indicates a p-value below the Bonferroni corrected significance threshold of 0.05/25=0.002

The activity score is associated with a reduction in BMI of 1.98% ([-2.07%,-1.90%], $p < 10^{-30}$) per S.D., and there is strong evidence that all the individual activity variables are also associated with reduced BMI in each sample (Figure 3.3A). We found that *FTO* interacts with the activity score (-0.23%, [-0.36%,-0.11%], $p = 3.1 \times 10^{-4}$), with *FTO* having a weaker effect in more physically active individuals.

Physical activity is the variable with the strongest prior evidence for an interaction with *FTO*[99, 31]. There is strong evidence for an interaction between *FTO* and physical activity from meta-analyses of North American cohorts and combined European and North American cohorts. Some but not all individual studies in European

cohorts have found statistically significant interactions between physical activity and *FTO* on BMI[119, 120, 121, 122, 123]. Large meta-analyses have not found statistically significant interactions between *FTO* and physical activity when restricted to European cohorts[99, 31]. To aid comparison with a previous meta-analysis[99], we fitted the 'Scores' model (Table 3.1) on untransformed BMI with the activity score dichotomised at its 20th percentile. Our estimate for the interaction between physical activity and *FTO* on BMI in this model is -0.19 kgm^{-2} difference in per copy *FTO* effect ([-0.29 kgm^{-2}, -0.08 kgm^{-2}], $p = 0.001$), which is larger ($p = 0.075$ for difference) than a meta-analysis estimate of European cohorts (n=164,307: -0.06 kgm^{-2}, [-0.16,0.03] kgm^{-2}, $p = 0.18$)[99]. We note that the estimate is very close to the estimate from the EPIC Norfolk cohort reported in the meta analysis (-0.18 kgm^{-2}, [-0.34 kgm^{-2},-0.02 kgm^{-2}])[99], which is a cohort of similar genetic, cultural and age composition to the one we analyse[119]. Our estimate is thus intermediate between a meta-analysis of North American cohorts estimate (n=47,938: -0.49 kgm^{-2}, [-0.65,-0.33] kgm^{-2}) and the meta-analysis of European cohorts estimate[99] and is consistent with the EPIC Norfolk estimate, another large study in a single British cohort. Our results thus support the picture in the literature of a larger interaction effect in North American cohorts compared to European cohorts.

3.3.2.3 Alcohol consumption

Frequency of alcohol consumption is associated with a decrease in BMI of 1.97% per S.D. (95% confidence interval [-2.06%,-1.88%], $p < 10^{-30}$, Figure 3.2A). This is in agreement with previous studies that have observed the number of days per week that an individual drinks alcohol is inversely associated with BMI, whereas total alcohol intake is positively associated[124, 125].

While *FTO* has been shown to affect alcohol consumption patterns[126], alcohol consumption patterns have not been previously found to modify the effect of *FTO*

on BMI. We found that the effect of *FTO* on BMI is diminished with more frequent consumption of alcohol (-0.24% ([-0.37%,-0.11%], $p = 3.0 \times 10^{-4}$).

3.3.2.4 Diet score

The diet score was associated with an increase in BMI of 2.56% per S.D. in the British Sample ([2.45%,2.67%], $p < 10^{-30}$) and 2.32% ([2.13%,2.50%], $p < 10^{-30}$) in the Diverse Sample, with some evidence for heterogeneity ($p = 0.024$). To better understand which nutritional properties of diet were driving the association between diet score and BMI, we took advantage of a small subsample ($\sim 12,500$ in the British Sample; $\sim 4,500$ in the Diverse Sample) of people who had a 24 hour diet recall questionnaire administered, from which UK Biobank estimated nutrient quantities[118]. Figure 3.4 shows the estimated effects of the nutrients on BMI and the diet score (see Methods for details). Protein, food weight, and saturated fat all had strong positive associations with both BMI and the diet score, and these associations were consistent across the two samples.

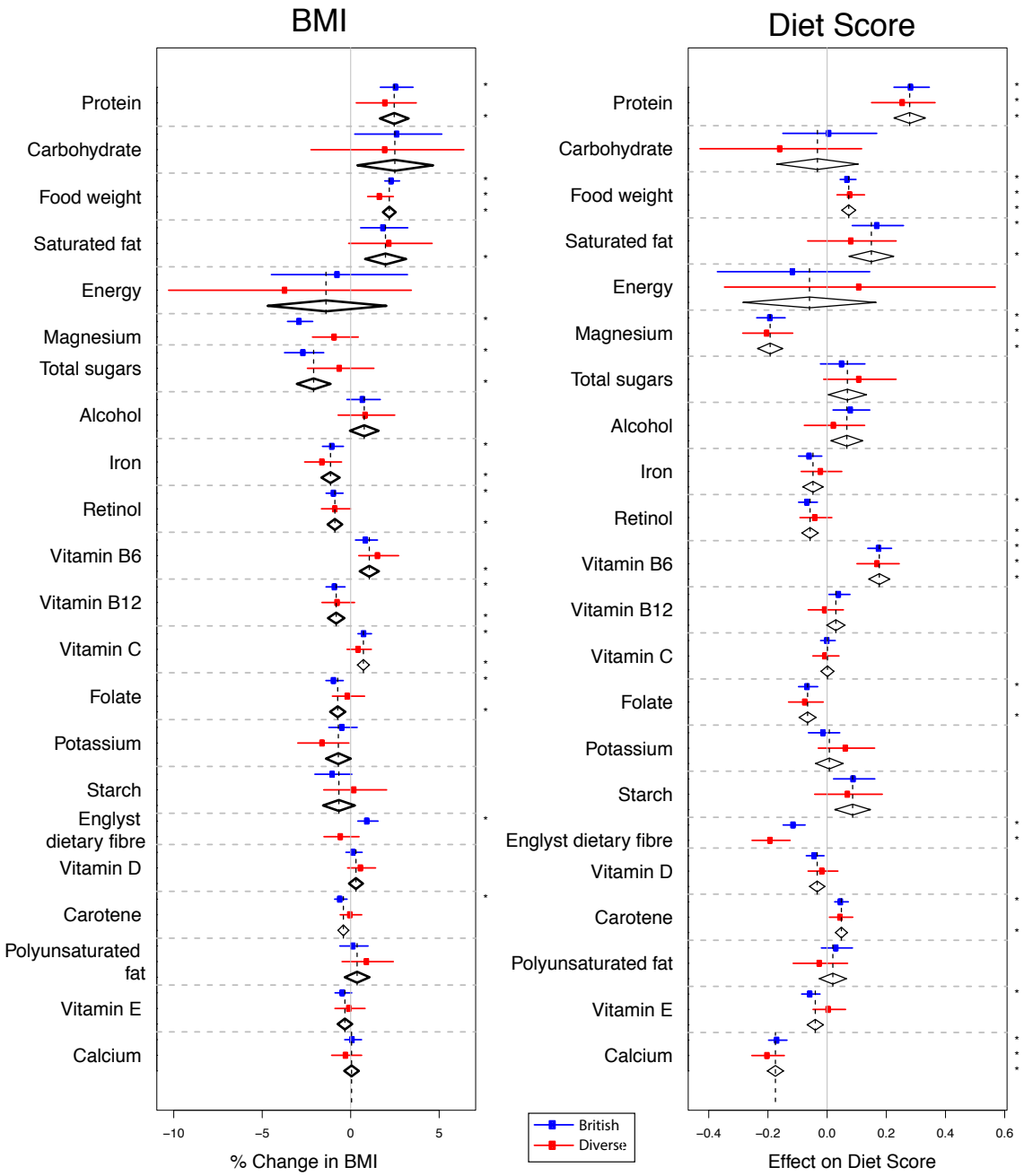


Figure 3.4: **The associations of different nutrient quantities with BMI and diet score**. Nutrient quantities were estimated from 24 hour dietary recall. Nutrients were fitted jointly along with variables from the 'BMI' model (Table 3.1 and Methods). For BMI, the effects are expressed as the percentage change in BMI per standard deviation of the nutrient, and for the diet score the effect is the standard deviation change in diet score per standard deviation of the nutrient. The estimated effects and 95% confidence intervals are plotted for each sample: the British Sample (n=12,747, blue) and the Diverse Sample (n=4,413, red). If there is no statistically significant heterogeneity ($p > 0.05$) between the samples, a combined estimate from a fixed effects meta-analysis is also plotted (diamonds). A star on the right indicates the p-value below the Bonferroni corrected significance threshold of 0.05/22.

We found that the effect of *FTO* on BMI is enhanced in individuals with a higher diet score (0.30% ([0.17%,0.43%], $p = 5.0 \times 10^{-6}$), the strongest estimated *FTO* interaction effect in the joint model.

3.3.2.5 Dietary components

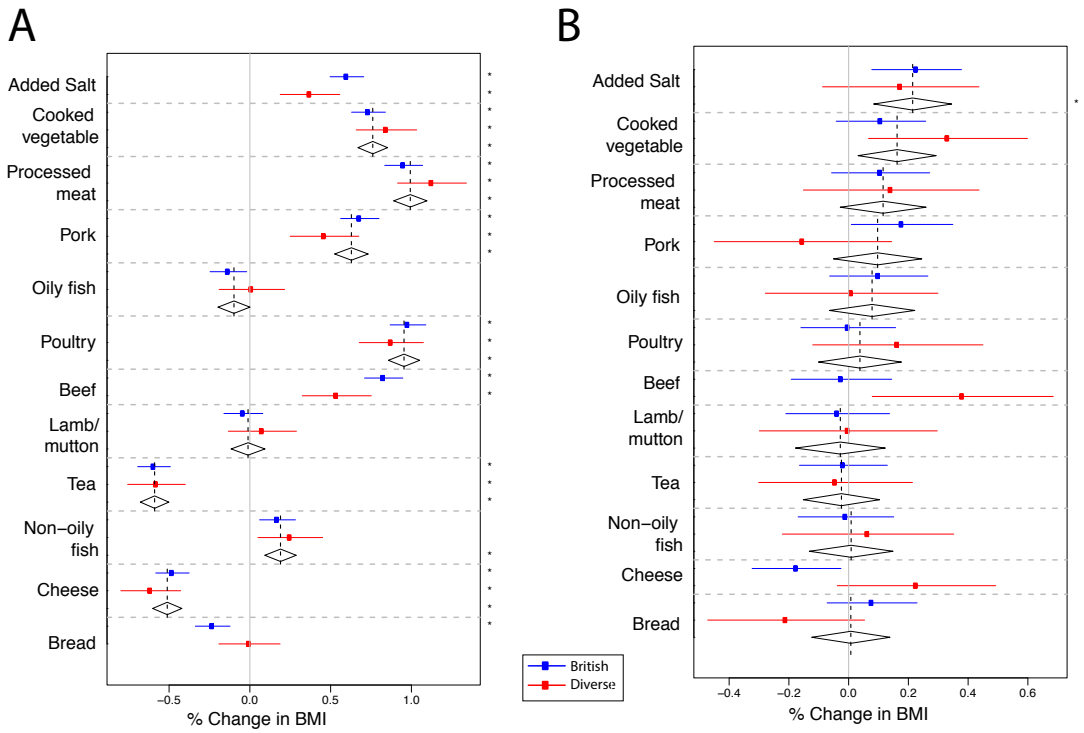


Figure 3.5: **Main effects and interactions with *FTO* of dietary variables**. For the components of the diet score, estimated A) main effects on BMI (% change in BMI per standard deviation) B) interaction effects with *FTO* on BMI (% change in BMI per *FTO* risk allele per standard deviation). All main and interaction effects were fitted jointly in the 'Diet' model (Table 3.1) in both the British ($n \sim 90,000$) and diverse ($n \sim 30,000$) samples. The estimated effects are shown along with their 95% confidence intervals in both the British (blue) and diverse (red) samples along with the combined estimate from a fixed effects meta-analysis when no significant heterogeneity between samples was observed (diamonds). A star on the right indicates a p-value below the Bonferroni corrected significance threshold of 0.05/25=0.002.

Next we investigated the relationship between specific components of the diet score and BMI (Figure 3.5A). We found that added salt (how frequently one adds salt to food) was associated with increased BMI in both the British Sample (0.60%,

[0.49%,0.71%], $p < 10^{-30}$) and the Diverse Sample (0.37%, [0.19%,0.56%], $p = 7.5 \times 10^{-5}$), with evidence that the size of the effect differs between the samples ($p = 0.034$). Added salt was associated with food energy estimated from 24-hour diet recall ($p = 1.2 \times 10^{-3}$, Figure 3.6A).

Cooked vegetable intake is consistently associated with increased BMI in both the British and Diverse Samples (combined estimate: 0.76%, [0.67%,0.85%], $p < 10^{-30}$). In the detailed nutrient study, cooked vegetable intake is associated with increased protein, carbohydrate, and food weight (Figure 3.6), which are all associated with increased BMI (Figure 3.4), possibly explaining the positive association between cooked vegetable intake and BMI.

Having observed the interaction between *FTO* and diet score, we investigated whether there was evidence for an interaction between *FTO* and any of the 12 variables comprising the diet score (Figure 3.5B). The strongest evidence for any particular dietary variable interacting with *FTO* is for added salt (0.21%, [0.08%,0.34%], $p = 1.2 \times 10^{-3}$), with the estimated interaction effect of similar size to the interaction with the activity score; again, the effect of *FTO* is increased for individuals who add salt to food more frequently.

We also tested for interactions between *FTO* and the estimated nutrient quantities for the subset of the British Sample for whom a 24-hour dietary recall questionnaire had been administered ($n \sim 12,500$, so substantially less powered than in our main analyses), and we did not find any statistically significant evidence for interactions (data not shown).

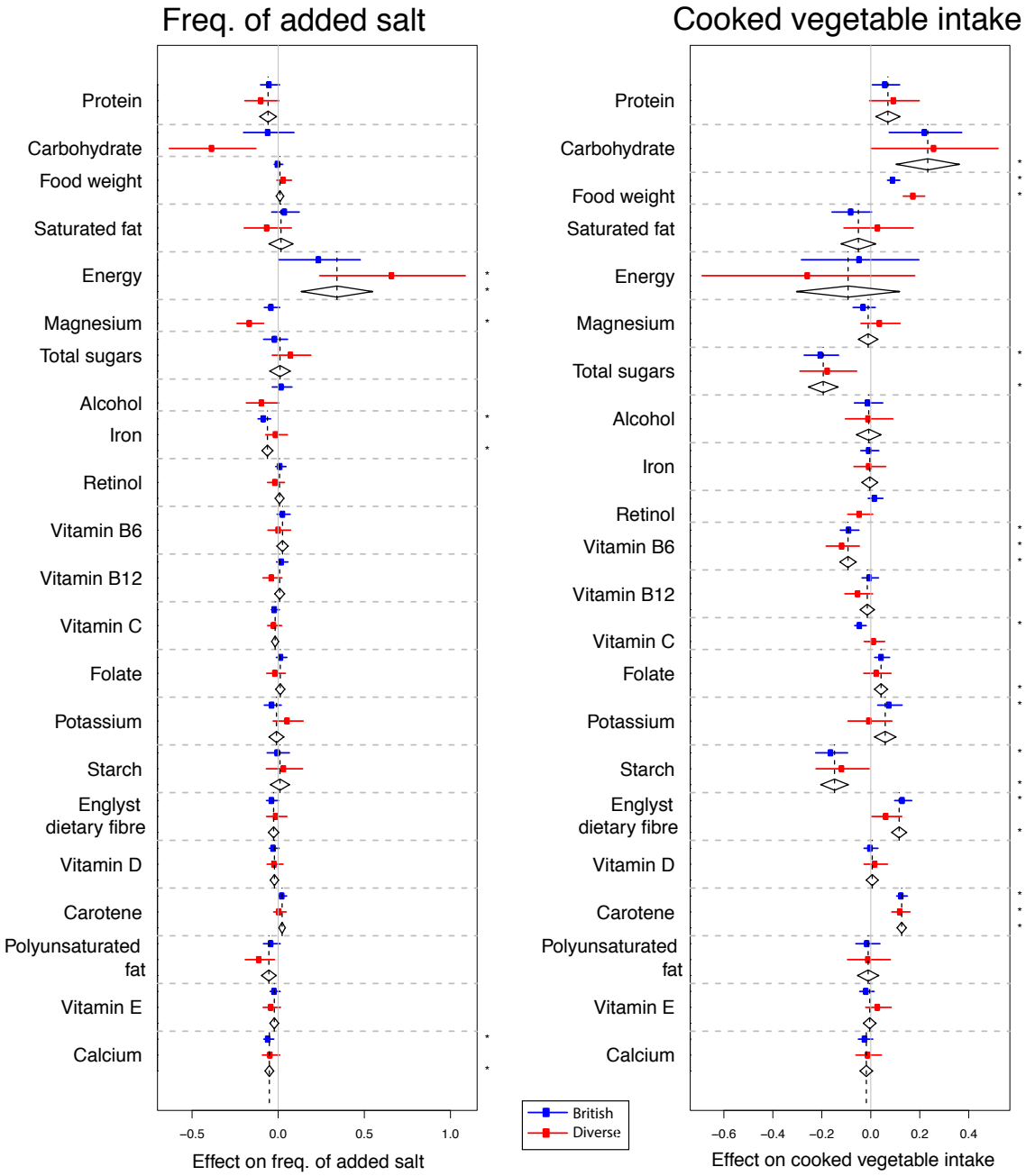


Figure 3.6: **The associations of different nutrient quantities with frequency of added salt and cooked vegetable intake**. Nutrients were fitted jointly along with variables from the 'BMI' model (Table 2 and Methods), excluding dietary variables and *FTO*. The effects are expressed as the standard deviation change per standard deviation of the nutrient. The estimated effects and 95% confidence intervals are plotted for each sample: the British Sample (n=12,716, blue) and the Diverse Sample (n=4,418, red). If there is no statistically significant heterogeneity ($p > 0.05$) between the samples, a combined estimate from a fixed effects meta-analysis is also plotted (diamonds). A star on the right indicates the p-value below the Bonferroni corrected significance threshold of 0.05/22.

3.3.2.6 Sleep

A non-linear U-shaped relationship between sleep duration and BMI has been observed[127]. We therefore fitted both sleep duration and the squared deviations from mean sleep duration as effects on BMI, finding that the linear term is associated with reduced BMI (British: -0.48%, [-0.58%,-0.38%], $p = 7.7 \times 10^{-20}$; Diverse: -0.36%, [-0.55%,-0.18%], $p = 1.1 \times 10^{-4}$. Figure 3.2A), while the squared deviation is associated with increased BMI (British: 0.42%, [0.36%,0.49%], $p < 10^{-30}$; Diverse: 0.47%, [0.36%,0.58%], $p = 4.7 \times 10^{-18}$.). This is in agreement with previous studies showing that more sleep is associated with lower BMI in a small range around the average, while more extreme deviations from the average amount of sleep are associated with increased BMI.

There was no evidence that linear variation in sleep duration modifies the effect of *FTO* on BMI (-0.02%, [-0.17%,0.13%], $p = 0.43$), whereas there was evidence of an interaction between *FTO* and the squared deviations from mean sleep duration, with increases in the squared deviation from mean sleep enhancing the effect of *FTO* on BMI (0.13%, [0.06%,0.21%], $p = 8.0 \times 10^{-4}$).

3.3.2.7 Townsend Deprivation Index

Lower socio-economic status is correlated with many lifestyle and environmental factors, and it has been shown to be associated with higher BMI in developed countries[128]. The Townsend Deprivation Index is a combined measure of indicators of socio-economic deprivation in a geographic region[110] and in our data it is associated with increased BMI (1.15%, [1.06%,1.25%], $p < 10^{-30}$). The estimated interaction between *FTO* and Townsend Deprivation Index was not significant after Bonferroni correction (p = 0.035).

3.3.2.8 Age

We found evidence that age is associated with reduced BMI in the joint model in the British Sample (-0.25% per decade, [-0.48%,-0.03%], $p = 0.026$, Figure 3.2A), although its univariate correlation with BMI is positive. In the Diverse Sample, age is associated with increased BMI in the joint model (0.44% per decade, [0.02%,0.86%], $p = 0.040$ in the Diverse Sample). The age range in the genotyped sample is between 39 and 70, so the effect of age estimated here reflects the difference between middle and older age, and is not informative for ages outside this range. While we saw evidence for an interaction between *FTO* and age, this was not significant after Bonferroni correction ($p = 0.006$).

3.3.2.9 TV watching

TV watching has been shown to strongly correlate with BMI[129], and does so in our data (2.76%, [2.66%,2.85%], $p < 10^{-30}$). In contrast to a previous study[129], we do not find strong evidence for an interaction between TV watching and *FTO* (0.11%, [-0.03%,0.24%], $p = 0.15$). However, we note that fitting a non-joint model with only the interaction between TV and *FTO* in the British Sample results in a much more statistically significant interaction estimate (0.21%, [0.07%,0.35%], $p = 0.003$), demonstrating that joint interaction modelling can prevent overestimation of interaction effects in the presence of multiple, correlated lifestyle factors.

3.3.2.10 Current smoking

While there is strong evidence that being a current regular smoker is associated with having a lower BMI than otherwise (-1.48%, [-1.57%,-1.39%], $p < 10^{-30}$), there is no evidence that the effect of *FTO* is different between current regular smokers and others (0.03%, [-0.11%,0.16%], $p = 0.69$).

3.3.3 Robustness of interaction effects

3.3.3.1 Confounding with diabetes and depression

To investigate whether our results may have been confounded by associations between BMI and either diabetes or depression, we conducted a sensitivity analysis by removing 12,891 individuals who had reported seeing a psychiatrist for depression and 5,888 individuals who reported having been diagnosed as diabetic. Most estimated interaction effects were effectively unchanged. The largest change was for the interaction with activity score, where the estimate reduced from -0.19% to -0.15% per SD per copy of *FTO*.

3.3.3.2 Reverse causation

Because we analysed one time point, the effects we estimated between BMI and lifestyle variables could have been caused in part by behavioural modifications in response to changes in BMI, that is instances of reverse causation. Assuming additivity, reverse causation would only generate a statistical interaction between a lifestyle factor and *FTO* on BMI if changes in the lifestyle factor in response to BMI depend on *FTO* genotype. We had limited power to address this issue, but there is potential information in participants responses to questions about changes in dietary and alcohol consumption. For example, if the observed interaction between *FTO* and frequency of alcohol consumption is a result of reverse causation, we would expect the interaction effect to be stronger in those who report having changed their alcohol consumption than in those that report no change in alcohol consumption.

To assess whether the interaction between *FTO* and alcohol consumption may reflect reverse causality, we repeated the analysis separately in two subsets of the British Sample: those that answered that their alcohol intake is about the same (n=37,534) as it was ten years ago and those that did not (n=57,193). We found no

significant difference in the estimated interaction effects in the two groups (z=-0.02, $p = 0.49$, one-sided test for stronger interaction in group reporting change in alcohol consumption). We also found no evidence that *FTO* genotype affected the probability of reporting a change in alcohol consumption ($p = 0.87$).

To assess whether the interaction between *FTO* and the diet score may reflect reverse causality, we repeated the analysis separately in two subsets of the British Sample: those that answered no to the question Have you made any major changes to your diet in the last five years? (n=33,781) and those that did not (n=55,675). The observed difference in effect was in the opposite direction to that predicted by reverse causation and was not significant (z=-0.63, $p = 0.73$, one-sided test).

3.3.3.3 Effects of *FTO* on lifestyle variables

The effect of *FTO* on a lifestyle variable can also suggest whether reverse causation may be occurring: if *FTO* affects a lifestyle variable without control for BMI, but *FTO* does not affect it when controlling for BMI, then this indicates that *FTO* may be affecting the lifestyle variable through BMI, an instance of reverse causation. We therefore tested whether *FTO* affected variables that it may interact with: alcohol frequency, squared deviations from mean sleep duration, added salt, and the activity and diet scores (Table 3.4). We regressed these variables onto *FTO*, the genotyping array, and the top 20 principal components in the British Sample. In a second set of analyses to see if *FTO* affected these variables independently of the influence of BMI, we also added log-BMI as a covariate.

	FTO		*FTO* (+BMI)	
	Estimate	p-value	Estimate	p-value
Alcohol score	-0.011	4.6e-02	-0.002	0.68
Activity score	-0.001	7.6e-01	0.009	0.05
Diet score	0.003	6.2e-01	-0.012	0.03
Frequency of added salt	-0.007	2.3e-01	-0.012	0.03
Sleep Squared	-0.008	4.8e-01	-0.017	0.03

Table 3.4: **The effect of *FTO* on selected variables**. Column 1 gives the estimated effect of *FTO* on the variable (expressed as SD change in response per copy of *FTO*), and column 2 gives the assocaited p-value. Columns 3 and 4 give the same when also fitting log-BMI as a covariate.

We found evidence that the *FTO* risk allele reduced the alcohol score ($p = 4.6 \times 10^{-2}$) without control for BMI, but found no evidence for this ($p = 0.68$) with control for BMI. This is consistent with the *FTO* risk allele reducing alcohol consumption as a consequence of increasing BMI, an instance of reverse causation. However, we did not find evidence that reverse causation affected the estimate of the interaction with *FTO* (above).

In contrast, we did not find any strong evidence that *FTO* affects the activity score, diet score, squared deviations from mean sleep duration, or frequency of adding salt to food when not controlling for BMI. When controlling for BMI, however, there is some evidence that *FTO* is associated with these variables (Table 3.4) in the direction that would be expected to decrease BMI according to the main effect of each lifestyle factor on BMI. This would be expected if *FTO* and a lifestyle variable are independent but both affect BMI, which would generate a correlation between *FTO* and the lifestyle variable conditional on BMI.

3.3.3.4 Confounding with overall health

Given that all of the environmental variables we tested for interactions with *FTO* are related to overall health, we attempted to test whether any of these interaction effects were being driven purely by the correlation of the variable with overall health. Again, we did not have perfect information with which to assess this, but we did have self-reported data. Specifically, we considered subjects answers to the question 'In general how would you rate your overall health?' These are encoded as: 1, excellent; 2, good; 3, fair; 4, poor.

It is likely that an individual's self-perception of their overall health is partially determined by their BMI. The correlation between this encoding and log-BMI was 0.27. Regressing overall health on principal components, genotyping array, and *FTO* gives a statistically significant effect of *FTO* ($p = 2 \times 10^{-4}$). However, adding log-BMI to the regression removes the evidence for the effect of *FTO* on overall health ($p = 0.29$), indicating that *FTO* likely effects self-reported overall health through its effect on BMI. We therefore chose not to include self-reported overall health as a covariate in our primary analyses. Doing so would in effect have focused the primary analyses on the component of BMI remaining after regressing out any effect of BMI on self-reported overall health. This would have complicated interpretation, and in addition precluded comparison with other studies based directly on BMI.

By way of a sensitivity analysis and to assess the possibility that our results may be driven by a latent variable related to overall health, we fitted the 'Scores' model (Table 3.1) in the British Sample along with overall health and its interaction with *FTO*. We found no significant evidence that overall health interacts with *FTO* (0.14%, [-0.01%,0.30%], $p = 0.10$). The other estimated interaction effects reduced slightly in magnitude: activity score (-0.19% to -0.16%) alcohol frequency (-0.28% to -0.25%), diet score (0.25% to 0.23%), squared deviations from mean sleep duration (0.13% to 0.11%), and age (-0.16% to -0.14%). We also note, from fitting the 'Diet' model

(Table 3.1) with overall health added, that the estimated interaction with frequency of added salt did not change from 0.23%. In summary, our results do not change much when controlling for overall health, arguing against the possibility that the interactions we report result from an interaction between *FTO* and a latent factor similar to overall health.

3.4 Discussion

We have jointly analysed interactions on BMI between a genetic risk variant in the first intron of the *FTO* gene (rs1421085) and several environmental and lifestyle factors. We undertook these joint analyses separately in two subsets of the UK Biobank data, a large ($n \sim 89,500$) British sample, and a somewhat smaller ($n \sim 29,500$) diverse sample. We found evidence that *FTO* interacts with physical activity, frequency of alcohol consumption, dietary variation, and squared deviations from mean sleep duration (Figures 3.2B, 3.3B, 3.5B, 3.7, and Table 3.3). We did not find statistically significant evidence for interactions with current smoking status, Townsend Deprivation Index, age, and TV watching. As our data relates mainly to individuals of European ancestry living in the United Kingdom, the results may not extend to other populations in different environments, or to children and adolescents.

As previous authors have argued[130], there are major advantages in being able to assess main and interaction effects in the context of joint models which simultaneously include many potential predictors and covariates. Whilst preferable, this approach has often not been possible in many earlier studies, either because a broad set of lifestyle factors have not been measured on study participants, or in the context of meta-analyses, because discretization of factors is essential to their combination, and because individual level data is not available[131, 105]. In our data, for example, we found that testing only one interaction at a time would have led to a large over-

estimation of the interaction between *FTO* and TV watching, potentially leading to a statistically significant result that is not present in the joint model.

Very large resources such as UK Biobank which simultaneously measure extensive genetic, lifestyle, and phenotypic information thus offer substantial promise to further our understanding of gene-by-environment interactions and interactions more generally. In our study, the joint interaction modelling gives us confidence that the interactions we find with physical activity, frequency of alcohol consumption, dietary variation, and squared deviations from mean sleep duration are not due to confounding with each other and with variables correlated with age, socio-economic status (Townsend Deprivation Index), current tobacco smoking, and TV watching. This is an advantage over previous large meta-analyses that have tested only one interaction at a time.

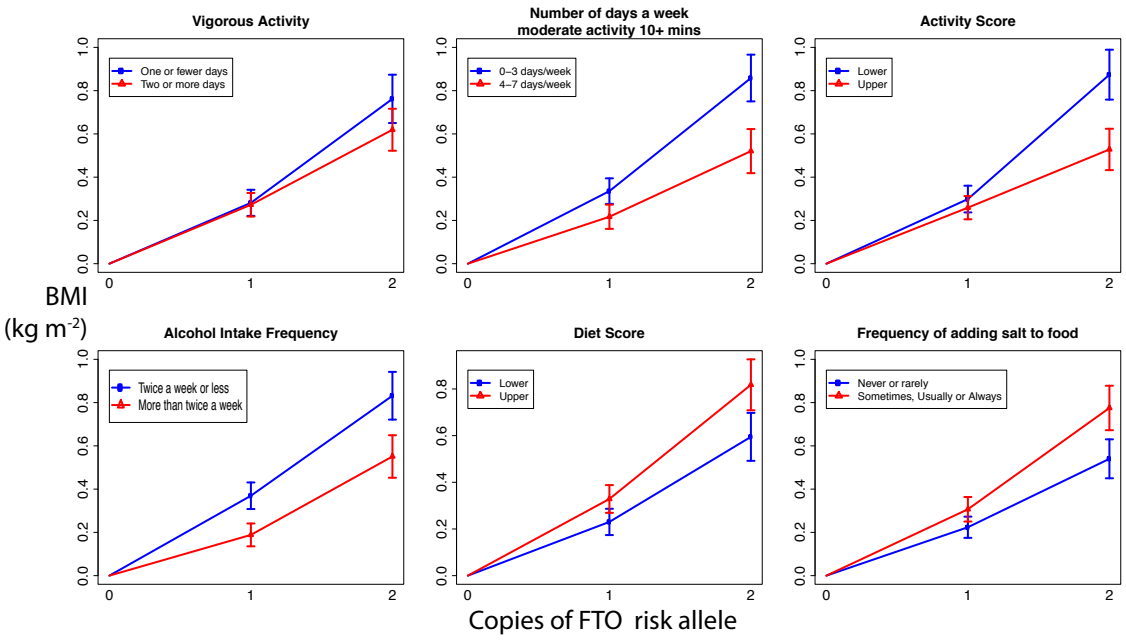


Figure 3.7: **The effect of the *FTO* risk allele for different levels of different lifestyle variables.** In the British subsample, we split each lifestyle variable into two roughly equally sized categories. For each category, we plot the mean BMI and its 95% confidence interval for 1 and 2 copies of the *FTO* risk allele relative to zero copies. If there is no interaction between *FTO* and the environmental variable, the effect of adding another copy of *FTO* should be the same whatever the value of the environmental variables, and the lines for different categories should have the same gradient. If there is an interaction, they should diverge.

All of the diet and lifestyle variables we analyse in the UK Biobank data (but not BMI) are self-reported. While we cannot exclude that self-reporting affected our results, we think this unlikely: while self-reported data may be noisy and biased, assuming additivity, it can only lead to spurious interactions with *FTO* on BMI if self-reporting as a function of BMI depends on *FTO* genotype, a phenomenon which seems a priori unlikely.

The individual components of the activity and diet scores could be viewed as noisy observations of underlying latent activity and diet factors affecting BMI that may interact with *FTO*. Our construction and use of a single summary score from variables in these categories can be seen as a way of estimating these latent factors, which can help overcome the lack of precision in estimates of the individual component measurements.

Evidence for a strong interaction between *FTO* and physical activity has been reported in several US-based studies, with the interaction estimated to be smaller in European cohorts[99]. We found evidence for a stronger interaction between *FTO* and physical activity than suggested by meta-analysis in European cohorts alone, but similar in magnitude to the interaction in another large British cohort (EPIC Norfolk)[99, 100]. We estimated an interaction effect of -0.23% per activity score S.D. per copy of *FTO* risk allele. For BMI of 25kgm^{-2}, this represents a change of 0.41kgm^{-2} per copy of *FTO* for -2 S.D. activity score versus 0.17kgm^{-2} for +2 S.D. activity score, more than a halving of the original *FTO* effect, on this scale.

In both our samples, we saw strong evidence of an interaction on BMI between *FTO* and diet score, which reflects variation in dietary intake of 12 different variables. The combined estimate was 0.30% per S.D. per *FTO* risk allele; for BMI=25kgm^{-2}, this represents a change of 0.44kgm^{-2} per copy of *FTO* for +2 S.D. diet score versus 0.17kgm^{-2} for -2 S.D. diet score, again more than doubling the *FTO* effect. The estimated effect was not stronger in those reporting dietary change in the last five

years, reducing the chance it is due primarily to reverse causation. Our estimates suggest that dietary variation is the strongest statistical modifier of the effect of *FTO* on BMI, out of the variables we investigated, including physical activity.

We found evidence that the effect of *FTO* on BMI is enhanced in those who add salt to food more frequently: combined estimate of 0.21% per S.D. per *FTO* risk allele; for BMI of 25kgm^{-2}, this represents a change of 0.25kgm^{-2} per copy of *FTO* for those who never or rarely add salt versus 0.43kgm^{-2} for those who always add salt. More frequent addition of salt to food is associated with increased energy intake (Figure 3.6). It is plausible that adding salt to energy dense foods increases their palatability and therefore intake, and that this effect may be stronger in *FTO* risk allele carriers and those at risk of obesity[132].

We find strong evidence that more frequent alcohol consumption is associated with decreased BMI in both samples (Figure 3.2A), which is in agreement with previous studies[124, 125], which have also shown a positive association between total alcohol consumption and BMI. Both these findings have been reported repeatedly, and so are likely real, even if, on the surface, they are not easy to reconcile. Data on total alcohol consumption was only available for a small minority of our sample, preventing a joint analysis with frequency of alcohol consumption.

As far as we are aware, the only previous study reporting a gene-alcohol interaction for obesity found evidence that genetic risk for greater central abdominal fat, defined using a twin design, was reduced by greater alcohol consumption within the moderate range ($p < 0.05$)[133]. Here we report evidence that the effect of *FTO* on BMI is reduced in more frequent consumers of alcohol. The combined estimate is -0.24% per S.D. per *FTO* risk allele; for BMI of 25kgm^{-2} this represents a change of 0.33kgm^{-2} per copy of *FTO* for those who drink two to three times a month versus 0.21kgm^{-2} for those who drink daily or almost daily. There was almost no difference in the estimated interaction in those reporting no change in alcohol con-

sumption over the last ten years compared to those reporting a change, increasing confidence that the result is not primarily due to reverse causation. However, we do find evidence that *FTO* reduces alcohol consumption frequency, in agreement with previous studies[126], possibly as a response to increased BMI. We therefore cannot rule out the possibility that the interaction we observe is due to a greater reduction of alcohol consumption frequency in response to higher BMI in *FTO* risk allele carriers compared to non-carriers. Nevertheless, our results highlight the need to further investigate the complex and statistically important relationship between alcohol consumption patterns, BMI, and *FTO*.

The heritability of BMI has been observed to be higher in people who sleep less than seven hours a night compared to those that sleep more than nine hours a night ($p < 0.05$)[134], implying genetic effects on BMI differ depending on sleep. We found evidence that squared deviations from mean sleep duration are associated with an enhanced effect of *FTO* on BMI (combined estimate: 0.13% per S.D. per *FTO* risk allele; for BMI of 25kgm^{-2}, this represents a change of 0.42kgm^{-2} per copy of *FTO* for those who sleep two hours more or less per night than the average versus 0.29kgm^{-2} for the average, 7.16 hours per night). This result implies an interaction that is symmetrical around mean sleep duration, with the influence of *FTO* increasing for both extremes of sleep duration, unlike the heritability result.

In common with other studies, it is possible that all the *FTO*-lifestyle interactions we report are driven by unobserved latent factors with which they are correlated. One specific possibility for such a latent factor is overall health. While we did not have perfect data with which to assess this, we tried to control for it by using a subjective self-reported measure of overall health. The interaction effect estimates reduced only slightly in magnitude when self-reported overall health was included in the model, increasing confidence they are not simply reflecting an interaction with an underlying factor related to overall health. While controlling for factors such as overall health

can increase confidence that an observed gene-lifestyle interaction is not due to a hypothesised confounding, there remains a need for randomization-based methods to demonstrate the causality of *FTO*-lifestyle interactions.

If particular interactions are proven to be causal, it then remains to elucidate the mechanism through which the interaction is mediated. A recent study suggested that the BMI increasing effect of the rs1421085 risk allele is mediated through decreased mitochondrial thermogenesis[94]. The suppressed effect of *FTO* in more physically active individuals could be due to increased thermogenesis as a result of increased exercise. The interaction with frequency of alcohol consumption could be related to increased thermogenesis with increased alcohol consumption[98]. If thermogenesis is the key to *FTO*-lifestyle interactions, then one could predict the interaction effect with lifestyle variables known to affect thermogenesis in a particular direction, such as protein intake[98].

Chapter 4

Heteroskedastic linear mixed models for detecting loci involved in interactions

4.1 Introduction

The problem of searching for interaction effects is harder than for additive effects in part because the number of possible interaction models grows super-linearly with the number of possible interacting variables. Recently it has been recognised that the phenotypic variance differs with genotype at loci involved in interactions[135, 136]. The variance effect of a genotype can be used to screen genome-wide loci for those likely to be involved in interactions, reducing the search space of interaction models. This has been exploited to discover interactions between genetic variants affecting gene expression[72]. The *FTO* locus provides an example of a genetic variant that affects variability in BMI[137] and interacts with lifestyle factors (Chapter 3).

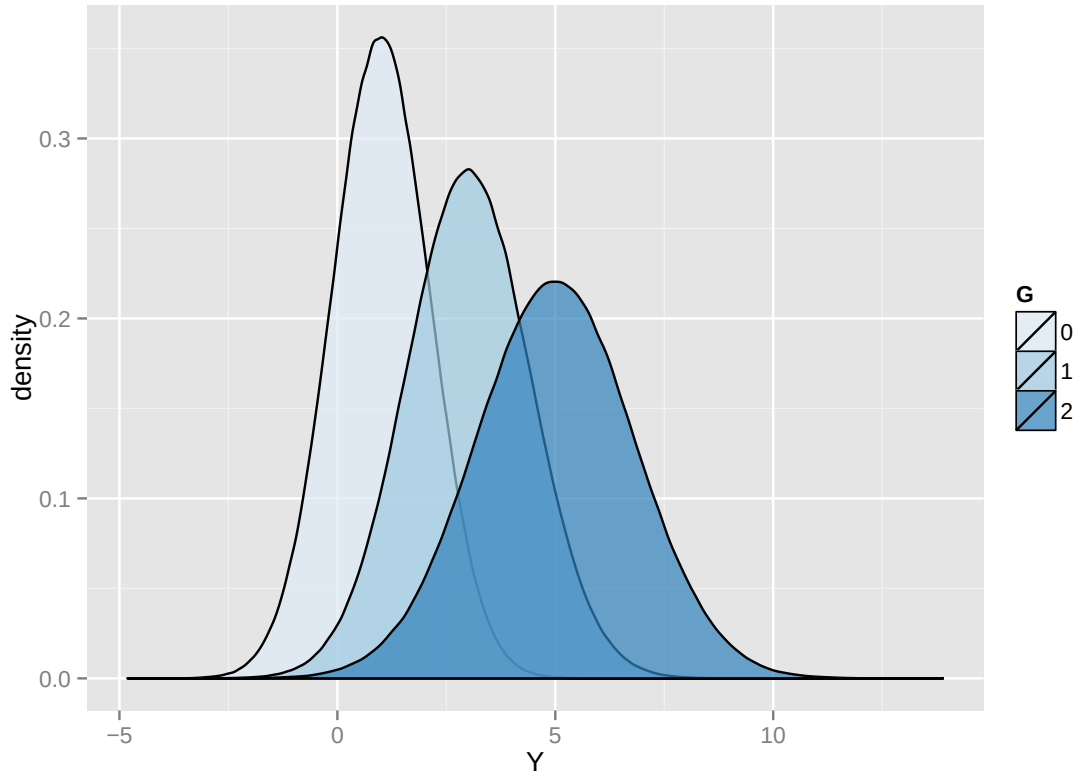


Figure 4.1: **Comparison of the phenotype distributions conditional on the state of a locus that interacts with the environment**.The phenotype is generated by an interaction between a genetic variant, G, and an environmental variable, E, plus some independent noise: $Y = G + E + \gamma(G \times E) + \epsilon$. We plot the distribution of $Y|G = 0, 1, 2$. Both the mean and the variance of Y increase with the number of copies of the G allele at the locus.

To illustrate that a locus that interacts with the environment will generate both mean and variance effects on the phenotype in a marginal model, we use a simple example for a continuous phenotype Y:

$$Y = G + E + \gamma(G \times E) + \epsilon, \tag{4.1}$$

where G is the number of copies of an allele at a locus, E is an environmental effect, and ϵ is independent noise. Because the genetic variant and the environmental variable interact, the sensitivity of the phenotype to variation in the environmental variable is increased with each copy of the allele. This can be seen in the conditional

distribution of $(Y|G = g) = g + (1 + \gamma g)E + \epsilon$, which is plotted for $G = 0, 1, 2$ in Figure 4.1. Both $\text{Var}(Y|G = g)$ and $\mathbb{E}[Y|G = g]$ are functions of g, suggesting that a test statistic for both mean and variance effects would be more powerful than an additive test alone.

Most methods that look at variance effects of loci have concentrated on testing for a variance effect alone[138, 139], even though it is unlikely that interacting loci have no marginal effects. While joint tests for mean and variance effects have been proposed[140, 141], the methods are not as well developed as those for additive association testing. In particular, the ability of linear mixed models to reduce confounding and increase power[41] has not been exploited in large datasets, with the only previous such method targeted to deal with moderately sized family pedigree structures[141].

We introduce the heteroskedastic linear mixed model (HLMM), which is flexible enough to model arbitrary mean and variance effects of a locus, the influence of known covariates and principal components on the mean and variance of a trait, while including the additive effects of many genetic variants as random effects. We have developed an efficient algorithm for fitting this model whose computations scale linearly with the sample size for a fixed number of loci in the random effect, allowing us to apply it to UK Biobank samples of over 100,000 individuals. First, we derive the test statistics that we aim to employ within the HLMM.

4.2 Test statistics for mean and variance effects

Assuming the phenotype (Y) distribution is normal conditional on the genotype (G), the most general model relating genotype to phenotype allows for the distribution conditional on each genotype at the locus to be any normal distribution:

$$\text{M}_\text{G}: \; Y|G = g \sim \mathcal{N}(\mu_g, \sigma_g^2). \tag{4.2}$$

To test for association between genotype and phenotype, one could compare the likelihood of model M_G to the null model,

$$M_0 : \ Y|G = g \sim \mathcal{N}(\mu, \sigma^2). \tag{4.3}$$

giving a likelihood ratio test on four degrees-of-freedom, which has been previously suggested[140].

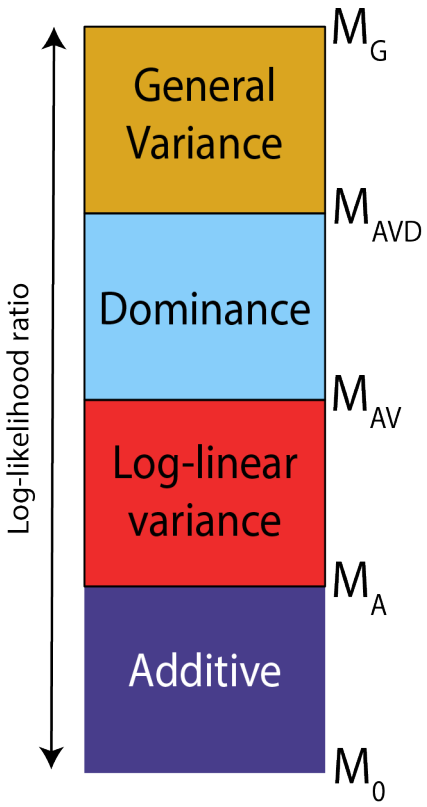


Figure 4.2: **Nested hierarchy of models**. The hierarchy builds from the null model (M_0 – no mean or variance effects) to the general model (M_G – arbitrary mean and variance effects). Effects are added successively at each level of the hierarchy (additive, log-linear variance, dominance, and general variance) with the model including all of the effects below it indicated on the right hand side. The overall height of the bar can be seen as the log-likelihood ratio test statistic comparing the general model (M_G) to the null model (M_0), with the heights of the components giving the corresponding log-likelihood ratio test statistics for the specified effects.

While model M_G can capture any mean and variance effects of a locus, it is possible to fit simpler models that capture mean and variance effects. We introduce a nested

hierarchy of models (Figure 4.2) that allows us to decompose the log-likelihood ratio of model $\mathrm{M_G}$ to the null model into components that give evidence for different mean and variance effects. From this hierarchy, simpler tests can be devised to improve power.

Most genetic association studies fit a model that only allows for an additive effect. The first model above the null model in our hierarchy is the additive model:

$$\mathrm{M_A}: \; Y|G = g \sim \mathcal{N}(\mu + \alpha g, \sigma^2), \tag{4.4}$$

where μ is the location parameter for the mean, and α is the additive effect of the genetic variant.

We now seek to introduce a variance effect that is analogous to the additive effect on the mean. Because the variance is always positive, one cannot use a linear model, which is unbounded; instead, we use a log-linear model. Let $\sigma_g^2 = \mathrm{Var}(Y|G = g)$. This model has the form

$$\log(\sigma_g^2) = \mu_v + \alpha_v g, \tag{4.5}$$

where μ_v corresponds to the scale of the variance, and α_v is termed the log-linear variance effect of the locus. The next model in our hierarchy incorporates a log-linear variance effect in addition to an additive effect:

$$\mathrm{M_{AV}}: \; Y|G = g \sim \mathcal{N}(\mu + \alpha g, \exp(\mu_v + \alpha_v g)), \tag{4.6}$$

which we call the additive-variance model or AV model for short.

We add a dominance effect to this model to allow for non-linearity in the relationship between the conditional means and the number of copies of an allele, giving:

$$\mathrm{M_{AVD}}: \; Y|G = g \sim \mathcal{N}(\mu_g, \exp(\mu_v + \alpha_v g)). \tag{4.7}$$

Similarly, we add a general variance effect to allow for non-linearity in the relationship between the conditional log-variances and the number of copies of an allele, which takes us to $\mathrm{M_G}$. The log-likelihood ratio between $\mathrm{M_G}$ and $\mathrm{M_0}$ can therefore be decomposed as the sum of the log-likelihood ratio test statistics for each of the mean and variance effects:

$$2[l(\mathrm{M_G}|y,g) - l(\mathrm{M_0}|y,g)] = 2[l(\mathrm{M_G}|y,g) - l(\mathrm{M_{AVD}}|y,g)] + \tag{4.8}$$
$$2[l(\mathrm{M_{AVD}}|y,g) - l(\mathrm{M_{AV}}|y,g)] + 2[l(\mathrm{M_{AV}}|y,g) - l(\mathrm{M_A}|y,g)] + 2[l(\mathrm{M_A}|y,g) - l(\mathrm{M_0}|y,g)].$$

The four components individually give evidence for additive ($\mathrm{M_A}$ vs. $\mathrm{M_0}$), log-linear variance ($\mathrm{M_{AV}}$ vs. $\mathrm{M_A}$), dominance ($\mathrm{M_{AVD}}$ vs. $\mathrm{M_{AV}}$), and general variance ($\mathrm{M_G}$ vs. $\mathrm{M_{AVD}}$) effects, which is illustrated in Figure 4.2.

4.2.1 Relation to mutual information

The mutual information between two random variables is a general measure of their dependence which is zero if and only if the two variables are independent, unlike linear correlation. It measures the amount of information that is shared between observations of the variables. The mutual information between a continuous phenotype Y and a genetic variant G is

$$I(Y;G) = H(Y) - H(Y|G); \tag{4.9}$$

where $H(Y)$ is the differential entropy of the phenotype Y,

$$H(Y) = -\int_{-\infty}^{\infty} f(y)\log(f(y))dy, \tag{4.10}$$

where $f(y)$ is the density function of the phenotype; and $H(Y|G)$ is the conditional entropy of Y given G,

$$H(Y|G) = \mathbb{E}_G[H(Y|G = g)], \tag{4.11}$$

where $H(Y|G = g)$ is the entropy of Y given that G takes a particular value, g.

We show that, in the infinitesimal genetic model, the likelihood ratio test statistic comparing $\mathrm{M_G}$ to $\mathrm{M_0}$ at the maximum likelihood parameter estimates is an estimator of the mutual information between Y and G, $I(Y;G)$.

4.2.1.1 General likelihood ratio test statistic

To derive the maximum likelihood of the data under $\mathrm{M_G}$, we parameterise the model as:

$$Y|G = g \sim \mathcal{N}(\mu_g, \sigma_g^2), \tag{4.12}$$

where $\mu_g = \mathbb{E}[Y|G = g]$ and $\sigma_g^2 = \mathrm{Var}(Y|G = g)$.

If there are n_g out of n genotypes in category g, and y_{gi} is the i^{th} phenotypic observation in category g, then

$$2l(\mathrm{M_G}|y, g) = -n\ln(2\pi) - \sum_{g=0}^{2} n_g \ln(\sigma_g^2) - \sum_{g=0}^{2}\sum_{i=1}^{n_g} \frac{(y_{gi} - \mu_g)^2}{\sigma_g^2}. \tag{4.13}$$

This implies that the maximum likelihood estimators are

$$\hat{\mu}_g = \frac{1}{n_g}\sum_{i=1}^{n_g} y_{gi};\ \hat{\sigma}_g^2 = \frac{1}{n_g}\sum_{i=1}^{n_g}(y_{gi} - \hat{\mu}_g)^2. \tag{4.14}$$

Let $\hat{l}$ be the value of the log-likelihood evaluated at the maximum likelihood estimator, then

$$2\hat{l}(\mathrm{M_G}|y, g) = -n\ln(2\pi) - \sum_{g=0}^{2} n_g \ln(\hat{\sigma}_g^2) - n. \tag{4.15}$$

For the null model, where $\bar{y}$ is the overall sample phenotype mean,

$$2\hat{l}(\mathrm{M}_0|y,g) = -n\ln(2\pi) - n\ln(\hat{\sigma}^2) - n; \hat{\sigma}^2 = \frac{1}{n}\sum_{g=0}^{2}\sum_{i=1}^{n_g}(y_{gi} - \bar{y})^2. \tag{4.16}$$

Therefore,

$$2[\hat{l}(\mathrm{M}_\mathrm{G}|y,g) - \hat{l}(\mathrm{M}_0|y,g)] = n\ln(\hat{\sigma}^2) - \sum_{g=0}^{2} n_g \ln(\hat{\sigma}_g^2). \tag{4.17}$$

4.2.1.2 Maximum likelihood estimator of the mutual information

We now derive the mutual information between a genotype and a continuous phenotype in the infinitesimal genetic model with Gaussian error[142]. In the infinitesimal model, the unconditional distribution of Y is normal:

$$Y \sim \mathcal{N}(\mu, \sigma^2). \tag{4.18}$$

The differential entropy of Y is therefore $H(Y) = 0.5\ln(2\pi e\sigma^2)$. The mutual information between Y and G, $I(Y;G)$, can be expressed as

$$I(Y;G) = H(Y) - \mathbb{E}_G[H(Y|G=g)]. \tag{4.19}$$

Under the infinitesimal genetic model, the conditional distribution $Y|G=g$ is also normal with variance σ_g^2. Therefore

$$\mathbb{E}_G[H(Y|G=g)] = 0.5\mathbb{E}_G[\ln(2\pi e\sigma_g^2)]. \tag{4.20}$$

The mutual information is therefore

$$I(Y;G) = H(Y) - H(Y|G) = 0.5\ln(2\pi e\sigma^2) - 0.5\mathbb{E}_G[\ln(2\pi e\sigma_g^2)]; \tag{4.21}$$

$$= 0.5\ln(\sigma^2) - 0.5\sum_{g=0}^{2}\mathbb{P}(G=g)\ln(\sigma_g^2). \tag{4.22}$$

If we estimate the mutual information with the maximum likelihood estimators of the parameters, this gives

$$2n\hat{I}(Y;G) = n\ln(\hat{\sigma}^2) - \sum_{g=0}^{2} n_g \ln(\hat{\sigma}_g^2) = 2[\hat{l}(\mathrm{M_G}|y,g) - \hat{l}(\mathrm{M_0}|y,g)]. \tag{4.23}$$

We have shown that the maximum likelihood estimator of the mutual information between a genotype and a phenotype in the infinitesimal genetic model is proportional to a general likelihood ratio test for dependence, which is on four degrees of freedom. Therefore, the asymptotic distribution of the maximum likelihood estimator of the mutual information between genotype and phenotype is $(2n)^{-1}\chi_4^2$. This can be seen as a case of the known relationship between mutual information and log-likelihood ratio test statistics in parametric models[143].

The mutual information between a genotype and phenotype is zero if and only if they are independent. We have therefore shown that, under the infinitesimal genetic model with Gaussian residual error, the likelihood ratio test comparing $\mathrm{M_G}$ to $\mathrm{M_0}$ will, for all fixed significance levels greater than zero and less than one, have power to detect an association that tends to 100% with sample size if and only if Y and G are dependent.

4.3 The Heteroskedastic Linear Model

All of the models in the above hierarchy can be incorporated into a class of models called heteroskedastic linear models, which allow for an arbitrary vector of covariates

to influence the residual variance of the response. Similar models and algorithms have a long history in the fields of heteroskedastic regression models[144] and econometrics.

Consider a phenotype Y with multivariate normal distribution:

$$Y \sim \mathcal{N}(X\alpha, D), \tag{4.24}$$

for some diagonal matrix D. A natural and simple way to model heteroskedasticity is to use a log-linear model. We can thereby model the diagonal elements of D as

$$D_{ii} = \exp(V_i\beta), \ \rightarrow \ \log(D_{ii}) = V_i\beta, \ i = 1, \ldots, n; \tag{4.25}$$

where V_i is a vector of v covariates measured for observation i, and β is a $[v \times 1]$ vector of coefficients which models the linear change in the log-residual-variance with that covariate vector. We can arrange the vectors of covariates V_i, $i = 1, \ldots, n$, into a design matrix for the residual variance, V, of dimension $[n \times v]$. We can then express D as

$$D = \exp(\text{diag}(V\beta)), \tag{4.26}$$

where $\text{diag}(V\beta)$ is the diagonal matrix with diagonal entry i equal to $V_i\beta$, and $\exp(\text{diag}(V\beta))$ is the matrix exponential of the diagonal matrix $\text{diag}(V\beta)$. A column of $1's$ models the scale of the residual variance. Alternatively, without a column of $1's$ in V, one could express D as

$$D = \sigma^2 \exp(\text{diag}(V\beta)), \tag{4.27}$$

which makes clear the effect of changing an element of V is to scale the residual variance up or down by some factor that depends on β.

The heteroskedastic linear model is therefore

$$Y \sim \mathcal{N}(X\alpha, \exp(\text{diag}(V\beta))). \tag{4.28}$$

4.3.1 Inference algorithm

We give the inference steps first, referencing the relevant equations where necessary, with detailed derivations in the relevant subsections. The approach we take is to optimise over the profile likelihood, $L_{\text{prof}}(\beta) = L(\hat{\alpha}_\beta, \beta)$, where $\hat{\alpha}_\beta$ is the value of α that maximises the likelihood for a particular β, the solution to (4.34).

1: $\alpha_{\text{OLS}} = (X^T X)^{-1} X^T y$. {Initialise α}

2: set $\hat{\beta}_*$ to the solution to (4.37) with $\alpha = \alpha_{\text{OLS}}$. {Initialise β}

3: Find $\hat{\beta} = \underset{\beta}{\text{argmax}}\, L_{\text{prof}}(\beta)$ by Newton's algorithm using $\hat{\beta}_*$ as the initial value.

4: Find $\hat{\alpha}$ as the solution to (4.34) for $\beta = \hat{\beta}$.

5: Compute the inverse Fisher Information Matrix (4.45) at $(\alpha, \beta) = (\hat{\alpha}, \hat{\beta})$ to obtain standard error estimates.

The gradients and Hessian used are derived below.

4.3.2 Likelihood

For convenience, instead of the full likelihood, we work with

$$L(\alpha, \beta | y, X, V) = 2 \log f(y | X, V, \alpha, \beta) + n \log(2\pi), \tag{4.29}$$

where $f(y|X, V, \alpha, \beta)$ is the multivariate normal density of the heteroskedastic linear model at y given X, V, α, β. Therefore, if y_i is the i^{th} observation of the phenotype and X_i is the i^{th} row of X,

$$L = L(\alpha, \beta | y, X, V) = -\sum_{i=1}^{n} V_i \beta - \sum_{i=1}^{n} (y_i - X_i \alpha)^2 \exp(-V_i \beta). \tag{4.30}$$

4.3.3 Gradient

For notation for the first derivative of a scalar function L of a $[k \times 1]$ vector x, we express the first (partial) derivative in terms of x^T:

$$\frac{\partial L}{\partial x^T} = \left[\frac{\partial L}{\partial x_1}, \ldots, \frac{\partial L}{\partial x_k}\right] \text{ is } [1 \times k]. \tag{4.31}$$

This has the advantage of writing the linear approximation from the Taylor series of the scalar function as

$$L(x) \approx L(x_0) + \frac{\partial L}{\partial x^T}(x - x_0). \tag{4.32}$$

4.3.3.1 With respect to mean effects

$$\frac{\partial L}{\partial \alpha^T} = 2(y - X\alpha)^T D^{-1} X \tag{4.33}$$

This implies the MLE for α, $\hat{\alpha}$ must satisfy the linear system:

$$X^T D^{-1} X \hat{\alpha} = X^T D^{-1} y, \tag{4.34}$$

for a given β.

4.3.3.2 With respect to variance effects

$$\frac{\partial L}{\partial \beta^T} = \sum_{i=1}^{n} ((y_i - X_i\alpha)^2 \exp(-V_i\beta) - 1)V_i \tag{4.35}$$

If we assume that $|V_i\beta|$ is small $\forall i$, then

$$\frac{\partial L}{\partial \beta^T} \approx \sum_{i=1}^{n} [(y_i - X_i\alpha)^2(1 - V_i\beta) - 1]V_i \tag{4.36}$$

If we set this approximation to zero, we can solve a linear system for an approximate MLE for β, $\hat{\beta}_*$:

$$V^T \text{diag}(e^2) V \hat{\beta}_* = \sum_{i=1}^{n} [(y_i - X_i\alpha)^2 - 1]V_i, \tag{4.37}$$

where e^2 is the element-wise square of the residuals: $e_i^2 = (y_i - X_i\alpha)^2$.

4.3.4 Second derivative and asymptotic covariance

The second derivative of L with respect to α is

$$\frac{\partial^2 L}{\partial\alpha\partial\alpha^T} = -2X^T D^{-1} X. \tag{4.38}$$

With respect to β, it is

$$\frac{\partial^2 L}{\partial\beta\partial\beta^T} = -\sum_{i=1}^{n}(y_i - X_i\alpha)^2 \exp(-V_i\beta)V_i^T V_i \tag{4.39}$$

$$= -V^T \delta V, \tag{4.40}$$

where δ is a diagonal matrix with diagonal element $\delta_{ii} = (y_i - X_i\alpha)^2 \exp(-V_i\beta)$. Delta is positive semi-definite $\forall\beta$ as $\delta_{ii} \geq 0\ \forall\beta$, implying that $-L$ is a convex function of β.

To complete the Hessian of the log-likelihood, we need

$$\frac{\partial^2 L}{\partial\alpha\partial\beta^T} = -2\sum_{i=1}^{n}(y_i - X_i\alpha)\exp(-V_i\beta)X_i^T V_i \tag{4.41}$$

$$= -2X^T \mathrm{diag}((y - X\alpha) \circ \exp(-V\beta))V, \tag{4.42}$$

where $(y - X\alpha) \circ \exp(-V\beta)$ is the element-wise product of $(y - X\alpha)$ and $\exp(-V\beta)$.

Therefore the Hessian of the log-likelihood with respect to (α, β) is

$$H = \begin{bmatrix} -X^T D^{-1} X & -X^T \mathrm{diag}((y - X\alpha) \circ \exp(-V\beta))V \\ -V^T \mathrm{diag}((y - X\alpha) \circ \exp(-V\beta))X^T & -V^T \delta V/2, \end{bmatrix} \tag{4.43}$$

where we have divided by 2 to make it correspond to the true log-likelihood, not L.

The negative expectation of the Hessian, the Fisher Information Matrix, is there-

fore

$$-\mathbb{E}[H] = \begin{bmatrix} X^T D^{-1} X & 0 \\ 0 & V^T V/2 \end{bmatrix}, \tag{4.44}$$

because $\mathbb{E}[\delta_{ii}] = \exp(-V_i\beta)\mathbb{E}[(y_i - X_i\alpha)^2] = 1 \; \forall \; i$, and $\mathbb{E}[(y - X\alpha)] = 0$. Therefore the inverse of the information matrix is

$$I((\alpha, \beta))^{-1} = \begin{bmatrix} (X^T D^{-1} X)^{-1} & 0 \\ 0 & 2(V^T V)^{-1} \end{bmatrix}. \tag{4.45}$$

This matrix will be invertible as long as X and V are of full column rank, which is also enough to ensure that the negative log-likelihood is asymptotically strictly convex, so that the maximum likelihood solutions are unique. Therefore, the asymptotic covariance of the maximum likelihood estimator of (α, β) is given by the inverse Fisher Information Matrix if X and V are of full column rank.

4.4 The heteroskedastic linear mixed model

We first consider a linear mixed model which allows for heteroskedasticity in both the random effects and the residual error:

$$Y = X\alpha + Z\gamma + \epsilon; \tag{4.46}$$

where X is the $[n \times c]$ design matrix for the fixed effects, α; Z is the $[n \times l]$ design matrix for the random effects, γ; and ϵ is the residual error vector. We define the covariance matrices:

$$H = \text{Cov}(\gamma); \text{ and } D = \text{Cov}(\epsilon). \tag{4.47}$$

We are interested in modelling the heteroskedasticity in both the random effects and the residual error. We model the heteroskedasticity in the residual error as in

the previous section:

$$D = \exp(\text{diag}(V\beta)). \tag{4.48}$$

The l random effects will in general have different variances, and the difference in variance between different random effects may depend on known covariates. If the random effects represent allelic substitution effects on a phenotype, we might expect non-synonomous coding variants to contribute more to the phenotypic variance than synonomous coding variants. In random effects models, heteroskedasticity is usually modelled by considering a partition of the l random effects into k discrete categories, with each random effect in each category having equal variance. This results in

$$ZHZ^T = \sum_{j=1}^{k} \sigma_j^2 Z_j Z_j^T. \tag{4.49}$$

While this can model heteroskedasticity coming from discrete, non-overlapping categories, it cannot model heteroskedasticity that follows continuous variables or multiple, overlapping variables.

The log-linear variance model offers greater flexibility in modelling the heteroskedasticity in the random effects. We consider uncorrelated random effects, so that H is diagonal:

$$H = \exp(\text{diag}(W\lambda)), \tag{4.50}$$

where W is a $[l \times w]$ design matrix for the log-variance of the random effects, with coefficient vector λ.

One disadvantage of this model is that it becomes impossible to test the hypothesis that a particular category of random effects contributes nothing to the variance, as a zero contribution to the variance corresponds to a coefficient in λ of negative infinity. If the random effects represent different allele substitution effects, however, this may not matter, as all variants can be expected to 'contribute' a small amount to the

phenotypic variance due to population stratification and confounding with shared environment. The interpretation of the coefficients in λ for particular covariates then becomes a variance contribution above or below the background level, which may be a more meaningful question than whether there is any contribution above zero.

Assuming that the random effects and the residual error are Gaussian, this gives

$$Y|X, Z, \alpha, \beta, \lambda \sim \mathcal{N}(X\alpha, ZHZ^T + D); \tag{4.51}$$

$$D = \exp(\text{diag}(V\beta));\ H = \exp(\text{diag}(W\lambda)). \tag{4.52}$$

In the empirical analyses and software implementation, we consider a simplification of this with $H = h^2 I$.

4.5 Efficient inference for the low-rank heteroskedastic linear mixed model

4.5.1 Algorithm overview

We implement the algorithm in Python using NumPy for linear algebra operations. To deal with missing data, we analyse only those individuals with complete observations of all the model variables. We note that our approach has similarities to computational approaches previously used in general linear mixed models[145].

Let $\theta = (\beta, h^2)$ be the vector of variance parameters of the simplified model with $H = h^2 I$. To fit the simplified model, we optimise over the profile likelihood, $L_{\text{prof}}(\theta) = L(\hat{\alpha}_\theta, \theta)$, where $\hat{\alpha}_\theta$ is the value of α that maximises the likelihood for a particular θ, the solution to (4.67).

1: Input an initial guess for h^2, h_0^2. {Initialise h^2}

2: Find $\hat{\beta}_{HLM}$, the maximum likelihood estimate of β in the model without the

random effects, by application of algorithm in Section 4.3.1. {Initialise β}

3: Initialise θ as $\theta_0 = (\hat{\beta}_{HLM}, h_0^2)$.

4: Use the L-BFGS-B algorithm to find the $\hat{\theta}$ that maximises $L_{\text{prof}}(\theta)$, using θ_0 as the initial value. The likelihood and its gradient are computed using the expressions derived below.

5: Find $\hat{\alpha}$, the α that maximises $L(\alpha, \hat{\theta})$.

6: Estimate standard errors from the negative inverse of a numerical approximation to the Hessian of the log-likelihood at $(\hat{\alpha}, \hat{\beta}, \hat{h}^2)$.

For each chromosome, we first fit a null model using a user input initial guess for h^2, and we use the resulting maximum likelihood estimate for h^2 as the initial guess for h^2 for all locus specific models. Because different loci in general have different sets of individuals with missing calls, we fit a locus specific null model first at each locus.

4.5.1.1 Overall complexity

The overall time complexity for the computation of the likelihood and gradients is

$$\mathrm{O}(nl^2 + l^3 + ncl + cl^2 + nc^2 + c^3 + nv + lw). \tag{4.53}$$

The overall space complexity is

$$\mathrm{O}(nl + l^2 + nc + c^2 + nv + lw). \tag{4.54}$$

Both the time and space complexity are linear in n when the other parameters are fixed.

4.5.2 Computation of the likelihood in O(n) Operations

As D is not proportional to the identity matrix, a rotation defined by the eigenvectors of Z does not diagonalise the system. However, the likelihood and its derivative can still be computed in O(n) by taking advantage of the structure of the covariance of Y, which is a diagonal matrix plus a low rank matrix.

Let

$$\mathrm{Cov}(Y) = \Sigma = ZHZ^T + D, \tag{4.55}$$

then the log-likelihood is

$$l = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}(y - X\alpha)^T\Sigma^{-1}(y - X\alpha) \tag{4.56}$$

Instead of maximising l, we equivalently maximise $L = 2l + n\log(2\pi)$:

$$L = -\ln|\Sigma| - (y - X\alpha)^T\Sigma^{-1}(y - X\alpha). \tag{4.57}$$

To naively compute the likelihood, one needs the inverse of Σ, computation of which requires O(n^3) operations. By application of the Woodbury Matrix Identity, the inverse of Σ can be reduced to the inverse of D, which is diagonal, plus a low rank correction:

$$\Sigma^{-1} = D^{-1} - D^{-1}Z(H^{-1} + Z^TD^{-1}Z)^{-1}Z^TD^{-1}. \tag{4.58}$$

Let $\Lambda = H^{-1} + Z^TD^{-1}Z$, then we also have, by the Matrix Determinant Lemma,

$$\log|\Sigma| = \log|\Lambda| + \log|H| + \log|D| \tag{4.59}$$

$$\log|\Sigma| = \log|\Lambda| + \sum_{j=1}^{l} W_j\lambda + \sum_{i=1}^{n} V_i\beta. \tag{4.60}$$

Therefore,

$$L = -\sum_{i=1}^{n} V_i\beta - \sum_{i=1}^{n}(y_i - X_i\alpha)^2 \exp(-V_i\beta) - \sum_{j=1}^{l} W_j\lambda \tag{4.61}$$

$$- \log|\Lambda| + [Z^T D^{-1}(y - X\alpha)]^T \Lambda^{-1} [Z^T D^{-1}(y - X\alpha)] \tag{4.62}$$

This can be computed in $\mathrm{O}(nl^2 + l^3)$ operations by precomputing the $[l \times 1]$ vector

$$r = Z^T D^{-1}(y - X\alpha). \tag{4.63}$$

The likelihood can thereby be expressed as

$$L = -\sum_{i=1}^{n} V_i\beta - \sum_{i=1}^{n}(y_i - X_i\alpha)^2 \exp(-V_i\beta) - \tag{4.64}$$
$$\sum_{j=1}^{l} W_j\lambda - \log|\Lambda| + r^T\Lambda^{-1}r,$$

where the first line is the likelihood for the diagonal system without the random effect, and the second line is the contribution to the likelihood of the random effects. The computation is dominated by calculation of Λ in $\mathrm{O}(nl^2)$ operations and its inverse and determinant in $O(l^3)$ operations.

4.5.3 Efficient computation of the maximum likelihood estimator of the fixed effects

The derivative of the log likelihood with respect to α is

$$\frac{\partial L}{\partial \alpha^T} = 2(y - X\alpha)^T D^{-1} X - 2(y - X\alpha)^T D^{-1} Z \Lambda^{-1} Z^T D^{-1} X \tag{4.65}$$

$$= 2(y - X\alpha)^T [D^{-1} X - D^{-1} Z \Lambda^{-1} Z^T D^{-1} X] \tag{4.66}$$

To find the MLE, we equate the derivative to zero and solve for α. We get that the MLE for α, $\hat{\alpha}$ must satisfy the linear system:

$$[X^T D^{-1} X - X^T D^{-1} Z \Lambda^{-1} Z^T D^{-1} X]\hat{\alpha} = X^T D^{-1} y - X^T D^{-1} Z \Lambda^{-1} Z^T D^{-1} y \quad (4.67)$$

$X^T D^{-1} X$ can be computed in $\mathrm{O}(nc^2)$ operations; $X^T D^{-1} Z$ is a $[c \times l]$ matrix which can be computed in $\mathrm{O}(ncl)$ operations; so assuming that Λ^{-1} has already been computed, and that $[X^T D^{-1} X - X^T D^{-1} Z \Lambda^{-1} Z^T D^{-1} X]$ is full rank, $\hat{\alpha}$ can be computed in $\mathrm{O}(nc^2 + ncl + cl^2 + c^3)$ operations.

4.5.4 Derivative with respect to variance parameters

In Appendix B, we derive the derivatives of the likelihood with respect to the variance parameters. We give the results here.

4.5.4.1 Derivative with respect to λ

$$\frac{\partial L}{\partial \lambda^T} = \sum_{j=1}^{l} [(\Lambda_{jj}^{-1} + \Gamma_{jj}) \exp(-W_j \lambda) - 1] W_j; \ \Gamma = \Lambda^{-1} r r^T \Lambda^{-1}. \quad (4.68)$$

Computing this gradient requires computation of Γ in $\mathrm{O}(l^3)$ operations, then an $O(lw)$ operation. It is therefore linear in the number of heteroskedasticity parameters for the random effect, w.

4.5.4.2 Derivative with respect to β

$$\frac{\partial L}{\partial \beta^T} = \sum_{i=1}^{n} (k_i \exp(-V_i \beta) - 1) V_i, \quad (4.69)$$

where k is a function of Λ, X, Z, D, and the residuals. To compute k requires $\mathrm{O}(nl^2)$ operations, then to complete the gradient computation requires an $\mathrm{O}(nv)$ operation, so the gradient computation is linear in the number of log-linear variance parameters.

Chapter 5

Genome-wide association analysis of body mass index with the heteroskedastic linear mixed model

5.1 Introduction

Previous work has indicated that gene-by-environment interactions may be important in explaining BMI variation (Introduction and Chapter 3). We therefore chose to analyse BMI in the subsample of the UK Biobank with predominantly European ancestry ($n \sim 142,000$) as a test case for application of the Heteroskedastic Linear Mixed Model (HLMM). Our analyses led to the discovery of eight novel BMI associations that fit a model with additive and variance effects better than a model with only additive effects, three of which would not have been discovered on the basis of an additive test alone (Figure 5.6 and Tables C.1, C.2, C.3, C.4, and C.5). Additionally, we found that five previously known loci also fit a model with additive and variance effects better than a model with only additive effects. After identifying a locus that is likely to be involved in interactions with the HLMM, one can test whether it inter-

acts with particular environmental or genetic variables. We did this for the type-II diabetes risk variant (rs7903146) at the *TCF7L2* locus, for which we discovered novel interactions with insulin treatment and age on BMI.

We anticipate based on simulations that many more possibly interacting loci with weak additive but moderate to strong variance effects remain to be discovered at larger sample sizes, and that insights into the importance of non-additivity for different traits can be obtained from examining the genome-wide test statistics.

5.2 Results

5.2.1 Choosing the most powerful test to detect loci involved in interactions

It has been shown that modeling dominance effects for common SNPs leads to little increase in variance explained for many traits, including BMI[146]. It is therefore unlikely that including a test for a dominance effect will increase power to detect loci involved in interactions.

If the variance effects of loci involved in interactions follow an approximate log-linear form, then a test that includes the log-linear variance effect but not the general variance effect should be more powerful. We now show that this is the case in a simple interaction model between the additive effect of a genetic variant G and an environmental variable E (Figure 4.1), although the same arguments would apply to interactions with a genetic variant. The model for the phenotype, Y, is

$$Y = G + E + \gamma GE + \epsilon, \tag{5.1}$$

where G is the number of copies of an allele at a locus, E is an environmental variable, and ϵ is independent noise with variance σ^2_ϵ.

The variance conditional on $G = g$ is

$$\begin{aligned}
\mathrm{Var}(Y|G=g) &= \sigma_\epsilon^2 + (1+\gamma g)^2 \mathrm{Var}(E) && (5.2)\\
&= (\sigma_\epsilon^2 + \mathrm{Var}(E)) + 2\mathrm{Var}(E)\gamma g + \mathrm{Var}(E)\gamma^2 g^2 && (5.3)\\
&= (\sigma_\epsilon^2 + \mathrm{Var}(E)) + 2\mathrm{Var}(E)\gamma g + \mathrm{O}(\gamma^2 g^2). && (5.4)
\end{aligned}$$

Therefore the conditional variance is a linear function of g up to a correction factor that is proportional to the square of the interaction effect size. Given that the effect sizes of common variants for complex traits in humans are generally small relative to the variance of the trait, the quadratic term is generally going to be too small to detect at current sample sizes. This also applies to the log-conditional variance. If we assume that $(\sigma_\epsilon^2 + \mathrm{Var}(E)) = 1$, then

$$\begin{aligned}
\log(\mathrm{Var}(Y|G=g)) &= \log(1 + 2\mathrm{Var}(E)\gamma g + \mathrm{O}(\gamma^2 g^2)) && (5.5)\\
&= 2\mathrm{Var}(E)\gamma g + \mathrm{O}(\gamma^2 g^2). && (5.6)
\end{aligned}$$

This implies that for the effect sizes of common loci on complex traits, a log-linear variance model should be accurate, unless the interaction model involves strongly non-linear functions of the genotype.

We therefore propose a test for powerful discovery of loci involved in interactions that compares the likelihood of the additive-variance model, $\mathrm{M_{AV}}$, to the likelihood of the null model. We call this the additive-variance test – or 'AV' test for short. This test statistic can be decomposed into test statistics for additive and log-linear variance effects.

5.2.2 Simulations

5.2.2.1 Power of the additive-variance test

To test the power of the AV test relative to the additive test, we simulated phenotypes according to the model

$$Y \sim \mathcal{N}(0.02 * g, \exp(\beta g)), \tag{5.7}$$

with β varying from 0 to 0.05 in increments of 0.005. The test locus, g, was simulated as a Binomial(2,0.5) random variable. The mean effect was chosen to be relatively weak to simulate the kinds of loci that may have been missed by previous genome wide association studies based on additive tests in large sample sizes. We tested sample sizes from 10,000 to 100,000, using 200 simulated phenotypes for each set of parameters (sample size and β). The expected negative log (base 10) p-values for the two tests were estimated from the 200 repetitions.

In Figure 5.1 we compare the expected negative log10 p-values from the two-degree of freedom additive-variance test to the one-degree of freedom additive test. For the sample sizes considered, the association signal from the additive test fails to be strong enough to pass the standard genome-wide significance threshold of 5×10^{-8}. However, for a sample size of 100,000, the significance threshold is crossed for a relatively modest variance effect of 0.025, where the p-value from the AV test is over 700 times smaller than from the additive test. This demonstrates that the additive-variance test has the ability to discover loci that may have been missed by previous additive genome-wide association studies in large sample sizes.

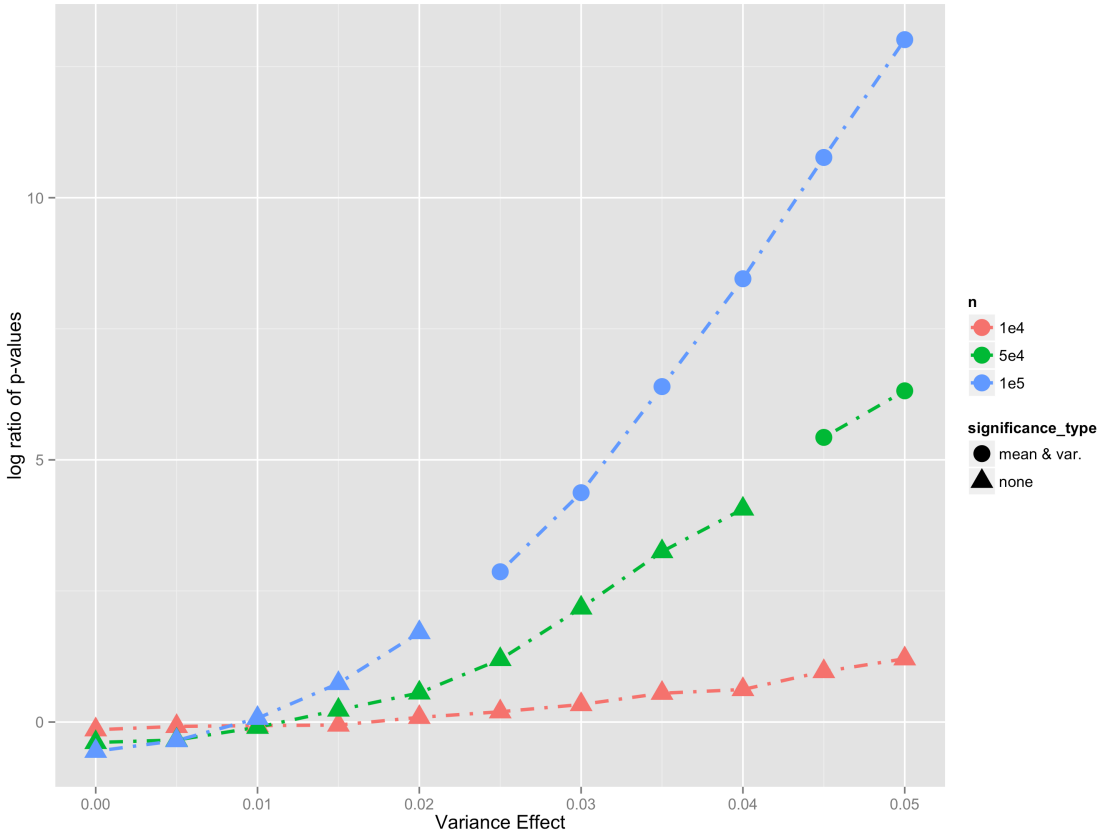


Figure 5.1: **Comparison of association signal for the additive-variance and additive tests**. The association signal when testing for both additive and log-linear variance effects (additive-variance test) compared to testing for only an additive effect (additive test) in simulations. The y-axis gives the expected log-ratio (base 10) of the p-value from the additive test to the additive-variance test for different variance effects of the test locus (x-axis), with values above zero indicating a stronger signal from the additive-variance test. The simulations were performed for sample sizes ranging from 10,000 to 100,000, indicated with the different coloured curves. The log-ratio is plotted as a circle if the expected p-value from the additive-variance test would pass the standard genome wide significance threshold of 5×10^{-8}, and it is plotted with a triangle if neither of the expected p-values from the two tests would pass the significance threshold. For these parameters, no test loci would be expected to pass the significance threshold under the additive test.

5.2.2.2 Population structure control

The mean of a phenotype may differ between populations that are genetically different, which can generate spurious additive associations[42]. To reduce this effect, genetic principal components are often included as mean covariates, which has been proven to be effective in many cases[41, 147].

Analogously, the variance of a phenotype may differ between populations that are genetically different. For example, in the UK Biobank, the variance of log-BMI in people of self-declared British ethnicity is 41% higher than those of self-declared Chinese ethnicity. This could lead to inflation of log-linear variance test statistics if not properly controlled for. We argue, by analogy to population structure affecting the mean, that using genetic principal components as variance covariates in a log-linear variance model can reduce the inflation of log-linear variance test statistics.

To simulate geographic structure in the mean and variance of a phenotype distribution, we used variables from the UK Biobank that give the north and east coordinates of the individuals place of birth in the UK as mean and variance covariates. We simulated phenotypes for the British subsample of the UK Biobank according to the model

$$Y \sim \mathcal{N}(\text{north} - \text{east}, \exp(0.2 * [\text{north} - \text{east}])). \tag{5.8}$$

We fitted models with linear mean and log-linear variance effects for each locus, with and without the top 20 principal components and the genotyping array as mean and variance covariates.

The mean log-likelihood ratio test statistic under the null should be 1, which is what could be achieved with perfect control of population structure in this simulation with no real genetic effects. Without fitting any mean and variance covariates, the mean log-likelihood ratio test statistics across loci on chromosome 22 were: 5.78 for the additive test, and 4.03 for the log-linear variance test. This indicates very strong mean and variance population structure. Fitting the top 20 principal components and the genotyping array as mean and variance covariates reduced the mean test statistics to 1.19 (additive test), and 1.13 (log-linear variance test). We have therefore shown that fitting principal components as variance covariates can be effective at reducing the inflation of log-linear variance test statistics.

For computational efficiency in additive genome-wide association studies, the maximum likelihood estimates of the mean covariates from the null model are often used to project out their effects before fitting models for specific loci. This enables the fitting of locus specific models with only a couple of mean parameters. If $\hat{\alpha}_0$ is the maximum likelihood estimate of the mean effects in the null model, then one transforms the phenotype, Y, to $Y - X\hat{\alpha}_0$. In a similar fashion, the phenotype can be rescaled so as to remove the influence of the variance covariates in the null model, reducing the number of variance parameters to fit for each locus. If $\hat{\beta}_0$ is the maximum likelihood estimate of the variance effects in the null model, the transform performed is

$$Y \to \exp\left(\text{diag}(-0.5V\hat{\beta}_0)\right)(Y - X\hat{\alpha}_0). \tag{5.9}$$

In our simulation of a trait with mean and variance structure, we tested if performing this transform and fitting only locus specific mean and variance effects was effective at controlling for population structure. We used the same mean and variance covariates as above (top 20 principal components and the genotyping array). The mean additive test statistic was 1.12, and the mean log-linear variance test statistic was 1.24. Performing this transform is therefore approximately as effective at controlling for the effects of structure on the test statistics as fitting the full model at each locus, while being computationally more efficient. There may be a loss in power, however, for causal loci that are correlated with mean and variance covariates in the null model.

5.2.3 Inflation control

While the null distributions of test statistics for mean effects are robust to departures from normality, this is not the case for variance test statistics that assume normality[148]. Ignoring other kinds of model misspecification, asymptotically the

null distribution of the test statistics takes the form[148]:

$$\left(1 + \frac{\gamma_2}{2}\right) \chi^2_k, \tag{5.10}$$

where γ_2 is the excess kurtosis of the phenotype distribution, and k is the degrees of freedom of the test. For a normal distribution, $\gamma_2 = 0$, and the test statistics follow the standard asymptotic null distribution. Note that for platykurtic distributions γ_2 is less than zero, leading to deflated test statistics.

One can therefore employ a technique similar to genomic control to obtain properly calibrated test statistics by estimating the excess kurtosis of the phenotype distribution, $\hat{\gamma}_2$, and multiplying the test statistics by $(1 + \hat{\gamma}_2/2)^{-1}$.

However, in real genetic data, other forms of confounding and model misspecification may also inflate test statistics. The mean test statistic across genome-wide SNPs will reflect additional sources of model misspecification, so will give properly calibrated test statistics in a wider set of scenarios than an inflation adjustment based on excess kurtosis alone. We therefore propose to divide variance test statistics by the genome-wide mean of these test statistics before comparing them to the χ^2_1 distribution, a technique similar in practice to genomic control[113]. If the mean test statistic is larger than would be expected due to excess kurtosis, this implies that there are other sources of model misspecification and possibly a true polygenic signal of variance effects.

We tested this approach by simulating non-normal traits with realistic distributions. We first simulated a trait with no real genetic effects by randomly permuting log-BMI among the British subset of the UK Biobank. We fitted models with linear mean and log-linear variance effects for each locus on chromosomes 22 with minor allele frequency greater than 5% and missingness less than 5%. We found that the mean log-linear variance test statistic was 1.290, very close to the prediction based

on the excess kurtosis of log-BMI, 1.295.

The more realistic null scenario is where there are additive effects on a non-normal trait but no true variance effects. To simulate this scenario, normally distributed additive effects were simulated for 1000 evenly spaced loci across chromosome 22. The simulated phenotypes were then constructed as:

$$Y = \sqrt{0.2}(\text{additive-genetic}) + \sqrt{0.8}(\text{permuted log-BMI}). \tag{5.11}$$

Ten phenotypes were simulated independently in this way.

We fitted models with linear mean and log-linear variance effects for each locus on chromosomes 21 and 22 with minor allele frequency greater than 5% and missingness less than 5%. We estimated the mean log-linear variance test statistic across all simulations to be 1.20, slightly higher than the mean expected from the kurtosis of the simulated traits, 1.18. It is possible that the difference is due to linkage with additive effect loci[140].

To control for inflation, for each simulated phenotype, we divided the test statistics by the estimated mean. We combined the inflation-adjusted test statistics across all ten simulations (116,003 in total), and we compare the inflation of the test statistics across the simulations before and after correction (Figure 5.2A). We tested whether the distribution differed from the Chi-Square distribution on 1 degree of freedom using a one-sided Kolmogorov-Smirnov test. While the Kolmogorov-Smirnov test assumes that test statistics are independent, ours are not (due to linkage disequilibrium). However, correlation between test statistics should increase the probability of rejecting the null when it is true. The p-value for the Kolmogorov-Smirnov test was 0.88 (D=0.0022), indicating the null is not rejected, despite the fact that there is an increased probability of rejecting the null when true due to correlated test statistics. We can therefore conclude that log-linear variance test statistics adjusted for inflation

in this way do not deviate significantly from the theoretical null distribution, at least for phenotypes with only additive genetic effects and error distributions similar to log-BMI. We have also shown that log-linear variance test statistics adjusted in this manner are not inflated due to linkage with causal, additive loci with normally distributed effect sizes and allele frequency greater than 5%.

We examined the distribution of log-linear variance effects and standard error estimates. Under the assumptions of the model, the t-statistics for a log-linear variance effect (the log-linear variance effect estimate divided by its standard error) should have a standard normal asymptotic distribution. We find that the mean of the t-statistics across all loci and all simulations is 0.0025 (p=0.45, two-sided t-test), giving no evidence for bias. The variance of the t-statistics was estimated to be 1.20, exactly the same as the inflation of the log-likelihood ratio test-statistics. This can be explained by the asymptotic equivalence of the square of the t-statistic and the log-likelihood ratio test statistic. This implies a simple adjustment to get properly calibrated standard errors for log-linear variance test statistics: multiply the standard errors by the square root of the mean of the log-likelihood ratio test statistics. We adjusted the simulated t-statistics in this way, giving a variance of 1.000131 (Figure 5.2B compares the distribution of t-statistics before and after inflation). The distribution of adjusted t-statistics was not significantly different from a standard normal distribution (p=0.49, Kolmogorov-Smirnov test).

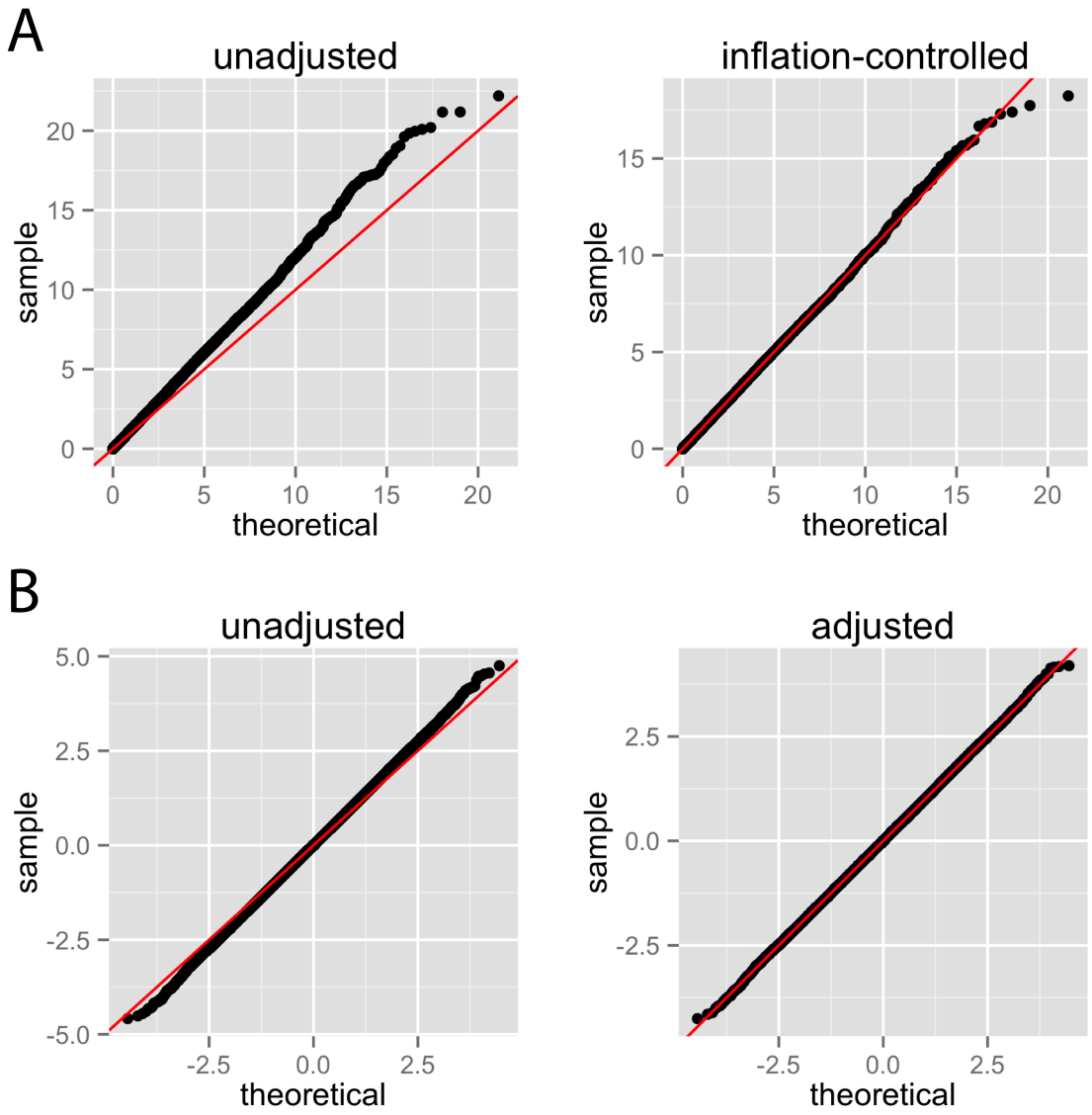


Figure 5.2: **Quantile-quantile plots comparing the theoretical quantiles to the sample quantiles from simulations of a non-normal trait**. To simulate a trait with a realistic non-normal distribution, we permuted log-BMI between unrelated British individuals and added to this a genetic component simulated from additive effects of 1000 loci on chromosome 22. The test statistics are from fitting a model with additive and log-linear variance effects to loci on chromosomes 21 and 22 to ten simulation runs (116,003 log-linear variance test statistics in total). A) the log-likelihood ratio test statistics for a log-linear variance effect, which theoretically are chi-square distributed on one degree of freedom asymptotically, before and after inflation adjustment B) The t-statistics for the log-linear variance effects, which theoretically have standard normal distribution asymptotically, before and after adjustment.

5.2.4 Analysis of body mass index with the heteroskedastic linear mixed model

Based on the power simulations (Figure 5.1), the analysis should be well powered to detect loci with weak additive and moderate variance effects that have been missed by previous additive association studies.

We took the genotyped sample of the UK Biobank split into 'British' and 'Diverse' subsamples, as outlined in the Introduction. To reduce the influence of very strong population structure on our analysis, we further removed individuals from the Diverse Sample who fell outside of the major European cluster in the space defined by the first two principal components (Figure 5.3). This excluded 7,493 individuals from the Diverse Sample, mainly comprised of those with at least some non-European ancestry. The final British Sample has 112,338 individuals, and the final Diverse Sample has 32,404 individuals.

We log-transformed BMI. This reduced the excess kurtosis from 2.56 to 0.57 in the British Sample, which would give a corresponding reduction in inflation of log-linear variance test statistics due to non-normality.

Throughout, we consider only loci with minor allele frequency greater than 5% and missingness less than 5%. We first fitted the full hierarchy of models (Figure 4.2) without a random effect but with age, sex, age^2, age^3, age x sex, age^2 x sex, age^3 x sex, genotyping array, and the top 20 principal components as mean and variance covariates for both the British and diverse subsamples.

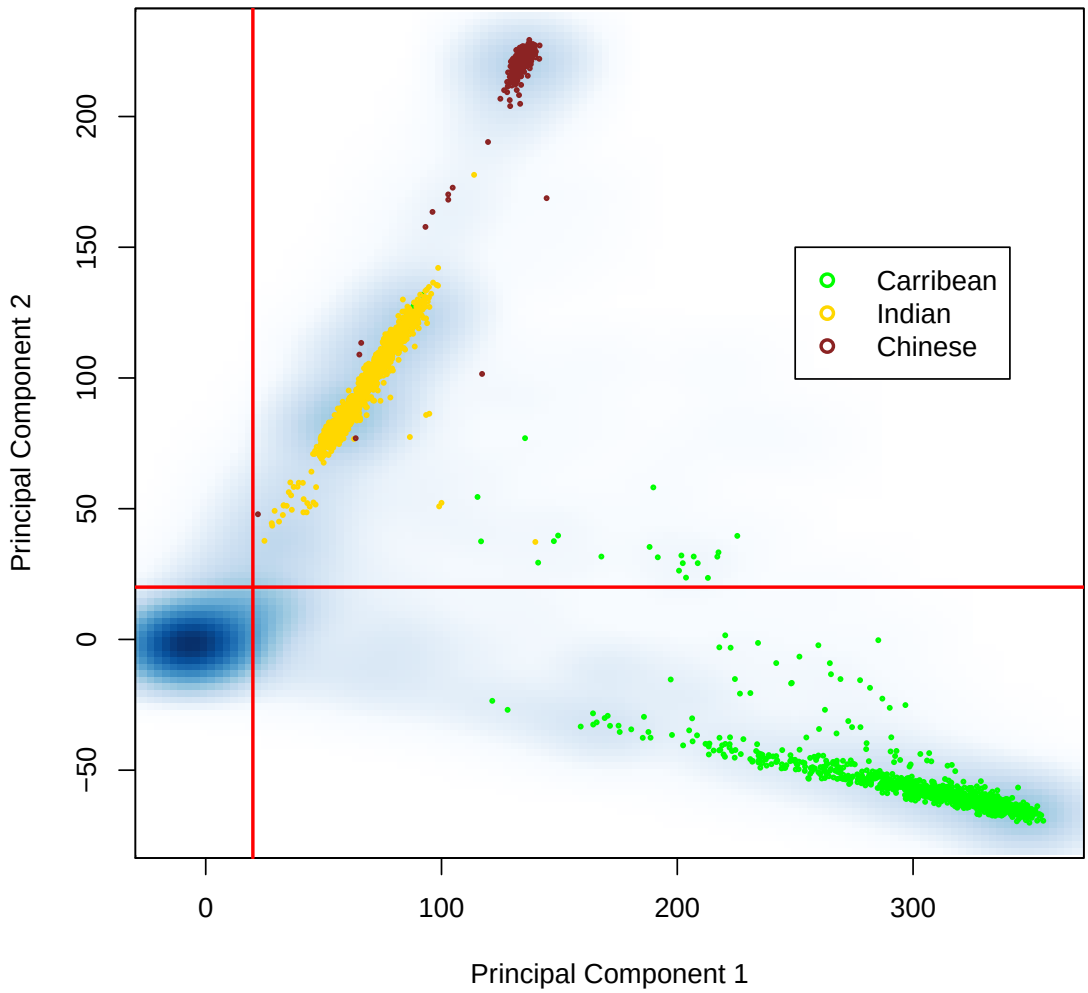


Figure 5.3: **Exclusion of non-European samples based on principal components**. The density of UK Biobank samples projected onto the first two principal components of genetic variation. Individuals with self-declared Chinese, Carribean, and Indian ancestry have been highlighted. The vertical and horizontal lines mark the boundaries for exclusion from the Diverse Sample, with only those in bottom left-hand quadrant retained.

To choose the loci for the random effect in each sample, we ranked the loci by their negative-log p-values for additive effects from fitting the model *without* the random effect. The genome-wide test statistics from the model without the random effect are displayed in Figure 5.4. The advantages of selecting loci in this way have been elucidated[51], although there may be disadvantages compared to using a full-rank

random effect when there is family relatedness[41]. We sequentially selected loci from this ranked list, excluding those that had a correlation of greater than 0.9 with any previously added locus. We chose 1000 loci in this way for the British Sample and 2000 for the Diverse Sample. We only chose 1000 for the British Sample for computational efficiency and because the British Sample has been pruned of relative pairs so should not require a random effect of as large rank to adjust for close relatedness[41]. We then refitted the additive and AV models at each locus with the random effects added. For computational efficiency, we first fitted a null model for each chromosome with all of the mean and variance covariates, obtaining $\hat{\alpha}_0$ and $\hat{\beta}_0$ as the maximum likelihood estimates of α and β. We then perform the transformation

$$Y \to \exp\left(\text{diag}(-0.5V\hat{\beta}_0)\right)(Y - X\hat{\alpha}_0). \tag{5.12}$$

to remove the influence of known covariates on the mean and the residual variance.

Within each sample, we estimated inflation factors for the log-likelihood ratio test statistics for additive effects and for log-linear variance effects. For the additive test statistics, we used LD-score regression[115] to infer an inflation factor, using the LD score statistics for Europeans from [149]. For the log-linear variance test statistics, we calculated the sample mean across all test loci to use as an inflation factor (see above). We formed two degree of freedom tests within each sample as

$$\chi^2_{\text{av}} = \frac{\chi^2_a}{\lambda_a} + \frac{\chi^2_v}{\lambda_v} \tag{5.13}$$

where χ^2_a is the unadjusted log-likelihood ratio test statistic for the additive effect, and χ^2_v is the unadjusted log-likelihood ratio test statistic for the log-linear variance effect; the inflation factor for the additive effect is λ_a, and the inflation factor for the log-linear variance effect is λ_v. Similarly, we multiplied the standard error estimates for the log-linear variance effects in each sample by $\sqrt{\lambda_v}$ to obtain properly calibrated

standard errors (see above).

To combine evidence against the null from the two subsamples, we added the χ^2_{av} statistics from each subsample together to create a four degree of freedom test statistic. For regions with multiple linked loci, we fitted joint models to determine if there was more than one independent signal. When linked loci did not exhibit independent effects, we took the locus with the strongest association as the representative locus.

Since we are interested in loci with variance effects, we further filtered our results to only report loci where the AV model had a lower Bayesian Information Criterion[150, 151] than the additive model. We used the Bayesian Information Criterion to ensure the greater model complexity of the AV model over the additive model was penalised.

To aid interpretation of the BMI results through comparison with another trait, we performed the same analysis for log-height as for log-BMI but without adding random effects.

5.2.5 Visualisation of genome-wide evidence for non-additivity

To visualise the decomposition of the test statistics, and to highlight any regions of the genome showing evidence for non-additivity, we introduce a generalisation of the standard 'Manhattan Plot'[152] that we call the 'Manhattan Information Plot', with 'Information' referencing the interpretation of the log-likelihood ratio test statistic as proportional to a mutual information estimate under certain assumptions. We show an example of this plot for log-BMI in the British subsample of the UK Biobank in Figure 5.4A. (Note that the test statistics in this plot come from fitting the heteroskedastic linear model, without a random effect.) At each locus, we plot the decomposition of the log-likelihood ratio of model M_G (Chapter 4), which allows arbitrary mean and variance effects, to the null model, as in Figure 4.2.

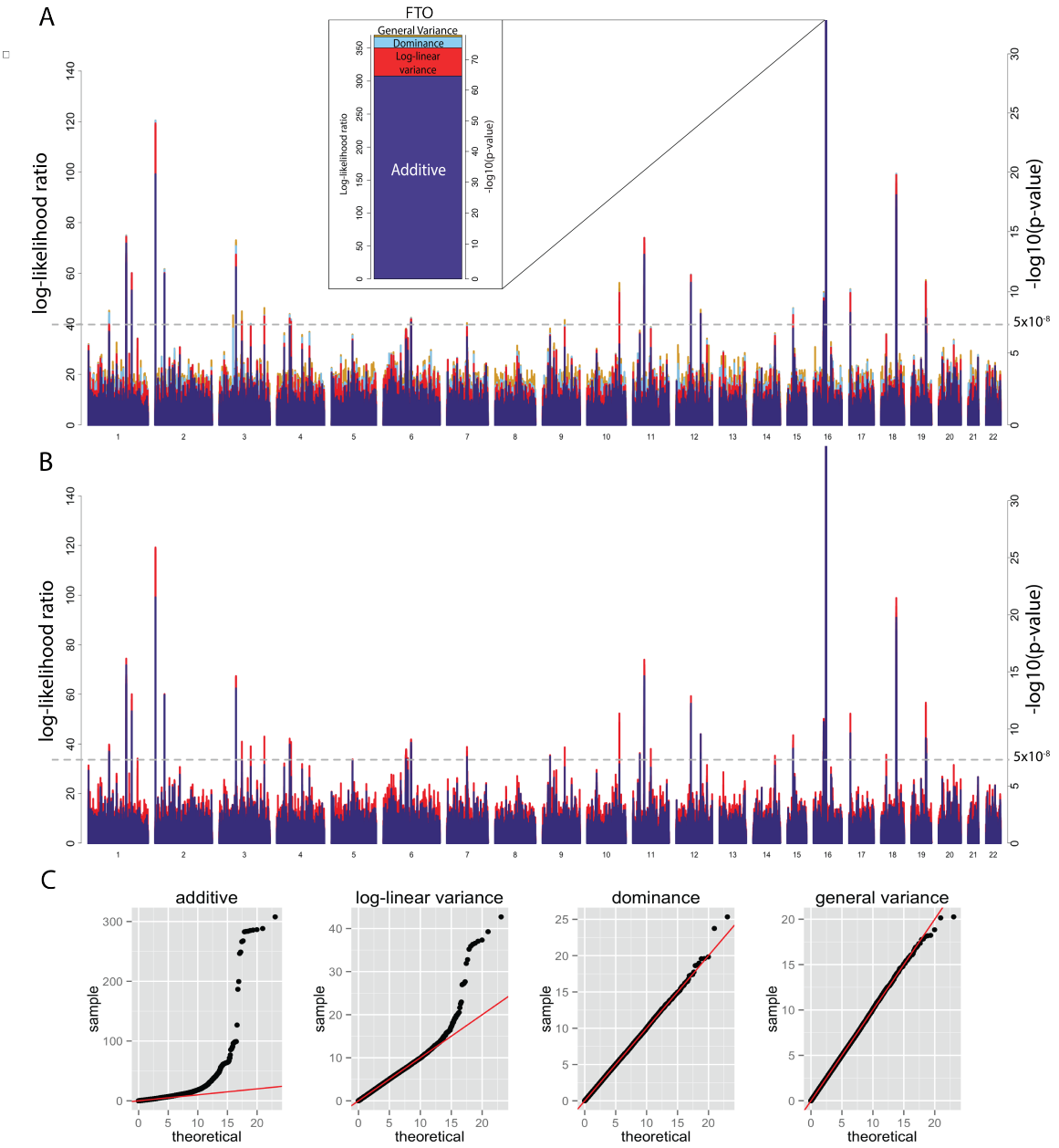


Figure 5.4: **Visualisation of the genome-wide test statistics from fitting the hierarchy of models** (Figure 4.2 illustrates the hierarchy) for log-BMI in the British subsample of the UK Biobank ($n \sim 112,000$). Test statistics were adjusted for inflation before plotting. A) 'Manhattan Information Plot' showing the additive, log-linear variance, dominance, and general variance log-likelihood ratio test statistics, the combined height of which gives a four degree-of-freedom test of association, the -log10(p-values) of which are marked on the right hand y-axis. B) 'Manhattan Sunset' plot showing only the additive and log-linear variance test statistics, the combined height of which gives a two degree-of-freedom test of association, the -log10(p-values) of which are marked on the right hand y-axis. C) The quantile-quantile plots comparing the quantiles of the sample test statistics, after inflation adjustment, to the theoretical quantiles of the χ^2_1 distribution.

Figure 5.4C shows that, while the additive and log-linear variance test statistics depart strongly from the null distribution, the dominance and general variance test statistics do not. This corroborates the argument that the additive-variance test should be more powerful than tests that include additional degrees of freedom, such as in [140].

To visualise the genome-wide test statistics for the additive-variance test, we introduce a simplification of the 'Manhattan Information Plot' that we call the 'Manhattan Sunset Plot', and we show an example of this in Figure 5.4B. The log-likelihood ratio boundary for genome wide significance (5×10^{-8}) is lower in Figure 5.4B than in Figure 5.4A, due to having reduced the degrees of freedom of the test statistic by two, illustrating the improvement in power that could be gained over other mean and variance tests.

5.2.6 Discovery of BMI loci with additive and variance effects in the UK Biobank

Fitting the mixed model reduced the inflation of log-linear variance statistics by 5% in the British Sample and 17% in the Diverse Sample. This shows the mixed model helps to reduce inflation of log-linear variance test statistics, justifying its use beyond the known benefits it brings to additive test statistics.

There is evidence that the distribution of log-linear variance test statistics deviates from the null (Figure 5.5), with large deviations for multiple strong effect loci visible in Figure 5.6 and Figure C.2. We note that 246 SNPs are genome-wide-significant under the additive test, whereas 243 are genome-wide-significant under the AV test, showing there is little loss in power genome-wide from adding the log-linear variance test statistic. While there may be a slight loss of power on average, there is a gain in power to detect loci that do not fit a purely additive model. Out of 13 loci passing the significance threshold (Tables C.1, C.2, C.3, C.4, and C.5) and fitting an AV model

better than an additive model, 8 are loci that have not previously been associated with BMI. Of these 8, three (rs2785980, rs4441044, and rs957919) would not be genome wide significant under an additive association test.

We replicate the previously observed increase in variability with each copy of the *FTO* risk allele[153]: with a combined estimate of 5.35% ([4.08%,6.63%] 95% confidence interval increase) increase in variance of log-BMI per copy of the *FTO* risk allele. Apart from *FTO*, the previously known locus with the strongest evidence for a variance effect is the TMEM18 locus, with a p-value of 1×10^{-6} for a log-linear variance effect.

Three (rs2814992, rs6831020, and rs2785980) of the novel loci have previously been found to have an additive effect on HDL levels[154], including our most striking result, rs2785980, which has an additive association p-value of only 4.7×10^{-4}. A novel locus on chromosome 18 (rs1652376) is an eQTL in adipose tissue ($p = 7.1 \times 10^{-15}$) for NPC1[155], a cholesterol transporter protein that has been implicated in extreme forms of obesity[156].

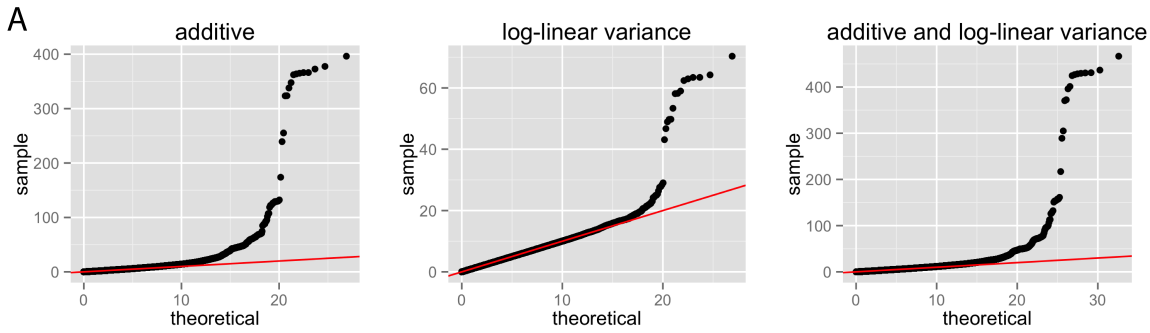


Figure 5.5: **QQ-plots for test statistics from the HLMM**. Comparison of inflation-adjusted test statistic quantiles for log-BMI to the quantiles of the theoretical, asymptotic null distribution, which is a Chi-Square distribution of appropriate degrees of freedom. See Figure 5.6 for the visualisation of these statistics genome-wide.

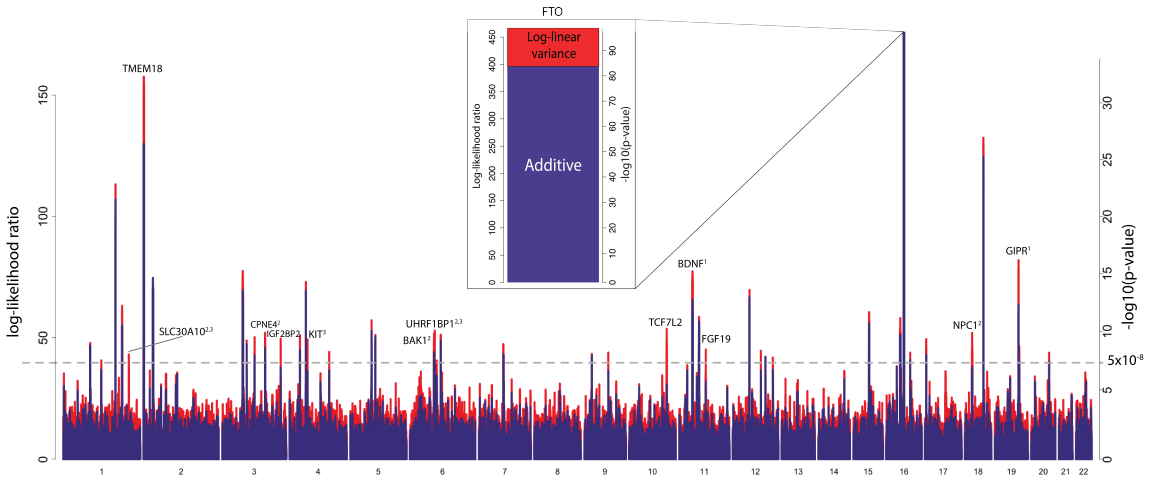


Figure 5.6: **'Manhattan Sunset' plot visualising the genome-wide additive and log-linear variance test statistics for log-BMI**. At each locus, the additive log-likelihood ratio test statistic is plotted as a blue bar, and the log-likelihood ratio test statistic for a log-linear variance effect is added on top of this, the combined height of which gives a two-parameter test of association, the -log10(p-values) of which are marked on the right hand y-axis. The test statistics from the two subsamples were combined after inflation adjustment. Names were added for loci that passed genome-wide-significance ($p = 5 \times 10^{-8}$) and that had a lower Bayesian Information Criterion for the model with both additive and log-linear variance effects than for the model with only additive effects (Tables C.1, C.2, C.3, C.4, and C.5). The name indicates the nearest gene and/or a gene that the variant controls expression of: 1) indicates the SNP is a missense variant; 2) indicates the SNP is a eQTL for the named gene according to the GTEX data[155]; and 3) indicates the variant has previously been associated with HDL levels[154]. Figure C.2 is a larger version of this figure.

5.2.7 *TCF7L2* interactions

Included in the known loci is rs7903146, one of the SNPs in the *TCF7L2* locus that tag a haplotype which confers the greatest increase in type-II-diabetes risk out of any common genetic variation[157]. While the fact that carriers of the *TCF7L2* risk alleles have reduced BMI has been noted before[157], we note for the first time that there is also a strong reduction in variation in BMI of around 3.6% ([-5.1%,-2.2%], 95% confidence interval) per copy in the British Sample ($p = 2.2 \times 10^{-6}$).

We attempted to investigate whether the variance effect of *TCF7L2* could be explained by interactions with age, sex, type-II-diabetes status, or insulin treatment.

We took the variables 'diabetes diagnosed by doctor' (data field 2443) and 'started insulin treatment within one year of diagnosis of diabetes' (data field 2986), using them as indicator variables after setting to missing those who answered 'Do not know' or 'Prefer not to answer'. We fitted HLMMs in both the British and diverse subsamples with the random effects and covariates used for the null model in the main log-BMI analysis; to the mean covariates, we added rs7903146 and its interactions with age and sex, and indicator variables for diabetes diagnosis and insulin treatment within one year of diagnosis, and their interactions with rs7903146. Using the *R* package meta[158], we tested for heterogeneity between the estimates in the British and Diverse Sample using the standard Q-statistic for heterogeneity in meta-analysis[159], and we combined estimates in a fixed effects meta-analysis if the p-value for the heterogeneity test was above 0.05.

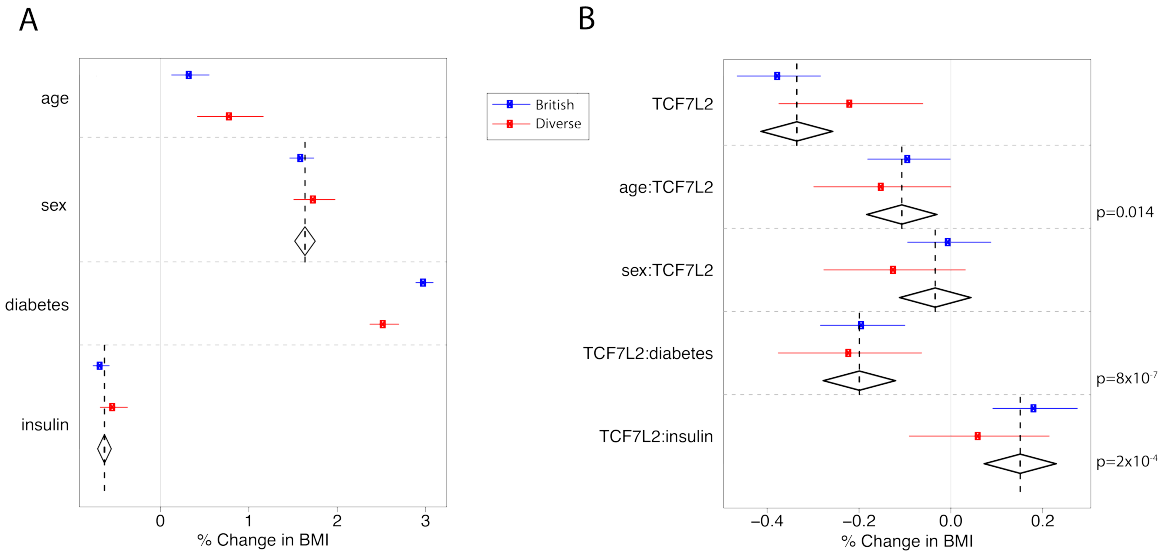


Figure 5.7: **Interaction analysis of the *TCF7L2* risk allele (rs7903146)**. Effects were estimated separately in the British (blue) and diverse (red) subsamples in heteroskedastic linear mixed models. If there was no statistically significant evidence for heterogeneity, effects were combined in a fixed effects meta analysis (diamonds). The width of the bars and diamonds indicate the 95% confidence intervals. Effects were transformed from the log-scale to give % change (per year for age, change from female to male for sex). A) main effects of age, sex, diabetes diagnosis by doctor, and insulin treatment started within one year of diabetes diagnosis. B) Main effect of the *TCF7L2* risk allele (rs7903146), and its interactions. We added p-values for the variables with significant interactions with *TCF7L2* variation.

We found that, in the joint model including diabetes status, insulin treatment and their interactions with rs7903146, diabetes is associated with higher BMI (2.99% in the British subsample versus 2.53% in the Diverse Sample, $p = 2.7 \times 10^{-6}$ for heterogeneity), while insulin treatment is associated with lower BMI (-0.63%, [-0.77%,-0.55%]; $p = 7.4 \times 10^{-56}$) (Figure 5.7A).

There is consistent evidence across the two subsamples that the reduction in BMI with each copy of the *TCF7L2* risk allele grows stronger with age (-0.10%, [-0.18%,-0.03%], per year; p=0.014), is stronger in those diagnosed with diabetes (combined estimate: -0.20%, [-0.28%,-0.12%] change in effect if diagnosis reported; $p = 8 \times 10^{-7}$), and is weaker in those who reported having had insulin treatment (combined estimate: 0.15%, [0.07%,0.23%] change in effect if insulin treatment reported; $p = 2 \times 10^{-4}$). While the stronger reduction in BMI in diabetics has been noted before[157], the interactions with insulin treatment and age are novel, as far as we are aware.

To test if these effects were due to censoring effects — individuals with higher BMI and the *TCF7L2* risk allele dying at higher rates than other individuals, for example — we reran the analysis in the younger half (below 58 years old) of the British subsample. The main effect estimate of the *TCF7L2* risk allele changed from -0.37% to -0.41%, showing that the estimated reduction in BMI per copy of *TCF7L2* is not being driven by older samples. The estimated interaction with age changed from -0.90% to -0.79%; for the interaction with diabetes diagnosis, from -0.19% to -0.21%; and for the interaction with insulin treatment, from 0.18% to 0.17%. It is therefore clear that the interaction effects are not overly influenced by the statistics of the older half of the sample, where censoring due to differential mortality will be much stronger.

The *TCF7L2* risk allele is associated with decreased proinsulin expression in human islet cells[160]. The interaction with insulin treatment could be showing that the reduction in BMI with the *TCF7L2* risk allele is suppressed by insulin treatment,

which may interfere with the effects of *TCF7L2* variants on regulation of insulin production and processing. However, it is impossible to determine causality from these data alone. It could also be that those carrying the *TCF7L2* risk variants and higher than expected BMI were more likely to be given insulin treatment than would be expected from looking at either factor independently.

5.3 Discussion

The best performing methods for additive association testing utilise the advantages of linear mixed models in a scalable and flexible way. We have extended the advantages of linear mixed model additive association testing to testing for both mean and variance effects, characteristic of loci involved in interactions. The hierarchy of models (Figure 4.2) can be incorporated into a generalisation of the standard mixed model, the heteroskedastic linear mixed model, for which we provide an algorithm that scales linearly with sample size for a fixed number of random effects.

A key assumption of our approach is that of normally distributed error terms. Variance test statistics are sensitive to departures from normality, and computationally expensive bootstrap procedures have been proposed for computing p-values[141]. We find that the inflation due to model misspecification follows simply from the excess kurtosis of the distribution, as expected from classical theory[148]. This justifies a simple and computationally efficient genomic-control like procedure for adjustment of test statistics and standard errors, which we show to be effective in simulations. This results in estimates of log-linear variance effects that are asymptotically normal with known standard error, opening up the possibility for meta-analysis using summary statistics. While we have concentrated on the log-transform here, other more complicated transforms may be preferred for other non-normal traits[161]. However, very complicated transformations make make interpretation of results difficult.

We found three novel BMI associations (rs2785980, rs4441044, and rs957919) through jointly testing for additive and log-linear variance effects that would not have been discovered by additive testing. This demonstrates that there exist associations that have been overlooked by previous genome-wide association studies that could have been discovered using this method. Simulations show that the relative gain of the joint AV test over the additive test increases with sample size (Figure 5.1). We therefore expect that there are many more loci affecting BMI with weak additive and moderate variance effects waiting to be discovered with this method at increased sample sizes.

A variance effect of a locus can in general be seen as a signature of model misspecification, which is useful to highlight as it prompts searching for the right model. There are multiple possibilities for the source of model misspecification, including: interactions with other variables influencing the trait, other non-linear models, and linkage with rare variants or haplotypes[140]. The 'Manhattan Sunset' plot (Figure 5.6 and Figure C.2) gives a way to visualize genome-wide evidence for any of these effects, highlighting any regions that may warrant further investigation. The method could be used to help discover interactions between genetic variants by first filtering on additive and variance test statistics before performing tests between all pairs of loci passing the filter, greatly reducing the number of interaction models considered.

We argue that our results are unlikely to be due to linkage with common, additive variants, because our simulations showed that our inflation correction procedure adjusts for this. It is harder to rule out linkage with unobserved rare variants or haplotypes. The novel associations on chromosome 6 (rs2814992 and rs9469488, Tables C.2 and C.5) are both near the HLA region, and may reflect complex haplotype structure in that region. It is unlikely that all of our results reflect this phenomenon, however, as they include known associations at the *FTO* and *TCF7L2* loci that are not caused

by these phenomena[157, 94]. Additionally, the distribution of log-linear variance test statistics for height does not show any deviation from the null (Figure C.1), and there is no reason to suspect that these phenomena are less common in the genetic architecture of height than in the genetic architecture of BMI. We therefore believe, given the strong prior plausibility for interactions between genes and environment in BMI variation and the example of *FTO*, that these results probably reflect that many BMI loci are involved in interactions with environmental and lifestyle factors.

A further possibility is that variance effects could reflect effects on intra-individual variation. Given that an individual's weight is not constant through their life, it is plausible that genetic variants affect how variable an individual's weight is over time, possibly in response to interactions with changing lifestyle habits.

While it is an advantage of the method that one does not need to have observed or modelled the other variables that interact with a locus to detect the resulting variance effect, it also implies that determining the particular interaction model could be challenging or impossible. Nevertheless, the strong variance effect of the *TCF7L2* risk variant (rs7903146) encouraged us to look for interactions with plausible candidates, discovering interactions with insulin treatment and age. The variance effect provides a measure of whether discovered interactions explain the observed variance effect, with the discovered interactions explaining some but not all of the variance effect of the *TCF7L2* risk variant. Similarly, the strong variance effects of the known, strong BMI loci *TMEM18*, *BDNF*, and *GIPR* suggest that we should investigate whether these loci are involved in interactions or other non-linear models. The fact that three of our novel associations have previously been associated with high density lipoprotein levels[154] also suggests there may be a HDL related pathway that is interacting with other factors to affect BMI.

While we have concentrated on variants that are genome-wide significant under the additive-variance test, there is information about non-additivity in the whole

distribution of the log-linear variance test statistics. It may be possible to extend methods that utilise the LD-score of variants[115] to investigate non-additive genetic architecture. This could characterise which traits show evidence for polygenic non-additivity, thereby helping to measure the importance of non-additivity in human genetics.

Chapter 6

Conclusions

6.1 Epistatic variance

6.1.1 Epistatic variance in model organisms

There are abundant examples of epistasis in model organisms[73]. While most of the examples are of pairs of interacting variants or genes, higher order interactions have been discovered[162]. The discovery of statistical interactions has been used to further understanding of biological mechanism[73] and improve prediction of traits[163] .

Direct estimation of epistatic variance components for complex traits in model organisms has only recently become possible due to the large sample sizes required. Similar analyses to the one performed in Chapter 2 have given similar estimates for the proportion of epistatic variance in haploid[88] and diploid[164] yeast growth traits. The proportion of variance due to third order interactions has been estimated to be much smaller than the variance due to pairwise interactions[88, 164]. This corroborates theoretical arguments that the amount of variance due to each order of interaction should sharply decline with the order of interaction[165]. It is possible that some of the difference between the broad sense heritability, as estimated by strain repeatability[84], and the variance explained by additive effects and pairwise

interaction effects could be due to non-genetic effects[88]. This could mean that the amount of higher order epistatic variance is somewhat smaller than inferred in the analysis of the yeast cross.

The experimental cross design generates large variation in kinship, which helps in fitting epistatic variance components. However, because only one or a few generations of recombination have taken place, there is substantial linkage disequilibrium in the genome, which prevents fine-mapping of causative loci and their interactions. There is therefore a need to generate experimental crosses that have been mated for sufficient numbers of generations to fine map interactions so that they can be functionally interpreted.

Epistasis can also be used to study the incompatibilities between hybrids. The classical Dobzhansky-Muller incompatibility in a cross between two diverged populations is a form of epistasis[166]. Furthermore, it is combinatorially easier to evolve higher order incompatibilities between populations than pairwise incompatibilities[92]. However, it is combinatorially challenging to map individual higher order interactions. The methods of Chapter 2 could be employed to investigate whether higher order incompatibilities are common in crosses of diverged populations by measuring how epistatic variance components change with hybrid incompatibility. This would give a measure of how many parts tend to be functionally linked in natural populations and how quickly, on an evolutionary timescale, these co-adaptations tend to evolve.

6.1.2 Epistatic variance in humans

Theoretical arguments against the importance of epistasis for complex traits in natural populations have been given[165, 89]. An important observation made by these authors is that the amount variance contributed by an interaction depends on the

product of the variances of the genotypes at the interacting loci:

$$V = f_1(1 - f_1)f_2(1 - f_2), \tag{6.1}$$

where V is proportional to the variance contributed by an interaction between two bi-allelic loci in linkage and Hardy-Weinberg equilibrium with frequencies f_1 and f_2. Assuming that $f_1 = f_2 = f$, the ratio of the variance contributed by the additive effects of the loci to the variance contributed by the interaction would be proportional to

$$\frac{f(1 - f)}{f^2(1 - f)^2} = \frac{1}{f(1 - f)}. \tag{6.2}$$

This shows that the relative amount of epistatic variance compared to additive variance tends to decline with average levels of heterozygosity of causative genetic variants. It is therefore likely that traits in natural, outbred populations such as humans exhibit less epistatic variance than traits in F1 crosses, such as the yeast cross analysed in Chapter 2, where all allele frequencies are 0.5 and heterozygosity is much higher.

Complex traits in humans tend to involve at least hundreds of causative genetic variants[114], whereas traits in organisms such as yeast tend to involve fewer genetic variants, some of very large effect[163]. It has been argued that epistatic variance is likely to tend to zero as the number of causative genetic loci tends to infinity[165]. The basis of this argument relies upon a model where single locus and interaction effects are of similar magnitude, in the same direction, and all pairs of loci interact with one another. While this is one possible model for a complex trait, many others are possible that do not necessarily result in a small amount of epistatic variance. There may be many independent pairs of loci affecting a trait with interactions within each pair but not between pairs. In this case, large amounts of epistatic variance relative to additive variance can be generated when the effect of one locus changes sign depending upon

the state of the other locus[73] (Figure 1.1A). It is therefore unconvincing to reject the possibility of substantial epistatic variance in complex traits in humans on the basis of *a priori* arguments alone.

It has been argued that twin correlations are incompatible with large amounts of epistatic variance[167]. If the environmental similarity between monozygotic and dizygotic twins is assumed to be equal (the 'equal environments assumption'[168]), then the correlations between monozygotic (MZ) and dizygotic (DZ) twins in an outbred population for a trait with only pairwise epistasis are[55]

$$r_{\mathrm{DZ}} = c + \frac{1}{2}h^2 + \frac{1}{4}h_2^2, \tag{6.3}$$

$$r_{\mathrm{MZ}} = c + h^2 + h_2^2, \tag{6.4}$$

where r_{MZ} is the correlation between monozygotic twins, r_{DZ} is the correlation between dizygotic twins, c is the correlation due to the shared environment, h^2 is the narrow-sense heritability, and h_2^2 is the proportion of phenotypic variance from pairwise interactions. It has been argued that, because in general $2r_{\mathrm{DZ}} > r_{\mathrm{MZ}}$, epistatic variance cannot be large[167]. However, the only precise conclusion that can be drawn from twin correlations in this model is about the relative size of shared environmental variance and epistatic variance:

$$2r_{\mathrm{DZ}} - r_{\mathrm{MZ}} = c - \frac{1}{2}h_2^2, \tag{6.5}$$

so if $2r_{\mathrm{DZ}} - r_{\mathrm{MZ}} = x$, then $h_2^2 = 2(c - x)$. It is therefore impossible to rule out substantial epistatic variance contributing to human traits on the basis of dizygotic and monozygotic twin correlations alone. If there is no epistasis or dominance, then $2r_{\mathrm{DZ}} - r_{\mathrm{MZ}} = c$, and this is the standard twin studies estimator for c. Therefore, if $h_2^2 > 0$, then $c > 2r_{\mathrm{DZ}} - r_{\mathrm{MZ}}$, and the contribution of the shared environment to

human traits has been underestimated.

While *a priori* arguments about the amount of epistatic variance and arguments based on twin studies are unsatisfactory, the empirical measurement of epistatic variance in humans is challenging. We have shown in Chapter 2 that precise measurement of epistatic variance in humans would require a large sample with high kurtosis in its kinship distribution. While this can be achieved by sampling a wide range of relatedness, the covariance between relatives due to epistasis then becomes strongly confounded with shared environment. If close relatives are excluded, then a sample of a few thousand from a strongly bottlenecked population such as the Hutterites would be required. Samples of that size do not yet exist for such strongly bottlenecked populations, but may do in the future.

Without direct measurement of epistatic variance components in humans, the most pertinent empirical data comes from studies that attempt to discover particular interactions between genetic variants. While discovery of interactions between genetic variants is difficult due to power considerations and other complexities, the 'variance prioritisation' approach has been employed with some success to discover interactions affecting human gene expression[72]. For close to half of the discovered interactions in a recent study, the amount of epistatic variance was larger than the amount of additive variance[72], implying that interactions that could generate substantial epistatic variance do exist in human genetics. For more complex traits, there is a lack of strong evidence for many interactions between genetic variants. It is possible that this is simply due to a lack of large enough sample sizes and appropriate methods. If interactions between genetic variants affecting complex traits are still impossible to find at larger sample sizes, however, that may enable the amount of pairwise epistatic variance to be bounded based on power calculations.

For now, we believe it is reasonable to conclude that the amount of epistatic variance in humans is likely to be much smaller than the amount of additive variance

and the amount observed in model organism studies, but it could still be substantial enough to matter for both genomic prediction and for methods than rely on an assumption of additivity. Finding the interactions underlying even a small amount of epistatic variance could lead to important improvements in understanding of mechanistic interactions underlying human physiology[169].

6.2 Gene-by-environment interactions

6.2.1 Methodology

The heteroskedastic linear mixed model we developed was able to discover variants affecting body mass index that would not be discovered based on additive testing alone. This raises the possibility that many such variants have been overlooked by additive genome-wide association studies, and that variants such as these could be discovered for BMI and other traits at similar or larger sample sizes.

While the current algorithm is efficient for low-rank random effects, it is not for full-rank random effects. A full-rank heteroskedastic linear mixed model constructed from dense genome-wide genetic variants would be particularly useful for datasets with many close relative pairs[41]. Furthermore, the heteroskedastic linear mixed model allows modelling of heteroskedasticity of genetic effects due to annotations. An efficient full-rank heteroskedastic linear mixed model would enable the assessment of which annotations are associated with larger magnitude effects on traits. It would have an advantage over previous methods[170] in that it could include continuous and overlapping annotations.

6.2.2 Prevalence and utility

We have provided strong evidence that lifestyle and environment can modify the effect of genetic variants on body mass index (BMI), primarily variants at the *FTO*

locus (Chapter 3). Furthermore, many variants affecting mean BMI also appear to affect variability in BMI, which is suggestive of widespread gene-by-environment interactions affecting BMI. This includes the type-II diabetes risk variant at the *TCF7L2* locus, whose effect on BMI we found to be modified by insulin treatment.

It is unlikely that BMI is an exceptional trait. It is a complex trait that is related to many other traits, so if gene-by-environment interactions affect BMI then they are likely to affect other related traits.

Factors known to affect body mass index, such as diet and exercise, have changed dramatically in the period leading to the current 'obesity epidemic'[60]. This may have revealed 'cryptic genetic variation' that previously had little statistical effect of body mass index variation due to lack of a suitable environment[171]. The changes in the environment may have resulted in an increase in the average effect of variants such as *FTO* on BMI.

If a changing and more variable environment is responsible for the presence or prominence of some gene-by-environment interactions affecting BMI, then it is possible that gene-by-environment interactions also affect other traits that have changed recently due to a changing environment. These traits could include other metabolic traits[172, 173, 174], fertility traits[175], psychiatric traits[176], and educational traits.

If gene-by-environment interactions are prevalent in traits of importance to health and education, then their effects should be precisely characterised to improve prediction. This could lead to personalised lifestyle interventions: for example, diet and exercise advice for weight loss could be based on genetic information.

A major obstacle to useful environmental intervention advice based on genetic information is the question of causality of gene-by-environment interactions. Power requirements imply that observational studies, such as those we conducted, will be necessary to discover gene-by-environment interactions. Before such gene-by-environment interactions can be used for interventions, they should be tested for

causality using different study designs, some of which can be borrowed from the social sciences[177].

We note that gene-by-lifestyle interactions are likely to imply interactions between genetic variants: this is because many lifestyle factors and behaviours, such as dietary behaviours and exercise, have a heritable component[167]. It is therefore likely that a variant such as *FTO* interacts with genetic variants affecting physical activity and dietary behaviours. This form of genetic interaction would be highly polygenic, however, making it hard to detect based on scans for pairwise interactions. It raises the intriguing possibility of testing for causality of gene-by-environment interactions by using a form of Mendelian randomisation[178, 179]: to construct polygenic scores for lifestyle factors and test if they interact with variants such as *FTO*. While there are problems with this approach as a method to positively prove causality, negative results could reveal that observational gene-by-environment results are due to confounding[180].

6.3 Missing heritability

This study has not provided evidence that epistatic variance contributes substantially to variation in complex human traits. It has, however, provided additional evidence that gene-by-environment interactions play a role in complex human traits, which can imply an element of epistasis when environmental/lifestyle factors are heritable. The evidence for gene-by-environment interactions is restricted to BMI in this study. While it is unlikely that BMI is exceptional, implying that gene-by-environment interactions are likely to affect other complex human traits, we did not see any evidence for such interactions in our preliminary analysis of height. The problem of missing heritability exists for height as well as for BMI, so this argues against non-additivity being a primary explanation for missing heritability.

It is generally harder to detect variants with non-additive effects than purely additive variants with standard additive association testing. We show that this theoretical argument has merit by providing empirical evidence in Chapter 5 that additive association testing has missed some variants affecting BMI. It is likely then that un-modelled non-additivity has contributed to lower rates of genetic discovery and variance explained by models of traits. It is also likely that non-additivity has contributed to overestimation of heritability[56, 55]. We therefore believe that it is reasonable to conclude that while interactions, and non-additivity more generally, are not primary explanations for missing heritability, they have contributed to the problem for at least some complex traits in humans.

Appendix A

A.1 Detailed theory

A.1.1 Covariance between individuals in a finite population

We model the phenotype of each diploid individual t as $Y_t = G_t + \epsilon_t$, where G_t is the effect of the genotype of individual t, and ϵ_t is the residual error arising from noise and the environment, with $\mathbb{E}[\epsilon_t] = 0 \; \forall t$. We do not model genotype-by-environment correlation or interaction here, so $\mathrm{Cov}(G_t, \epsilon_t) = 0 \; \forall t$.

We model G_t to accommodate any possible interaction effects between loci and to give an orthogonal partitioning of the genetic variance. The model allows every subset of the causal loci L to interact in an arbitrary way to produce an effect, X_{tL}, on the phenotype of individual t. Within each subset L, each possible sequence of alleles across the loci, $s \in S_L$, can have a different effect on the phenotype, β_{Ls}. Therefore, the model is general for any genotype-phenotype map without dominance effects.

For a set of causal loci N, which we assume are inherited independently and therefore in linkage equilibrium,

$$G_t = \sum_{L \subseteq N} X_{tL}; \; X_{tL} = \sum_{s \in S_L} \beta_{Ls} \prod_{l \in L} x_{tls[l]}; \tag{A.1}$$

where $X_\emptyset$ is the phenotypic mean; $S_L = \{A, T, G, C\}^{|L|}$ is the set of possible sequences

of alleles across the loci in L; $s[l]$ is the allelic state of the locus l in the sequence s; and

$$x_{tlA} = g^p_{tlA} + g^m_{tlA} - 2f_{lA}, \tag{A.2}$$

where g^p_{tlA} is an indicator variable for the presence of allele A at locus l on the paternally inherited chromosome of individual t, g^m_{tlA} indicates the equivalent on the maternally inherited chromosome, and f_{lA} is the frequency of the A allele at locus l.

Because $\mathbb{E}[x_{tlk}] = 0\ \forall t, l, k$ and the loci are in linkage equilibrium, $\mathrm{Cov}(X_{tL}, X_{tL'}) = 0$ for $L \neq L'$. Proof: without loss of generality, $\exists\, d \in L \backslash L'$, implying

$$\begin{aligned}\mathrm{Cov}(X_{tL}, X_{tL'}) &= \sum_{s\in S_L}\sum_{s'\in S_{L'}} \beta_{Ls}\beta_{L's'} &\text{(A.3)}\\ &\quad \mathbb{E}\left[\prod_{l\in L\backslash\{d\}} x_{tls[l]} \prod_{l'\in L'} x_{tl's'[l']}\right] \mathbb{E}[x_{tds[d]}]\\ &= 0,\ \text{as } \mathbb{E}[x_{tds[d]}] = 0. &\text{(A.4)}\end{aligned}$$

The covariance between arbitrary relatives t and u with kinship coefficient $K_{t,u}$ is therefore

$$\begin{aligned}\mathrm{Cov}(G_t, G_u) &= \sum_{L\subseteq N} \mathrm{Cov}(X_{tL}, X_{uL}); &\text{(A.5)}\\ &= \sum_{s\in S_L}\sum_{s'\in S_L} \beta_{Ls}\beta_{Ls'}\mathbb{E}[\prod_{l\in L} x_{tls[l]}x_{uls'[l]}]. &\text{(A.6)}\end{aligned}$$

To evaluate $\mathbb{E}[\prod_{l\in L} x_{tls[l]}x_{uls'[l]}]$, we further assume that the IBD sharing events are independent at different loci. Let $\mathrm{IBD}^l_{t,u}$ represent the IBD sharing state of the pair s, t at locus l. The joint distribution of $x_{tls[l]}, x_{uls'[l]}$ is determined only by the allele frequencies at the locus in the ancestral population and $\mathrm{IBD}^l_{t,u}$. Therefore, under the assumption that alleles at different causal loci are inherited independently, $x_{tls[l]}, x_{uls'[l]} \perp x_{tl's[l]}, x_{ul's'[l]} | \mathrm{IBD}^l_{t,u}, \mathrm{IBD}^{l'}_{t,u}$ for $l \neq l'$. Generalising this to all causal loci,

this implies

$$\mathbb{E}\left[\prod_{l\in L} x_{tls[l]}x_{uls'[l]}|\mathrm{IBD}^1_{t,u},\mathrm{IBD}^2_{t,u},\ldots,\mathrm{IBD}^N_{t,u}\right] = \tag{A.7}$$

$$\prod_{l\in L}\mathbb{E}[x_{tls[l]}x_{uls'[l]}|\mathrm{IBD}^1_{t,u},\mathrm{IBD}^2_{t,u},\ldots,\mathrm{IBD}^N_{t,u}] \tag{A.8}$$

$$=\prod_{l\in L}\mathbb{E}[x_{tls[l]}x_{uls'[l]}|\mathrm{IBD}^l_{t,u}]. \tag{A.9}$$

We therefore have that, by the Law of Total Expectation,

$$\mathbb{E}\left[\prod_{l\in L} x_{tls[l]}x_{uls'[l]}\right] = \sum\prod_{l\in L}\mathbb{E}[x_{tls[l]}x_{uls'[l]}|\mathrm{IBD}^l_{t,u}]\mathbb{P}(\mathrm{IBD}^1_{t,u},\mathrm{IBD}^2_{t,u},\ldots,\mathrm{IBD}^N_{t,u}), \tag{A.10}$$

where the sum is over all possible IBD sharing states at all causal loci. If the IBD sharing events at different causal loci are independent, then

$$\mathbb{P}(\mathrm{IBD}^1_{t,u},\mathrm{IBD}^2_{t,u},\ldots,\mathrm{IBD}^N_{t,u}) = \prod_{l\in L}\mathbb{P}(\mathrm{IBD}^l_{t,u}). \tag{A.11}$$

Therefore,

$$\mathbb{E}\left[\prod_{l\in L} x_{tls[l]}x_{uls'[l]}\right] = \sum\prod_{l\in L}\mathbb{E}[x_{tls[l]}x_{uls'[l]}|\mathrm{IBD}_{t,u}]\mathbb{P}(\mathrm{IBD}^l_{t,u}) \tag{A.12}$$

$$=\prod_{l\in L}\mathbb{E}[x_{tls[l]}x_{uls'[l]}]. \tag{A.13}$$

We now evaluate each term in the product for the case when $s[l] = s'[l]$:

$$\mathbb{E}[x_{tls[l]}x_{uls'[l]}] = \sum_{i=m,p}\sum_{j=m,p}\mathrm{Cov}(g^i_{tls[l]},g^j_{uls'[l]}). \tag{A.14}$$

$$\mathrm{Cov}(g^i_{tls[l]},g^j_{uls'[l]}) = \mathbb{E}[g^i_{tls[l]}g^j_{uls'[l]}] - f_{ls[l]}f_{ls'[l]}; \tag{A.15}$$

$$= \left(\frac{K^{i,j}_{t,u}-K_0}{1-K_0}\right) f_{ls[l]}(1-f_{ls[l]}), \tag{A.16}$$

from the genotypic covariance in a founder population (2.13).

When $s[l] \neq s'[l]$, assuming no mutation, $\mathbb{E}[g^i_{tls[l]} g^j_{uls'[l]}]$ is only non-zero when the haplotypes are not IBD, because the alleles are different. Given that the haplotypes are not IBD, $\mathbb{E}[g^i_{tls[l]} g^j_{uls'[l]}]$ is the probability that t inherits allele $s[l]$ from one of the A ancestral haplotypes and u inherits allele $s'[l]$ from one of the other $A-1$ ancestral haplotypes. Therefore, if $c_{ls'[l]}$ is the number of ancestral haplotypes carrying the $s'[l]$ allele,

$$\mathbb{E}[g^i_{tls[l]} g^j_{uls'[l]}] = (1 - K^{i,j}_{t,u}) f_{ls[l]} \frac{c_{ls'[l]}}{A-1} = \frac{(1 - K^{i,j}_{t,u})}{1 - K_0} f_{ls[l]} f_{ls'[l]}. \tag{A.17}$$

Therefore,

$$\mathrm{Cov}(g^i_{tls[l]}, g^j_{uls'[l]}) = -f_{ls[l]} f_{ls'[l]} \left(\frac{K^{i,j}_{t,u} - K_0}{1 - K_0} \right). \tag{A.18}$$

Therefore,

$$\mathbb{E}[x_{tls[l]} x_{uls'[l]}] = 2 \left(\frac{K_{t,u} - K_0}{1 - K_0} \right) \xi_{l:s[l],s'[l]}, \tag{A.19}$$

where $\xi_{l:s[l],s'[l]} = -2 f_{ls[l]} f_{ls'[l]}$ is the covariance between $x_{tls[l]}$ and $x_{tls'[l]}$ for the distinct alleles $s[l]$ and $s'[l]$ in an outbred population with the same allele frequencies. The variance of $x_{tls[l]}$ in an outbred population is $\xi_{l:s[l],s[l]} = 2 f_{ls[l]}(1 - f_{ls[l]})$. The outbred allele count variances and covariances are equivalent to those for a multinomial distribution with two trials and with event probabilities equal to the allele frequencies at the locus.

Therefore,

$$\mathrm{Cov}(X_{tL}, X_{tL'}) = 2^{|L|} \left(\frac{K_{t,u} - K_0}{1 - K_0} \right)^{|L|} \xi_L, \tag{A.20}$$

where ξ_L is the variance of X_L in an outbred population, and is equal to

$$\sum_{s \in S_L} \sum_{s' \in S_L} \beta_{Ls} \beta_{Ls'} \prod_{l \in L} \xi_{l:s[l],s'[l]}. \tag{A.21}$$

Therefore,

$$\mathrm{Cov}(G_t, G_u) = \sum_{\tau=1}^{|N|} 2^\tau \left(\frac{K_{t,u} - K_0}{1 - K_0} \right)^\tau v_\tau, \tag{A.22}$$

where v_τ is the variance from genetic interactions between τ loci in an outbred population, and is the sum of ξ_L over all subsets L of size τ.

Using the fact that $K_{tt} = (1 + F_t)/2$, where F_t is the inbreeding coefficient of individual t, and setting $t = u$ gives

$$\mathrm{Var}(G_t) = \sum_{\tau=1}^{|N|} \left(1 + \frac{F_t - K_0}{1 - K_0} \right)^\tau v_\tau. \tag{A.23}$$

The population variance is derived by the law of total variance,

$$Var(G) = \mathbb{E}_t[\mathrm{Var}(G_t)] + \mathrm{Var}_t(\mathbb{E}[G_t]). \tag{A.24}$$

Because there is no dominance, the phenotypic mean does not change with inbreeding. In a random-mating population, the mean inbreeding coefficient is equal to the mean kinship coefficient: $\mathbb{E}_t(F_t) = K_0$. Therefore,

$$\begin{aligned} \mathrm{Var}(G) = \mathbb{E}_t[\mathrm{Var}(G_t)] = v_1 + \left(1 + \frac{\mathrm{Var}(F_t)}{(1 - K_0)^2} \right) v_2 + \\ \sum_{\tau=3}^{|N|} \left(1 + \binom{\tau}{2} \frac{\mathrm{Var}(F_t)}{(1 - K_0)^2} + \sum_{i=3}^{\tau} \binom{\tau}{i} \frac{\mathbb{E}[(F_t - K_0)^i]}{(1 - K_0)^i} \right) v_\tau. \end{aligned} \tag{A.25}$$

A.1.2 Dominance variance in a finite population

We consider a phenotype whose genetic contribution is determined by two bi-allelic loci in linkage equilibrium that have non-zero dominance deviations as well as an interaction between their additive effects. The phenotype of an individual s, Y_s, is the sum of the additive contributions of the loci, a_s, the interaction between the additive effects, $(a \times a)_s$, and the sum of the dominance deviations of the loci d_s,

giving

$$Y_s = a_s + (a \times a)_s + d_s + \epsilon_s; \text{ where } d_s = \delta_1 \gamma_{s1m} \gamma_{s1p} + \delta_2 \gamma_{s2m} \gamma_{s2p}, \tag{A.26}$$

and $\gamma_{sim} = g^m_{si} - f_i$ is the mean normalised indicator variable for the presence of the minor allele at locus i, with frequency f_i, on the maternal chromosome, and $\gamma_{sip} = g^p_{si} - f_i$ is the corresponding variable for the paternal chromosome.

The additive and additive-by-additive components are orthogonal, as shown above. The additive-by-additive and the dominance components are orthogonal, because the dominance deviation of each locus is uncorrelated with the additive effect of the other locus. Inbreeding, however, induces a correlation between the additive effect and dominance deviation at a locus, implying that

$$\text{Var}(Y_s) = \text{Var}(a_s) + \text{Var}((a \times a)_s) + \text{Var}(d_s) + \text{Cov}(a_s, d_s). \tag{A.27}$$

The additive and additive-by-additive variance components are as derived above.

$\text{Var}(d_s)$ is derived by applying the law of total variance to $d_{si} = \delta_i \gamma_{sim} \gamma_{sip}$, the contribution of locus $i \in \{1, 2\}$ to d_s.

$$\text{Var}(d_{si}) = \mathbb{E}_{I_{si}}[\text{Var}(d_{si}|I_{si})] + \text{Var}_{I_{si}}(\mathbb{E}[d_{si}|I_{si}]), \tag{A.28}$$

where I_{si} is the indicator variable for whether individual s is inbred or not at locus i. The conditional expectation of d_{si} depends on

$$\mathbb{E}[\gamma_{sim}\gamma_{sip}|I_{si}] = I_{si}(f_i(1 - f_i) - \mathbb{E}[\gamma_{sim}\gamma_{sip}|I_{si} = 0]) + \mathbb{E}[\gamma_{sim}\gamma_{sip}|I_{si} = 0]. \tag{A.29}$$

Using the expression for the genotypic covariance in a founder population derived in

(2.13),

$$\mathbb{E}[\gamma_{sim}\gamma_{sip}|I_{si}=0] = \frac{-K_0}{1-K_0} f_i(1-f_i), \tag{A.30}$$

where K_0 is the mean inbreeding (and kinship) coefficient. Therefore,

$$f_i(1-f_i) - \mathbb{E}[\gamma_{sim}\gamma_{sip}|I_{si}=0] = \frac{f_i(1-f_i)}{1-K_0}, \tag{A.31}$$

and therefore

$$\mathrm{Var}_{I_{si}}(\mathbb{E}[d_{si}|I_{si}]) = \frac{F_s(1-F_s)}{(1-K_0)^2}\mu_{hi}^2, \tag{A.32}$$

where $\mu_{hi} = \delta_i f_i(1-f_i)$ is the inbreeding depression at locus i.

We now calculate $\mathbb{E}_{I_{si}}[\mathrm{Var}(d_{si}|I_{si})]$. First, in the inbred case,

$$\mathrm{Var}(d_{si}|I_{si}=1) = \delta_i^2 f_i(1-f_i)(1-2f_i)^2 = v_{hi}, \tag{A.33}$$

which is the dominance variance at locus i in the homozygous population. When there is no inbreeding at locus i,

$$\mathrm{Var}(d_{si}|I_{si}=0) = \delta_i^2(\mathbb{E}[\gamma_{sim}^2\gamma_{sip}^2|I_{si}=0] - \mathbb{E}[\gamma_{sim}\gamma_{sip}|I_{si}=0]^2) \tag{A.34}$$

By expanding the squares in $\gamma_{sim}^2\gamma_{sip}^2$ and using the result for the genotypic covariance when there is no IBD sharing, it can be shown that

$$\mathbb{E}[\gamma_{sim}^2\gamma_{sip}^2|I_{si}=0] = f_i^2(1-f_i)^2 - \frac{K_0}{1-K_0}(1-2f_i)^2 f_i(1-f_i). \tag{A.35}$$

Using the result for genotypic covariance when there is no IBD again gives

$$\mathbb{E}[\gamma_{sim}\gamma_{sip}|I_{si}=0]^2 = \frac{K_0^2}{(1-K_0)^2} f_i^2(1-f)^2. \tag{A.36}$$

Therefore,

$$\mathrm{Var}(d_{si}|I_{si}=0)=\mu_{hi}^2\left(1-\frac{K_0^2}{(1-K_0)^2}-\frac{K_0}{1-K_0}\frac{(1-2f_i)^2}{f_i(1-f_i)}\right)=v_{\delta_i}, \tag{A.37}$$

which we have defined to be $v_{\delta i}$.

Combining the results gives

$$\mathrm{Var}(d_{si})=(1-F_s)v_{\delta i}+F_s v_{hi}+\frac{F_s(1-F_s)}{(1-K_0)^2}\mu_{hi}^2, \tag{A.38}$$

where F_s is the inbreeding coefficient of individual s. Summing across the loci gives

$$\mathrm{Var}(d_s)=(1-F_s)v_{\delta}+F_s v_{h}+\frac{F_s(1-F_s)}{(1-K_0)^2}SS_{\mu_h}, \tag{A.39}$$

where $v_\delta=v_{\delta 1}+v_{\delta 2}$, $v_h=v_{h1}+v_{h2}$, and $SS_{\mu_h}=\mu_{h1}^2+\mu_{h2}^2$ is the sum of the squared inbreeding depressions at the loci.

In a founder population, the maternal and paternal alleles are not independent, which implies that

$$\mathrm{Cov}(\gamma_{s1m},\gamma_{s1m}\gamma_{s1p})=\mathbb{E}[\gamma_{s1m}^2\gamma_{s1p}]\neq 0. \tag{A.40}$$

This implies that there is covariance between an individual's additive effect and dominance deviation, depending on their inbreeding coefficient. The above expectation can be evaluated by conditioning on whether or not individual s is inbred or not, giving

$$\mathrm{Cov}(\gamma_{s1m},\gamma_{s1m}\gamma_{s1p})=f_1(1-f_1)(1-2f_1)\frac{F_s-K_0}{1-K_0}. \tag{A.41}$$

Summing the contributions of the four possible covariances within the locus and

summing across loci gives

$$\mathrm{Cov}(a_s, d_s) = 4\frac{F_s - K_0}{1 - K_0} C_{a,d}, \tag{A.42}$$

where $C_{a,d} = \sum_{i=1}^{2} \beta_i \delta_i f_i (1 - f_i)(1 - 2f_i)$ parameterises the strength of the covariance between additive and dominance effects.

Combining these results with those from (2.20):

$$\mathrm{Var}(Y_s) = \sum_{\tau=1}^{2} \left(1 + \frac{F_t - K_0}{1 - K_0}\right)^{\tau} v_\tau + (1 - F_s) v_\delta + \tag{A.43}$$

$$4\frac{F_s - K_0}{1 - K_0} C_{a,d} + F_s v_h + \frac{F_s(1 - F_s)}{(1 - K_0)^2} SS_{\mu_h} + \sigma^2_\epsilon; \tag{A.44}$$

where v_1 and v_2 are as defined in (2.18) and (2.19). v_δ is the covariance between two individuals' dominance deviations conditional on both alleles of one individual being IBD to distinct alleles of the other individual, implying that neither individual is inbred. v_δ differs slightly from the dominance variance in an infinite, outbred population, where $v_{\delta 1} = \mu^2_{h1}$. Similarly, $v_\delta \approx SS_{\mu_h}$. These difference will be very small apart from for populations descending from a very small number of founders (large K_0), such as in certain cross designs.

The variance in the population is, by the Law of Total Variance,

$$\mathrm{Var}(Y) = \mathbb{E}_s[\mathrm{Var}(Y_s)] + \mathrm{Var}_s(\mathbb{E}[Y_s]). \tag{A.45}$$

Because the mean inbreeding coefficient is K_0 in an outbred population,

$$\mathbb{E}_s[\mathrm{Var}(Y_s)] = v_1 + \left(1 + \frac{\mathrm{Var}(F)}{(1 - K_0)^2}\right) v_2 + (1 - K_0) v_\delta + K_0 v_h \tag{A.46}$$

$$+ \frac{K_0(1 - K_0) - \mathrm{Var}(F)}{(1 - K_0)^2} SS_{\mu_h} + \sigma^2_\epsilon. \tag{A.47}$$

The expectation of Y_s only depends on the expectation of d_s, the others being zero.

$$\mathbb{E}[d_{s1}] = \frac{F_s}{1-K_0}\mu_{h1} + C, \tag{A.48}$$

where C is a constant that does not depend on s. Therefore,

$$\mathbb{E}[d_s] = \frac{F_s}{1-K_0}\mu_h + 2C, \tag{A.49}$$

where $\mu_h = \mu_{h1} + \mu_{h2}$ is the inbreeding depression of the phenotype. Therefore

$$\mathrm{Var}_s(\mathbb{E}[Y_s]) = \frac{\mathrm{Var}(F)}{(1-K_0)^2}\mu_h^2 \tag{A.50}$$

Therefore,

$$\mathrm{Var}(Y) = v_1 + \left(1 + \frac{\mathrm{Var}(F)}{(1-K_0)^2}\right) v_2 + (1-K_0)v_\delta + K_0 v_h + \tag{A.51}$$

$$\frac{K_0}{(1-K_0)} SS_{\mu_h} + \frac{\mathrm{Var}(F)}{(1-K_0)^2}\left(\mu_h^2 - SS_{\mu_h}\right) + \sigma_\epsilon^2. \tag{A.52}$$

A.2 Asymptotic variance of fitting a quadratic

We extend the analogy of fitting a quadratic to derive an analytic approximation of the standard error of the estimator of the variance from pairwise interactions. We imagine fitting the off diagonal elements of the covariance matrix as a quadratic function of $R_{s,t} = 2(K_{s,t} - K_0)/(1-K_0)$, with normal error:

$$\Sigma_{st} \sim N(v_1 R_{st} + v_2 R_{st}^2, \sigma^2), \tag{A.53}$$

for all $\eta = N(N-1)/2$ pairs st. This assumes that the off diagonal elements of the covariance matrix are independent, which may be problematic for samples which contain large sets of closely related individuals. The homoscedasticity assumption

could also be problematic when there are many levels of relatedness present in the sample.

We invert the information matrix to obtain the asymptotic error of the maximum likelihood estimator of v_2. If we define $e_i = \Sigma_i - v_1 R_i - v_2 R_i^2$ to be the i^{th} residual, then the matrix of the second derivatives of the log-likelihood is

$$H = -\sigma^{-2} \begin{bmatrix} S_{R^2} & S_{R^3} & \sigma^{-2} \sum_{i=1}^{\eta} e_i R_i \\ S_{R^3} & S_{R^4} & \sigma^{-2} \sum_{i=1}^{\eta} e_i R_i^2 \\ \sigma^{-2} \sum_{i=1}^{\eta} e_i R_i & \sigma^{-2} \sum_{i=1}^{\eta} e_i R_i^2 & -2\sigma^{-4} \sum_{i=1}^{\eta} e_i^2 \end{bmatrix}, \tag{A.54}$$

where $S_{R^c} = \sum_{i=1}^{\eta} R_i^c$.

We now take the negative expectation of H to obtain the Fisher information matrix. Because $R_i = 2(K_i - K_0)/(1 - K_0)$, $\mathbb{E}[R] = 0$. Therefore $\mathbb{E}[S_{R^c}] = \eta \mu_c$, where μ_c is the c^{th} central moment of the distribution of R.

Therefore, the information matrix is

$$\mathbb{I} = \eta \sigma^{-2} \begin{bmatrix} \mathrm{Var}(R) & \mu_3 & 0 \\ \mu_3 & \mu_4 & 0 \\ 0 & 0 & \sigma^{-2} \end{bmatrix}. \tag{A.55}$$

Using elimination to invert the matrix gives

$$\mathbb{I}^{-1} = \frac{\sigma^2}{\eta} \begin{bmatrix} \frac{1}{\mathrm{Var}(R)} + \frac{\mu_3^2}{\mu_4 - \mu_3/\mathrm{Var}(R)} & \frac{-\mu_3}{\mathrm{Var}(R)\mu_4 - \mu_3^2} & 0 \\ \frac{-\mu_3}{\mathrm{Var}(R)\mu_4 - \mu_3^2} & \left(\mu_4 - \frac{\mu_3^2}{\mathrm{Var}(R)}\right)^{-1} & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix} \tag{A.56}$$

This implies that the asymptotic standard error of the estimate of the variance

from pairwise interactions is

$$\frac{\sigma}{\sqrt{\eta(\mu_4 - \mu_3^2/\mathrm{Var}(R))}} \tag{A.57}$$

If the phenotype has been normalised to have variance one, then σ should be approximately 1.

A.3 Tables

K_0	1/240	1/120	1/60	1/30
$\hat{v_1}$	0.41	0.41	0.41	0.41

Table A.1: **Bias in estimation of additive variance as a function of kinship.** The mean estimates of the additive variance, v_1, for populations with different mean kinship, K_0. The true value of v_1 is 0.4. The variance from pairwise interactions, v_2, is 0.2, leading to a slight upward bias in the estimates of v_1. The bias does not depend on the mean kinship for these populations.

Trait	h^2	s.e.	h_2^2	s.e.	Z_2	$h_>^2$	s.e.	$Z_>$	H^2
Cadmium Chloride	0.83	0.03	0.00	0.04	0.00	0.50	0.05	2.81	0.96
Maltose	0.72	0.03	0.02	0.04	0.45	0.34	0.05	3.85	0.94
Tunicamycin	0.73	0.03	0.09	0.04	2.15	0.01	0.05	1.75	0.91
Zeocin	0.54	0.04	0.04	0.07	0.59	0.26	0.08	4.14	0.90
4-Hydroxybenzaldehyde	0.52	0.04	0.08	0.07	1.07	0.11	0.08	3.55	0.89
Lactose	0.62	0.04	0.09	0.06	1.57	0.04	0.07	2.69	0.89
YPD:4C	0.51	0.05	0.19	0.09	2.16	0.01	0.10	1.77	0.88
Lithium Chloride	0.77	0.03	0.11	0.04	2.74	0.00	0.05	0.00	0.87
Manganese Sulfate	0.42	0.04	0.32	0.08	4.08	0.00	0.09	1.48	0.87
4NQO	0.75	0.03	0.05	0.04	1.39	0.07	0.05	1.13	0.86
Neomycin	0.69	0.03	0.02	0.05	0.46	0.32	0.06	2.59	0.86
Cobalt Chloride	0.55	0.04	0.07	0.07	1.13	0.11	0.08	2.95	0.85
YPD:15C	0.47	0.04	0.06	0.07	0.79	0.22	0.08	3.78	0.84
Lactate	0.60	0.04	0.10	0.06	1.67	0.04	0.07	1.95	0.84
YPD:37C	0.63	0.04	0.09	0.06	1.64	0.04	0.07	1.64	0.83
SDS	0.52	0.05	0.10	0.08	1.31	0.09	0.10	1.98	0.81
Indoleacetic Acid	0.41	0.04	0.13	0.08	1.62	0.04	0.09	2.91	0.81
YNB	0.58	0.04	0.09	0.06	1.41	0.06	0.07	1.80	0.80
Paraquat	0.58	0.04	0.03	0.06	0.58	0.28	0.07	2.53	0.79
5-Fluorouracil	0.67	0.03	0.12	0.05	2.14	0.01	0.07	0.00	0.78
YNB:ph8	0.47	0.04	0.14	0.08	1.76	0.03	0.09	1.80	0.78
Copper	0.38	0.04	0.23	0.08	2.66	0.00	0.10	1.81	0.78
Trehalose	0.54	0.04	0.16	0.07	2.40	0.00	0.08	0.81	0.77
E6 Berbamine	0.53	0.04	0.16	0.07	2.36	0.00	0.08	0.90	0.77
YPD	0.60	0.04	0.08	0.06	1.38	0.07	0.07	1.07	0.76
Xylose	0.49	0.04	0.17	0.08	2.25	0.01	0.09	1.15	0.76
Formamide	0.38	0.05	0.18	0.08	2.20	0.01	0.10	2.14	0.76
Ethanol	0.49	0.04	0.17	0.08	2.26	0.01	0.09	0.93	0.75
Hydroxyurea	0.48	0.05	0.27	0.08	3.29	0.00	0.10	0.00	0.74
Cycloheximide	0.59	0.04	0.07	0.06	1.20	0.10	0.07	0.80	0.72
Congo red	0.61	0.04	0.03	0.06	0.47	0.32	0.07	1.03	0.71
Sorbitol	0.45	0.05	0.19	0.13	1.46	0.06	0.15	0.38	0.70
5-Fluorocytosine	0.54	0.04	0.10	0.07	1.44	0.06	0.09	0.67	0.70
Diamide	0.51	0.04	0.00	0.07	0.00	0.48	0.08	2.35	0.70
Menadione	0.43	0.04	0.10	0.08	1.34	0.07	0.09	1.34	0.65
Hydrogen Peroxide	0.51	0.05	0.12	0.09	1.29	0.09	0.11	0.00	0.63
Raffinose	0.47	0.05	0.04	0.11	0.40	0.35	0.13	0.91	0.63
Galactose	0.28	0.04	0.15	0.10	1.57	0.05	0.11	1.74	0.63
Hydroquinone	0.24	0.04	0.00	0.10	0.00	0.50	0.11	3.31	0.60
Magnesium Chloride	0.29	0.04	0.01	0.09	0.11	0.47	0.10	2.55	0.57
6-Azauracil	0.35	0.04	0.00	0.08	0.00	0.48	0.09	2.22	0.56
Cisplatin	0.33	0.04	0.03	0.09	0.35	0.34	0.10	1.25	0.49
Calcium Chloride	0.32	0.04	0.13	0.09	1.39	0.07	0.11	0.20	0.47
Caffeine	0.25	0.04	0.18	0.10	1.91	0.01	0.11	0.11	0.44
Mannose	0.24	0.04	0.07	0.10	0.70	0.24	0.12	0.94	0.42
YNB:ph3	0.18	0.04	0.00	0.10	0.00	0.49	0.11	1.95	0.40

Table A.2: **Numerical estimates of heritability components.** The table gives the estimates of the heritability components followed by their standard errors in the right adjacent column. Z_2 is the estimate of h_2^2 divided by its estimated standard error; $Z_>$ is the estimate of $h_>^2$ divided by its estimated standard error.

Appendix B

B.1 Computation of the derivative with respect to the variance parameters

To compute the derivative with respect to the variance parameters, we use the method of differentials[181] to compute the infinitesimal change in the log-likelihood, dL, with respect to infinitesimal changes in λ or β.

We illustrate this with an example. For a scalar function, L, of a column vector, x, one computes the infinitesimal change in L, dL, with an infinitesimal change in a the vector, dx:

$$dL = \frac{\partial L}{\partial x^T} dx, \tag{B.1}$$

which corresponds to the linear term in the Taylor expansion of L:

$$L = L(0) + \frac{\partial L}{\partial x^T} dx + \dots \tag{B.2}$$

Note that the infinitesimal change, dx, is of the same dimension as x, whereas $\frac{\partial L}{\partial x^T}$ has dimension equal to x^T.

In deriving the differentials with respect to the variance parameters, we utilise the

differential formulae:

$$d\Lambda^{-1} = -\Lambda^{-1}d\Lambda\Lambda^{-1},\ d\log|\Lambda| = \mathrm{tr}(\Lambda^{-1}d\Lambda). \tag{B.3}$$

B.1.1 Derivative with respect to λ

We compute the infinitesimal change in L with respect to an infinitesimal change in λ, $d\lambda$. The differential of the log-likelihood with respect to λ relies upon the differential of Λ with respect to λ:

$$dL = -\sum_{j=1}^{l} W_j d\lambda - \mathrm{tr}(\Lambda^{-1}d\Lambda) - r^T\Lambda^{-1}(d\Lambda)\Lambda^{-1}r. \tag{B.4}$$

The differential of Λ with respect to λ is

$$d\Lambda = dH^{-1} = -H^{-1}\mathrm{diag}(Wd\lambda) \tag{B.5}$$

Therefore,

$$\begin{aligned} -\mathrm{tr}(\Lambda^{-1}d\Lambda) &= \mathrm{tr}(\Lambda^{-1}H^{-1}\mathrm{diag}(Wd\lambda)) \\ &= \sum_{j=1}^{l} \Lambda_{jj}^{-1}\exp(-W_j\lambda)W_j d\lambda. \end{aligned} \tag{B.6}$$

For the other component of the differential, we have

$$-r^T\Lambda^{-1}(d\Lambda)\Lambda^{-1}r = r^T\Lambda^{-1}H^{-1}\mathrm{diag}(Wd\lambda)\Lambda^{-1}r \tag{B.7}$$

$$= \mathrm{tr}(\Lambda^{-1}rr^T\Lambda^{-1}H^{-1}\mathrm{diag}(Wd\lambda)). \tag{B.8}$$

Let $\Gamma = \Lambda^{-1} r r^T \Lambda^{-1}$, then

$$- r^T \Lambda^{-1} (d\Lambda) \Lambda^{-1} r = \sum_{j=1}^{l} \Gamma_{jj} \exp(-W_j \lambda) W_j d\lambda. \tag{B.9}$$

Therefore,

$$dL = - \sum_{j=1}^{l} W_j d\lambda + \sum_{j=1}^{l} \Lambda_{jj}^{-1} \exp(-W_j \lambda) W_j d\lambda + \sum_{j=1}^{l} \Gamma_{jj} \exp(-W_j \lambda) W_j d\lambda. \tag{B.10}$$

Therefore,

$$\frac{\partial L}{\partial \lambda^T} = \sum_{j=1}^{l} [(\Lambda_{jj}^{-1} + \Gamma_{jj}) \exp(-W_j \lambda) - 1] W_j. \tag{B.11}$$

B.1.2 Derivative with respect to β

To aid differentiation, we rewrite Λ to make its reliance on β explicit:

$$\Lambda = H^{-1} + Z^T D^{-1} Z = H^{-1} + \sum_{i=1}^{n} Z_i^T Z_i \exp(-V_i \beta), \tag{B.12}$$

where Z_i is the i^{th} $[1 \times l]$ row vector of Z.

We also rewrite r to make its dependence on β explicit.

$$r = \sum_{i=1}^{n} Z_i^T (y_i - X_i \alpha) \exp(-V_i \beta). \tag{B.13}$$

The differential of the likelihood with respect to a change in β is

$$dL = - \sum_{i=1}^{n} V_i d\beta + \sum_{i=1}^{n} (y_i - X_i \alpha)^2 \exp(-V_i \beta) V_i d\beta - \tag{B.14}$$
$$\text{tr}(\Lambda^{-1} d\Lambda) + d(r^T \Lambda^{-1} r).$$

The differential of Λ with respect to β is

$$d\Lambda = -\sum_{i=1}^{n} Z_i^T Z_i \exp(-V_i\beta) V_i d\beta. \tag{B.15}$$

It can therefore be shown that,

$$-\mathrm{tr}(\Lambda^{-1} d\Lambda) = \sum_{i=1}^{n} Z_i \Lambda^{-1} Z_i^T \exp(-V_i\beta) V_i d\beta. \tag{B.16}$$

We use the fact that

$$d(r^T \Lambda^{-1} r) = 2r^T \Lambda^{-1} dr - r^T \Lambda^{-1} d\Lambda \Lambda^{-1} r \tag{B.17}$$

to derive the differential of the inner product $r^T \Lambda^{-1} r$.

The differential of r with respect to β is

$$dr = -\sum_{i=1}^{n} Z_i^T (y_i - X_i\alpha) \exp(-V_i\beta) V_i d\beta. \tag{B.18}$$

The differential of Λ^{-1} is

$$\begin{aligned} d\Lambda^{-1} &= -\Lambda^{-1} d\Lambda \Lambda^{-1} \\ &= \sum_{i=1}^{n} \Lambda^{-1} Z_i^T Z_i \Lambda^{-1} \exp(-V_i\beta) V_i d\beta. \end{aligned} \tag{B.19}$$

It can then be shown that

$$d(r^T \Lambda^{-1} r) = \sum_{i=1}^{n} r^T \Lambda^{-1} Z_i^T \left(Z_i \Lambda^{-1} r - 2(y_i - X_i\alpha) \right) \exp(-V_i\beta) V_i d\beta. \tag{B.20}$$

This can be calculated efficiently by realising that $Z_i \Lambda^{-1} r = r^T \Lambda^{-1} Z_i^T$, and that this

is the i^{th} element of the vector

$$a = Z\Lambda^{-1}[Z^T D^{-1}(y - X\alpha)]. \tag{B.21}$$

Therefore, the differential is

$$d(r^T\Lambda^{-1}r) = \sum_{i=1}^{n} a_i \left(a_i - 2(y_i - X_i\alpha)\right) \exp(-V_i\beta)V_i d\beta \tag{B.22}$$

Therefore,

$$\begin{aligned} dL = &- \sum_{i=1}^{n} V_i d\beta + \sum_{i=1}^{n} (y_i - X_i\alpha)^2 \exp(-V_i\beta)V_i d\beta + \\ &\sum_{i=1}^{n} Z_i\Lambda^{-1}Z_i^T \exp(-V_i\beta)V_i d\beta + a_i \left(a_i - 2(y_i - X_i\alpha)\right) \exp(-V_i\beta)V_i d\beta \end{aligned} \tag{B.23}$$

Therefore,

$$\frac{\partial L}{\partial \beta^T} = \sum_{i=1}^{n} \{(y_i - X_i\alpha)^2 + Z_i\Lambda^{-1}Z_i^T + a_i \left(a_i - 2(y_i - X_i\alpha)\right)\} \exp(-V_i\beta)V_i - \sum_{i=1}^{n} V_i. \tag{B.24}$$

Let

$$k_i = (y_i - X_i\alpha)^2 + Z_i\Lambda^{-1}Z_i^T + a_i \left(a_i - 2(y_i - X_i\alpha)\right), \tag{B.25}$$

then

$$\frac{\partial L}{\partial \beta^T} = \sum_{i=1}^{n} (k_i \exp(-V_i\beta) - 1)V_i. \tag{B.26}$$

Appendix C

C.1 Tables

					British		Diverse	
rsid	**genes**	**chr**	**pos. (bp)**	**maf**	**est.**	**s.e.**	**est.**	**s.e.**
Novel								
rs2814992	UHRF1BP1, SNRPC2, C6orf1062	6	34,617,144	0.33	0.028	0.0045	0.016	0.009
rs1652376	NPC12 C18orf82	18	21,109,466	0.46	0.023	0.0043	0.029	0.008
rs61587156	IGF2BP2	3	185,831,583	0.19	-0.032	0.0054	-0.025	0.01
rs957919	CPNE42	3	131,629,716	0.28	-0.026	0.0048	-0.018	0.009
rs6831020	KIT	4	55,500,226	0.29	0.027	0.0046	0.023	0.009
rs9469488	BAK12	6	33,517,117	0.27	-0.023	0.0048	-0.034	0.009
rs4441044	FGF19, ORAOV1	11	69,500,363	0.35	-0.024	0.0044	-0.02	0.008
rs2785980	SLC30A102	1	219,700,519	0.34	0.016	0.0043	0.018	0.009
Known								
rs1421085	FTO	16	53,800,954	0.4	0.083	0.0043	0.064	0.008
rs6548238	TMEM18	2	634,905	0.17	0.06	0.0056	0.057	0.011
rs1800437	GIPR1	19	46,181,392	0.19	0.037	0.0053	0.048	0.01
rs6265	BDNF1	11	27,679,916	0.19	0.043	0.0054	0.015	0.01
rs7903146	TCF7L2	10	114,758,349	0.29	-0.026	0.0047	-0.016	0.009

Table C.1: **Estimated additive effects.** The estimated additive effects and their inflation adjusted standard errors in the British and diverse subsamples for loci passing genome-wide significance and fitting an additive-variance model better than an additive model.

					British		Diverse	
rsid	**genes**	**chr**	**pos. (bp)**	**maf**	**est.**	**s.e.**	**est.**	**s.e.**
Novel								
rs2814992	UHRF1BP1, SNRPC2, C6orf1062	6	34,617,144	0.33	0.017	0.0074	0.0393	0.0136
rs1652376	NPC12 C18orf82	18	21,109,466	0.46	0.026	0.007	0.0102	0.0129
rs61587156	IGF2BP2	3	185,831,583	0.19	-0.031	0.0089	-0.0045	0.0165
rs957919	CPNE42	3	131,629,716	0.28	-0.027	0.0077	-0.0326	0.0141
rs6831020	KIT	4	55,500,226	0.29	0.027	0.0076	0.0141	0.0141
rs9469488	BAK12	6	33,517,117	0.27	-0.028	0.0077	0	0.0143
rs4441044	FGF19, ORAOV1	11	69,500,363	0.35	-0.026	0.0073	0	0.0134
rs2785980	SLC30A102	1	219,700,519	0.34	0.033	0.0073	0.0368	0.0135
Known								
rs1421085	FTO	16	53,800,954	0.4	0.05	0.007	0.0594	0.0131
rs6548238	TMEM18	2	634,905	0.17	0.049	0.0092	-0.0008	0.017
rs1800437	GIPR1	19	46,181,392	0.19	0.033	0.0088	0.0353	0.0161
rs6265	BDNF1	11	27,679,916	0.19	0.022	0.0089	0.0417	0.0167
rs7903146	TCF7L2	10	114,758,349	0.29	-0.037	0.0077	0.0111	0.0141

Table C.2: **Estimated log-linear variance effects.** The estimated log-linear variance effects and their inflation adjusted standard errors in the British and diverse subsamples for loci passing genome-wide significance and fitting an additive-variance model better than an additive model.

					British		
rsid	**genes**	**chr**	**pos. (bp)**	**maf**	**add**	**llv**	**av**
novel							
rs2814992	UHRF1BP1, SNRPC2, C6orf1062	6	34,617,144	0.33	8.74	1.66	8.99
rs1652376	NPC12 C18orf82	18	21,109,466	0.46	6.41	3.59	8.49
rs61587156	IGF2BP2	3	185,831,583	0.19	7.8	3.24	9.51
rs957919	CPNE42	3	131,629,716	0.28	7	3.22	8.71
rs6831020	KIT	4	55,500,226	0.29	7.28	3.28	9.04
rs9469488	BAK12	6	33,517,117	0.27	5.39	3.55	7.47
rs4441044	FGF19, ORAOV1	11	69,500,363	0.35	6.49	3.55	8.53
rs2785980	SLC30A102	1	219,700,519	0.34	3.08	5.17	6.83
known							
rs1421085	FTO	16	53,800,954	0.4	74.52	11.71	83.91
rs6548238	TMEM18	2	634,905	0.17	23.33	8.81	28.21
rs1800437	GIPR1	19	46,181,392	0.19	10.08	3.65	12.12
rs6265	BDNF1	11	27,679,916	0.19	13.22	1.8	13.51
rs7903146	TCF7L2	10	114,758,349	0.29	6.78	5.65	10.82

Table C.3: **Association statistics in the British subsample.** Negative log (base 10) p-values of the loci passing genome-wide significance and fitting an additive-variance model better than an additive model. The p-values for additive effects (add), log-linear variance effects (llv), and for the additive-variance test (av) are given for the British subsample.

					Diverse		
rsid	**genes**	**chr**	**pos. (bp)**	**maf**	**add**	**llv**	**av**
novel							
rs2814992	UHRF1BP1, SNRPC2, C6orf1062	6	34,617,144	0.33	1.15	2.4	2.51
rs1652376	NPC12 C18orf82	18	21,109,466	0.46	3.33	0.37	2.8
rs61587156	IGF2BP2	3	185,831,583	0.19	1.78	0.11	1.27
rs957919	CPNE42	3	131,629,716	0.28	1.32	1.67	2
rs6831020	KIT	4	55,500,226	0.29	2	0.5	1.67
rs9469488	BAK12	6	33,517,117	0.27	3.64	0.02	2.95
rs4441044	FGF19, ORAOV1	11	69,500,363	0.35	1.79	0.07	1.27
rs2785980	SLC30A102	1	219,700,519	0.34	1.39	2.2	2.53
known							
rs1421085	FTO	16	53,800,954	0.4	13.9	5.29	17.43
rs6548238	TMEM18	2	634,905	0.17	6.84	0.02	6.01
rs1800437	GIPR1	19	46,181,392	0.19	5.43	1.54	5.69
rs6265	BDNF1	11	27,679,916	0.19	0.82	1.9	1.8
rs7903146	TCF7L2	10	114,758,349	0.29	1.14	0.36	0.84

Table C.4: **Association statistics in the diverse subsample.** Negative log (base 10) p-values of the loci passing genome-wide significance and fitting an additive-variance model better than an additive model. The p-values for additive effects (add), log-linear variance effects (llv), and for the additive-variance test (av) are given for the diverse subsample.

					Combined		
rsid	**genes**	**chr**	**pos. (bp)**	**maf**	**add**	**llv**	**av**
novel							
rs2814992	UHRF1BP1, SNRPC2, C6orf1062	6	34,617,144	0.33	8.56	2.94	10.06
rs1652376	NPC12 C18orf82	18	21,109,466	0.46	8.25	3.03	9.85
rs61587156	IGF2BP2	3	185,831,583	0.19	8.19	2.59	9.36
rs957919	CPNE42	3	131,629,716	0.28	7	3.7	9.3
rs6831020	KIT	4	55,500,226	0.29	7.88	2.83	9.3
rs9469488	BAK12	6	33,517,117	0.27	7.56	2.86	9.02
rs4441044	FGF19, ORAOV1	11	69,500,363	0.35	6.93	2.87	8.43
rs2785980	SLC30A102	1	219,700,519	0.34	3.33	6.03	8
known							
rs1421085	FTO	16	53,800,954	0.4	86.04	15.27	98.94
rs6548238	TMEM18	2	634,905	0.17	28.23	5.98	32.32
rs1800437	GIPR1	19	46,181,392	0.19	13.8	4	16.18
rs6265	BDNF1	11	27,679,916	0.19	12.69	2.62	13.75
rs7903146	TCF7L2	10	114,758,349	0.29	6.65	5	10.21

Table C.5: **Combined association statistics.** Negative log (base 10) p-values of the loci passing genome-wide significance and fitting an additive-variance model better than an additive model. The p-values from combining evidence from the two subsamples are given for additive effects (add), log-linear variance effects (llv), and for the additive-variance test (av). The p-values are in Table C.3 for the British subsample and Table C.4 for the diverse subsample.

C.2 Figures

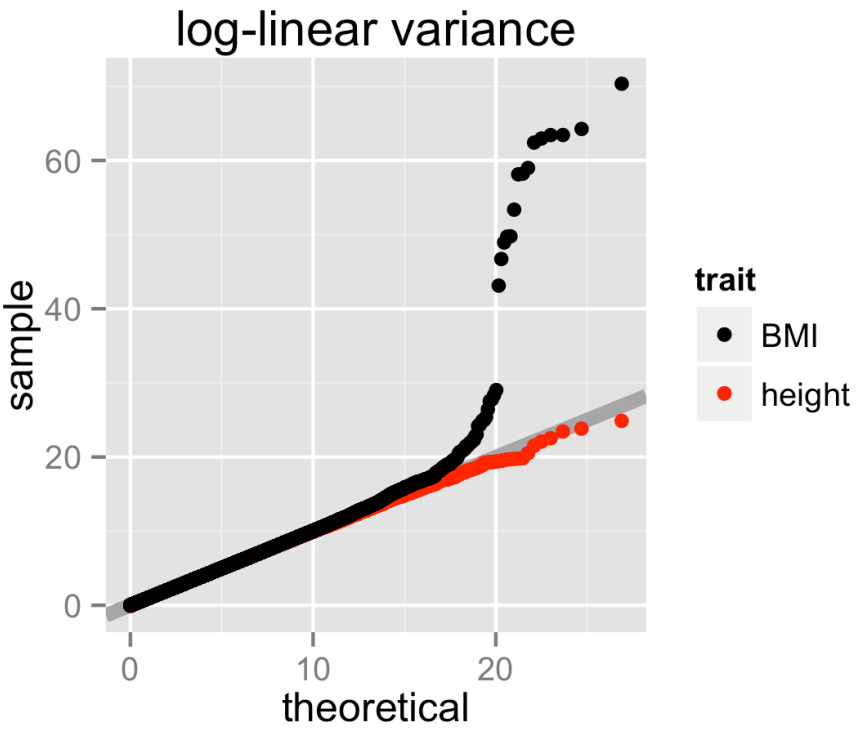


Figure C.1: **Comparison of height and BMI log-linear variance test statistics.** Comparison of the QQ-plots for the inflation adjusted log-linear variance test statistics for log-height (red) and log-BMI (black) when compared to the asymptotic null distribution, which is a Chi-Square distribution of appropriate degrees of freedom.

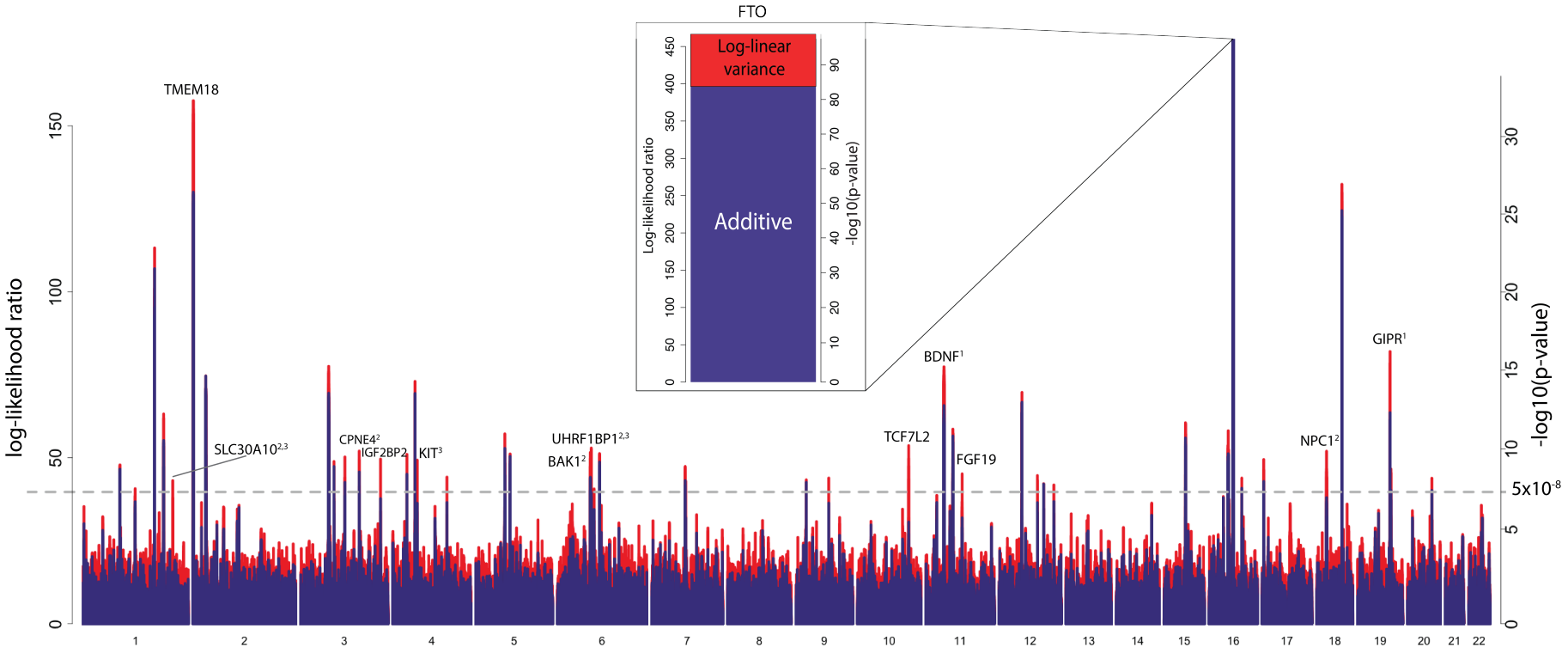


Figure C.2: **'Manhattan Sunset' plot visualising the genome-wide additive and log-linear variance test statistics for log-BMI**. Names were added for loci that passed genome-wide-significance ($p = 5 \times 10^{-8}$) and that had a lower Bayesian Information Criterion for the model with both additive and log-linear variance effects than for the model with only additive effects (Tables C.1, C.2, C.3, C.4, and C.5). The name indicates the nearest gene and/or a gene that the variant controls expression of: 1) indicates the SNP is a missense variant; 2) indicates the SNP is a eQTL for the named gene according to the GTEX data[155]; and 3) indicates the variant has previously been associated with HDL levels[154].

Bibliography

[1] Galton, F. *Natural Inheritance*. Natural Inheritance. Macmillan and Company 1889.

[2] Provine, W. B. *The Origins of Theoretical Population Genetics: With a New Afterword*. Chicago history of science and medicine. University of Chicago Press 2001. ISBN 9780226684635.

[3] Fisher, R. The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh*, **52**:399–433 1918.

[4] Cockerham, C. C. An Extension of the Concept of Partitioning Hereditary Variance for Analysis of Covariances Among Relatives when Epistasis is Present. *Genetics*, **39**(November) 1954.

[5] Kempthorne, O. The Correlation between Relatives in a Random Mating Population. *Proceedings of the Royal Society B: Biological Sciences*, **143**(910):103–113 1954. ISSN 0962-8452. doi:10.1098/rspb.1954.0056.

[6] Kempthorne, O. The Theoretical Values of Correlations between Relatives in Random Mating Populations. *Genetics*, **40**(2):153–67 1955. ISSN 0016-6731.

[7] Cox, D. Interaction. *International Statistical Review/Revue Internationale de Statistique*, **52**(1):1–24 1984.

[8] González, A. B. and Cox, D. R. Interpretation of interactions: a review. *Annals of Applied Statistics*, **1**(2):371–385 2007. ISSN 1932-6157. doi: 10.1214/07-AOAS124.

[9] Sesardic, N. *Making Sense of Heritability*. Cambridge Studies in Philosophy and Biology. Cambridge University Press 2005. ISBN 9781139445672.

[10] Cordell, H. J. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, **11**(20):2463–8 2002. ISSN 0964-6906. doi:10.1093/hmg/11.20.2463.

[11] Bateson, W. and Mendel, G. *Mendel's Principles of Heredity*. Putnam's 1909.

[12] Phillips, P. C. Epistasis the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, **9**(11):855–867 2008. ISSN 1471-0056. doi:10.1038/nrg2452.

[13] Hemani, G., Knott, S., and Haley, C. An Evolutionary Perspective on Epistasis and the Missing Heritability. *PLoS Genetics*, **9**(2) 2013. doi:10.1371/journal.pgen.1003295.

[14] Fisher, R. A. and Mackenzie, W. A. Studies in Crop Variation. II. The manurial response of different potato varieties. *Journal of Agricultural Science*, **13**:311–320 1923. ISSN 0021-8596. doi:10.1017/S0021859600003592.

[15] Freeman, G. Statistical methods for the analysis of genotype-environment interactions. *Heredity*, **31**(3):339–354 1973. ISSN 0018-067X. doi:10.1038/hdy.1973.90.

[16] Paaby, A. B. and Rockman, M. V. Cryptic genetic variation: evolution's hidden substrate. *Nature Reviews Genetics*, **15**(4):247–58 2014. ISSN 1471-0064. doi: 10.1038/nrg3688.

[17] Vieira, C., et al. Genotype-environment interaction for quantitative trait loci affecting life span in Drosophila melanogaster. *Genetics*, **154**:213–227 2000. ISSN 0016-6731.

[18] Ungerer, M. C., Halldorsdottir, S. S., Purugganan, M. D., and Mackay, T. F. C. Genotype-environment interactions at quantitative trait loci affecting inflorescence development in Arabidopsis thaliana. *Genetics*, **165**(1):353–365 2003. ISSN 00166731.

[19] Thomas, D. Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies. *Annual Review of Public Health*, **31**:21–36 2010. ISSN 1545-2093. doi:10.1146/annurev.publhealth.012809.103619.

[20] Hodgins-Davis, A., Adomas, A. B., Warringer, J., and Townsend, J. P. Abundant gene-by-environment interactions in gene expression reaction norms to copper within saccharomyces cerevisiae. *Genome Biology and Evolution*, **4**(11):1061–1079 2012. ISSN 17596653. doi:10.1093/gbe/evs084.

[21] Eichelbaum, M., Ingelman-Sundberg, M., and Evans, W. E. Pharmacogenomics and individualized drug therapy. *Annu Rev Med*, **57**:119–137 2006. ISSN 0066-4219. doi:10.1146/annurev.med.56.082103.104724.

[22] Marchini, J., Donnelly, P., and Cardon, L. R. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics*, **37**(4):413–417 2005. ISSN 1061-4036. doi:10.1038/ng1537.

[23] Wang, Y., Liu, G., Feng, M., and Wong, L. An empirical comparison of several recent epistatic interaction detection methods. *Bioinformatics*, **27**(21):2936–2943 2011. ISSN 13674803. doi:10.1093/bioinformatics/btr512.

[24] Cordell, H. J. Detecting gene-gene interactions that underlie human diseases. *Nature Reviews Genetics*, **10**(6):392–404 2009. ISSN 1471-0064. doi:10.1038/nrg2579.

[25] Zhang, Y. and Liu, J. S. Bayesian inference of epistatic interactions in case-control studies. *Nature Genetics*, **39**(9):1167–1173 2007. ISSN 1061-4036. doi:10.1038/ng2110.

[26] Bien, J., Simon, N., and Tibshirani, R. Convex hierarchical testing of interactions. *Annals of Applied Statistics*, **9**(1):27–42 2015. ISSN 19417330. doi:10.1214/14-AOAS758.

[27] Li, J., Zhong, W., Li, R., and Wu, R. A fast algorithm for detecting gene-gene interactions in genome-wide association studies. *Annals of Applied Statistics*, **8**(4):2292–2318 2014. ISSN 19417330. doi:10.1214/14-AOAS771.

[28] Fan, J. and Lv, J. Sure Independence Screening for Ultra-High Dimensional Feature Space. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, **70**(5):849–911 2008. ISSN 1369-7412. doi:10.1111/j.1467-9868.2008.00674.x.

[29] Duncan, L. E. and Keller, M. C. A critical review of the first 10 years of candidate gene-by-environment interaction research in psychiatry. *American Journal of Psychiatry*, **168**(10):1041–1049 2011.

[30] Ahmad, S., et al. Gene x Physical Activity Interactions in Obesity: Combined Analysis of 111,421 Individuals of European Ancestry. *PLoS Genetics*, **9**(7):1–9 2013. ISSN 15537390. doi:10.1371/journal.pgen.1003607.

[31] Ahmad, S., Varga, T. V., and Franks, P. W. Gene x environment interactions in obesity: The state of the evidence. *Human Heredity*, **75**(2-4):106–115 2013. ISSN 00015652. doi:10.1159/000351070.

[32] Manning, A. K., et al. Meta-analysis of gene-environment interaction: Joint estimation of SNP and SNP x environment regression coefficients. *Genetic Epidemiology*, **35**(1):11–18 2011. ISSN 07410395. doi:10.1002/gepi.20546.

[33] Hancock, D. B., et al. Genome-Wide Joint Meta-Analysis of SNP and SNP-by-Smoking Interaction Identifies Novel Loci for Pulmonary Function. *PLoS Genetics*, **8**(12) 2012. ISSN 15537390. doi:10.1371/journal.pgen.1003098.

[34] Thomas, D. Gene–environment-wide association studies: emerging approaches. *Nature Reviews Genetics*, **11**(4):259–272 2010. ISSN 1471-0056. doi:10.1038/nrg2764.

[35] Searle, S. R., Casella, G., and McCulloch, C. E. *Variance Components* 2006. ISBN I 3 978-0-470-00959-8.

[36] Graybill, F. A. and Hultquist, R. A. Theorems Concerning Eisenhart's Model II. *Ann. Math. Statist.*, **32**(1):261–269 1961. doi:10.1214/aoms/1177705158.

[37] Mrode, R. A. and Thompson, R. *Linear Models for the Prediction of Animal Breeding Values*. CAB books. CABI Pub. 2005. ISBN 9781845931025.

[38] Boomsma, D., Busjahn, A., and Peltonen, L. Classical twin studies and beyond. *Nature Reviews Genetics*, **3**(11):872–82 2002. ISSN 1471-0056. doi:10.1038/nrg932.

[39] Yang, J., et al. Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, **42**(7):565–9 2010. ISSN 1546-1718. doi:10.1038/ng.608.

[40] Yang, J., et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature Genetics*, **47**(10):1114–1120 2015. ISSN 1061-4036.

[41] Yang, J., Zaitlen, N. a., Goddard, M. E., Visscher, P. M., and Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics*, **46**(2):100–6 2014. ISSN 1546-1718. doi:10.1038/ng.2876.

[42] Astle, W. and Balding, D. J. Population Structure and Cryptic Relatedness in Genetic Association Studies. *Statistical Science*, **24**(4):451–471 2009. ISSN 0883-4237. doi:10.1214/09-STS307.

[43] Novembre, J., et al. Genes mirror geography within Europe. *Nature*, **456**(7218):98–101 2008. ISSN 1476-4687. doi:10.1038/nature07331.

[44] Young, G. A. and Smith, R. L. *Essentials of statistical inference*. Cambridge University Press 2005.

[45] Zhang, Z., et al. Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, **42**(4):355–60 2010. ISSN 1546-1718. doi: 10.1038/ng.546.

[46] Loh, P.-R., et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, **47**(3):284–290 2015. ISSN 1061-4036. doi:10.1038/ng.3190.

[47] Gilmour, A. R., Thompson, R., and Cullis, B. R. Average Information REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models. *Biometrics*, **51**(4):1440–1450 1995.

[48] Kang, H. M., et al. Efficient Control of Population Structure in Model Organism Association Mapping. *Genetics*, **178**(3):1709–1723 2008. ISSN 0016-6731. doi: 10.1534/genetics.107.080101.

[49] Lippert, C., et al. FaST linear mixed models for genome-wide association studies. *Nature Methods*, **8**(10):833–5 2011. ISSN 1548-7105. doi: 10.1038/nmeth.1681.

[50] Listgarten, J., Lippert, C., and Heckerman, D. FaST-LMM-Select for addressing confounding from spatial structure and rare variants. *Nature Genetics*, **45**(5):470–1 2013. ISSN 1546-1718. doi:10.1038/ng.2620.

[51] Lippert, C., et al. The benefits of selecting phenotype-specific variants for applications of mixed models in genomics. *Scientific Reports*, **3**:1815 2013. ISSN 2045-2322. doi:10.1038/srep01815.

[52] Matilainen, K., Mantysaari, E. A., Lidauer, M. H., Stranden, I., and Thompson, R. Employing a Monte Carlo algorithm in Newton-type methods for restricted maximum likelihood estimation of genetic parameters. *PLoS ONE*, **8**(12):2–8 2013. ISSN 19326203. doi:10.1371/journal.pone.0080821.

[53] Visscher, P. M., Brown, M. a., McCarthy, M. I., and Yang, J. Five years of GWAS discovery. *American Journal of Human Genetics*, **90**(1):7–24 2012. ISSN 1537-6605. doi:10.1016/j.ajhg.2011.11.029.

[54] Eichler, E. E., et al. Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, **11**(6):446–50 2010. ISSN 1471-0064. doi:10.1038/nrg2809.

[55] Zuk, O., Hechter, E., Sunyaev, S. R., and Lander, E. S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences of the United States of America*, **109**(4):1193–8 2012. ISSN 1091-6490. doi:10.1073/pnas.1119675109.

[56] Purcell, S. Variance components models for gene-environment interaction in twin analysis. *Twin Research*, **5**(6):554–571 2002. ISSN 1369-0523. doi:10.1375/twin.5.6.554.

[57] Allen, N., et al. UK Biobank: Current status and what it means for epidemiology. *Health Policy and Technology*, **1**(3):123–126 2012. ISSN 22118837. doi:10.1016/j.hlpt.2012.07.003.

[58] Genotyping and quality control of UK Biobank, a large-scale, extensively phenotyped prospective resource: http://biobank.ctsu.ox.ac.uk/crystal/docs/genotyping_qc.pdf.

[59] Leslie, S., et al. The fine-scale genetic structure of the British population. *Nature*, **519**(7543):309–314 2015. ISSN 0028-0836. doi:10.1038/nature14230.

[60] Malik, V. S., Willett, W. C., and Hu, F. B. Global obesity: trends, risk factors and policy implications. *Nature Reviews Endocrinology*, **9**(1):13–27 2013. ISSN 1759-5037. doi:10.1038/nrendo.2012.199.

[61] Ntuk, U. E., Gill, J. M. R., Mackay, D. F., Sattar, N., and Pell, J. P. Ethnic-Specific Obesity Cutoffs for Diabetes Risk: Cross-sectional Study of 490,288 UK Biobank Participants. *Diabetes Care*, **37**(September):1–8 2014. ISSN 1935-5548. doi:10.2337/dc13-2966.

[62] Ogden, C. L., Yanovski, S. Z., Carroll, M. D., and Flegal, K. M. The Epidemiology of Obesity. *Gastroenterology*, **132**(6):2087–2102 2007. ISSN 00165085. doi:10.1053/j.gastro.2007.03.052.

[63] Elks, C. E., et al. Variability in the heritability of body mass index: A systematic review and meta-regression. *Frontiers in Endocrinology*, **3**(FEB):1–16 2012. ISSN 16642392. doi:10.3389/fendo.2012.00029.

[64] Rich, J. D., Allen, S. A., Williams, B. A., and Chin, J. The Biology and Genetics of Obesity A Century of Inquiries. *New England Journal of Medicine*, **370**(20):1874–1877 2014. ISSN 00284793. doi:10.1056/NEJMp1400613.

[65] O'Rahilly, S. and Farooqi, I. S. Human obesity: A heritable neurobehavioral disorder that is highly sensitive to environmental conditions. *Diabetes*, **57**(11):2905–2910 2008. ISSN 00121797. doi:10.2337/db08-0210.

[66] El-Sayed Moustafa, J. S. and Froguel, P. From obesity genetics to the future of personalized obesity therapy. *Nature Reviews Endocrinology*, **9**(7):402–13 2013. ISSN 1759-5037. doi:10.1038/nrendo.2013.57.

[67] Speliotes, E. K., et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics*, **42**(11):937–948 2010. ISSN 1061-4036. doi:10.1038/ng.686.

[68] Locke, A. E., et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, **518**(7538):197–206 2015. ISSN 0028-0836. doi:10.1038/nature14177.

[69] Huang, T. and Hu, F. B. Gene-environment interactions and obesity: recent developments and future directions. *BMC Medical Genomics*, **8**(Suppl 1):S2 2015. ISSN 1755-8794. doi:10.1186/1755-8794-8-S1-S2.

[70] Visscher, P. M., et al. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genetics*, **2**(3):e41 2006. ISSN 1553-7404. doi:10.1371/journal.pgen.0020041.

[71] Strange, A., et al. A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. *Nature Genetics*, **42**(11):985–90 2010. ISSN 1546-1718. doi:10.1038/ng.694.

[72] Brown, A. A., et al. Genetic interactions affecting human gene expression identified by variance association mapping. *eLife*, **2014**(3):1–16 2014. ISSN 2050084X. doi:10.7554/eLife.01381.

[73] Mackay, T. F. C. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nature Reviews Genetics*, **15**(1):22–33 2014. ISSN 1471-0064. doi:10.1038/nrg3627.

[74] Vukcevic, D., Hechter, E., Spencer, C., and Donnelly, P. Disease model distortion in association studies. *Genetic Epidemiology*, **290**:278–290 2011. ISSN 1098-2272. doi:10.1002/gepi.20576.

[75] Gallais, A. Covariances between Arbitrary Relatives with Linkage and Epistasis in the Case of Linkage Disequilibrium. *Biometrics*, **30**(3):429–446 1974.

[76] Abney, M., McPeek, M. S., and Ober, C. Estimation of variance components of quantitative traits in inbred populations. *American Journal of Human Genetics*, **66**(2):629–50 2000. ISSN 0002-9297. doi:10.1086/302759.

[77] Carmi, S., et al. The variance of identity-by-descent sharing in the Wright-Fisher model. *Genetics*, **193**(3):911–28 2013. ISSN 1943-2631. doi:10.1534/genetics.112.147215.

[78] Cockerham, C. C. and Tachida, H. Permanency of response to selection for quantitative characters in finite populations. *Proceedings of the National Academy of Sciences of the United States of America*, **85**(5):1563–5 1988. ISSN 0027-8424.

[79] Tachida, H. and Cockerham, C. C. Effects of identity disequilibrium and linkage on quantitative variation in finite populations. *Genetical Research*, **53**(1):63–70 1989.

[80] Baud, A., et al. Combined sequence-based and genetic mapping analysis of complex traits in outbred rats. *Nature Genetics*, **45**(7):767–75 2013. ISSN 1546-1718. doi:10.1038/ng.2644.

[81] Browning, S. R. and Browning, B. L. Identity-by-descent-based heritability analysis in the Northern Finland Birth Cohort. *Human Genetics*, **132**(2):129–38 2013. ISSN 1432-1203. doi:10.1007/s00439-012-1230-y.

[82] Zaitlen, N., et al. Using Extended Genealogy to Estimate Components of Heritability for 23 Quantitative and Dichotomous Traits. *PLoS Genetics*, **9**(5):e1003520 2013. ISSN 1553-7404. doi:10.1371/journal.pgen.1003520.

[83] Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *American Journal of Human Genetics*, **88**(1):76–82 2011. ISSN 1537-6605. doi:10.1016/j.ajhg.2010.11.011.

[84] Bloom, J. S., Ehrenreich, I. M., Loo, W. T., Lite, T.-L. V., and Kruglyak, L. Finding the sources of missing heritability in a yeast cross. *Nature*, **494**(7436):234–7 2013. ISSN 1476-4687. doi:10.1038/nature11867.

[85] Gauvin, H., et al. Genome-wide patterns of identity-by-descent sharing in the French Canadian founder population. *European Journal of Human Genetics*, pages 1–8 2013. ISSN 1476-5438. doi:10.1038/ejhg.2013.227.

[86] Khoury, M. J., Cohen, B. H., Diamond, E. L., Chase, G. A., and McKusick, V. A. Inbreeding and prereproductive mortality in the Old Order Amish. I. Genealogic epidemiology of inbreeding. *American Journal of Epidemiology*, **125**(3):453–61 1987. ISSN 0002-9262.

[87] Visscher, P. M., et al. Statistical Power to Detect Genetic (Co)Variance of Complex Traits Using SNP Data in Unrelated Samples. *PLoS Genetics*, **10**(4) 2014. ISSN 15537404. doi:10.1371/journal.pgen.1004269.

[88] Bloom, J. S., et al. Genetic interactions contribute less than additive effects to quantitative trait variation in yeast. *Nature Communications*, **6**:8712 2015. ISSN 2041-1723. doi:10.1038/ncomms9712.

[89] Hill, W. G., Goddard, M. E., and Visscher, P. M. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genetics*, **4**(2):e1000008 2008. ISSN 1553-7404. doi:10.1371/journal.pgen.1000008.

[90] Liti, G., et al. Population genomics of domestic and wild yeasts. *Nature*, **458**(7236):337–41 2009. ISSN 1476-4687. doi:10.1038/nature07743.

[91] Anderson, J. B., et al. Determinants of divergent adaptation and Dobzhansky-Muller interaction in experimental yeast populations. *Current Biology*, **20**(15):1383–8 2010. ISSN 1879-0445. doi:10.1016/j.cub.2010.06.022.

[92] Orr, H. A. The Population Genetics of Speciation: The Evolution of Hybrid Incompatibilities. *Genetics*, **139**(April):1805–1813 1995.

[93] Loos, R. J. F. and Yeo, G. S. H. The bigger picture of FTO–the first GWAS-identified obesity gene. *Nature Reviews Endocrinology*, **10**(1):51–61 2014. ISSN 1759-5037. doi:10.1038/nrendo.2013.227.

[94] Claussnitzer, M., et al. FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *The New England Journal of Medicine*, **373**(10):895–907 2015. ISSN 1533-4406 (Electronic). doi:10.1056/NEJMoa1502214.

[95] Bell, C. G., et al. Integrated genetic and epigenetic analysis identifies haplotype-specific methylation in the FTO type 2 diabetes and obesity susceptibility locus. *PLoS ONE*, **5**(11) 2010. ISSN 19326203. doi:10.1371/journal.pone.0014040.

[96] Stratigopoulos, G., LeDuc, C. A., Cremona, M. L., Chung, W. K., and Leibel, R. L. Cut-like homeobox 1 (CUX1) regulates expression of the fat mass

and obesity-associated and retinitis pigmentosa GTPase regulator-interacting protein-1-like (RPGRIP1L) genes and coordinates leptin receptor signaling. *Journal of Biological Chemistry*, **286**(3):2155–2170 2011. ISSN 00219258. doi: 10.1074/jbc.M110.188482.

[97] Qi, Q., et al. FTO genetic variants, dietary intake, and body mass index: insights from 177,330 individuals. *Human Molecular Genetics*, **23**(25):1–12 2014. ISSN 1460-2083. doi:10.1093/hmg/ddu411.

[98] Westerterp, K. R. Diet induced thermogenesis. *Nutrition & Metabolism*, **1**:5 2004. ISSN 1743-7075. doi:10.1186/1743-7075-1-5.

[99] Kilpeläinen, T. O., et al. Physical activity attenuates the influence of FTO variants on obesity risk: A meta-analysis of 218,166 adults and 19,268 children. *PLoS Medicine*, **8**(11) 2011. ISSN 15491277. doi:10.1371/journal.pmed.1001116.

[100] Li, S., et al. Physical activity attenuates the genetic predisposition to obesity in 20,000 men and women from EPIC-Norfolk prospective population study. *PLoS Medicine*, **7**(8):1–9 2010. ISSN 15491277. doi:10.1371/journal.pmed.1000332.

[101] Phillips, C. M., et al. High Dietary Saturated Fat Intake Accentuates Obesity Risk Associated with the Fat Mass and Obesity-Associated Gene in Adults. *Journal of Nutrition*, **142**(5):824–831 2012. ISSN 0022-3166. doi:10.3945/jn.111.153460.

[102] Moleres, A., et al. Dietary fatty acid distribution modifies obesity risk linked to the rs9939609 polymorphism of the fat mass and obesity-associated gene in a Spanish casecontrol study of children. *British Journal of Nutrition*, **107**(04):533–538 2012. ISSN 0007-1145. doi:10.1017/S0007114511003424.

[103] Corella, D., et al. A High Intake of Saturated Fatty Acids Strengthens the Association between the Fat Mass and Obesity-Associated Gene and

BMI. *Journal of Nutrition*, **141**(12):2219–2225 2011. ISSN 0022-3166. doi: 10.3945/jn.111.143826.

[104] Qi, Q., et al. Fried food consumption, genetic risk, and body mass index: gene-diet interaction analysis in three US cohort studies. *BMJ*, **348**(March):g1610 2014. ISSN 1756-1833. doi:10.1136/bmj.g1610.

[105] Ragland, D. R. Dichotomizing continuous outcome variables: dependence of the magnitude of association and statistical power on the cutpoint. *Epidemiology*, **3**(5):434–440 1992. ISSN 1044-3983. doi:10.1097/00001648-199209000-00009.

[106] Allen, N. E., Sudlow, C., Peakman, T., and Collins, R. UK Biobank Data: Come and Get It. *Science Translational Medicine*, **6**(224):224ed4 2014. ISSN 1946-6242. doi:10.1126/scitranslmed.3008601.

[107] UK Biobank. Genotype imputation and genetic association studies of UK Biobank Interim Data Release: http://biobank.ctsu.ox.ac.uk/crystal/docs/impute_ukb_v1.pdf 2015.

[108] Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference on Artificial Intelligence*, **14**(12):1137–1143 1995. ISSN 10450823. doi:10.1067/mod.2000.109031.

[109] UK Biobank. UK Biobank Anthropometry: http://biobank.ctsu.ox.ac.uk/crystal/docs/Anthropometry.pdf 2014.

[110] Townsend, P. Deprivation. *Journal of Social Policy*, **16**(02):125–146 1987. ISSN 1469-7823. doi:10.1017/S0047279400020341.

[111] Patterson, N., Price, A. L., and Reich, D. Population structure and eigenanalysis. *PLoS Genetics*, **2**(12):2074–2093 2006. ISSN 15537390. doi: 10.1371/journal.pgen.0020190.

[112] Galinsky, K. J., et al. Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia. *American Journal of Human Genetics*, **98**(3):456–472 2016. ISSN 15376605. doi:10.1016/j.ajhg.2015.12.022.

[113] Devlin, B. and Roeder, K. Genomic control for association studies. *Biometrics*, **55**(4):997–1004 1999. ISSN 0006-341X. doi:10.1111/j.0006-341X.1999.00997.x.

[114] Yang, J., et al. Ubiquitous polygenicity of human complex traits: genome-wide analysis of 49 traits in Koreans. *PLoS Genetics*, **9**(3):e1003355 2013. ISSN 1553-7404. doi:10.1371/journal.pgen.1003355.

[115] Bulik-Sullivan, B. K., et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, **47**(3):291–295 2015. ISSN 1061-4036. doi:10.1038/ng.3211.

[116] WTCCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**(7145):661–78 2007. ISSN 1476-4687. doi:10.1038/nature05911.

[117] International Stroke Genetics Consortium (ISGC), et al. Genome-wide association study identifies a variant in HDAC9 associated with large vessel ischemic stroke. *Nature Genetics*, **44**(3):328–333 2012. ISSN 1061-4036. doi:10.1038/ng.1081.

[118] UK Biobank. UK Biobank 24-hour dietary recall questionnaire: http://biobank.ctsu.ox.ac.uk/crystal/docs/DietWebQ.pdf 2012.

[119] Vimaleswaran, K. S., et al. Physical activity attenuates the body mass index increasing influence of genetic variation in the FTO gene. *The American Journal of Clinical Nutririon*, **90**:425–428 2009. ISSN 1938-3207. doi:10.3945/ajcn.2009.27652.

[120] Cauchi, S., et al. Combined effects of MC4R and FTO common genetic variants on obesity in European general populations. *Journal of Molecular Medicine*, **87**(5):537–546 2009. ISSN 09462716. doi:10.1007/s00109-009-0451-6.

[121] Andreasen, C. H., et al. Low physical activity acentuates the effect of rs9939609 polymorphism. *Diabetes*, **57**(January):95–101 2008. ISSN 1939-327X. doi: 10.2337/db07-0910.

[122] Ruiz, J. R., et al. Attenuation of the effect of the FTO rs9939609 polymorphism on total and central body fat by physical activity in adolescents: The Helena Study. *Archives of Pediatrics & Adolescent Medicine*, **164**(4):328–333 2010. doi:10.1001/archpediatrics.2010.29.

[123] Scott, R. A., et al. FTO genotype and adiposity in children: physical activity levels influence the effect of the risk genotype in adolescent males. *European Journal of Human Genetics*, **18**(12):1339–43 2010. ISSN 1476-5438. doi:10. 1038/ejhg.2010.131.

[124] Breslow, R. a. and Smothers, B. a. Drinking patterns and body mass index in never smokers: National Health Interview Survey, 1997-2001. *American Journal of Epidemiology*, **161**(4):368–376 2005. ISSN 00029262. doi:10.1093/ aje/kwi061.

[125] Tolstrup, J. S., et al. The relation between drinking pattern and body mass index and waist and hip circumference. *International Journal of Obesity*, **29**(5):490–497 2005. ISSN 0307-0565. doi:10.1038/sj.ijo.0802874.

[126] Sobczyk-Kopciol, A., et al. Inverse association of the obesity predisposing FTO rs9939609 genotype with alcohol consumption and risk for alcohol dependence. *Addiction*, **106**(4):739–748 2011. ISSN 09652140. doi:10.1111/j.1360-0443.2010. 03248.x.

[127] Taheri, S., Lin, L., Austin, D., Young, T., and Mignot, E. Short sleep duration is associated with reduced leptin, elevated ghrelin, and increased body mass index. *PLoS Medicine*, **1**(3):210–217 2004. ISSN 15491277. doi:10.1371/journal.pmed.0010062.

[128] McLaren, L. Socioeconomic status and obesity. *Epidemiologic Reviews*, **29**(1):29–48 2007. ISSN 0193936X. doi:10.1093/epirev/mxm001.

[129] Qi, Q., et al. Television watching, leisure time physical activity, and the genetic predisposition in relation to body mass index in women and men. *Circulation*, **126**(15):1821–1827 2012. ISSN 00097322. doi:10.1161/CIRCULATIONAHA.112.098061.

[130] Holford, T. R. *Multivariate Methods in Epidemiology*. Monographs in Epidemiology and Biostatistics. Oxford University Press, USA 2002. ISBN 9780195124408.

[131] Ahmad, T., et al. Lifestyle interaction with fat mass and obesity-associated (FTO) genotype and risk of obesity in apparently healthy U.S. women. *Diabetes Care*, **34**(3):675–680 2011. ISSN 01495992. doi:10.2337/dc10-0948.

[132] Cox, D. N., Perry, L., Moore, P. B., Vallis, L., and Mela, D. J. Sensory and hedonic associations with macronutrient and energy intakes of lean and obese consumers. *International Journal of Obesity*, **23**(4):403–410 1999. ISSN 0307-0565. doi:10.1038/sj.ijo.0800836.

[133] Greenfield, J. R., et al. Moderate alcohol consumption, dietary fat composition, and abdominal obesity in women: evidence for gene-environment interaction. *The Journal of Clinical Endocrinology and Metabolism*, **88**(11):5381–5386 2003. ISSN 0021972X. doi:10.1210/jc.2003-030851.

[134] Watson, N. F., et al. Sleep Duration and Body Mass Index in Twins: A Gene-Environment Interaction. *Sleep*, **35**(5):597–603 2016. ISSN 0161-8105. doi: 10.5665/sleep.1810.

[135] Paré, G., Cook, N. R., Ridker, P. M., and Chasman, D. I. On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the Women's Genome Health Study. *PLoS Genetics*, **6**(6):e1000981 2010. ISSN 1553-7404. doi:10.1371/journal.pgen.1000981.

[136] Struchalin, M. V., Dehghan, A., Witteman, J. C., van Duijn, C., and Aulchenko, Y. S. Variance heterogeneity analysis for detection of potentially interacting genetic loci: method and its limitations. *BMC Genetics*, **11**(1):92 2010. ISSN 1471-2156. doi:10.1186/1471-2156-11-92.

[137] Yang, J., et al. FTO genotype is associated with phenotypic variability of body mass index. *Nature*, **490**(7419):267–272 2013. doi:10.1038/nature11401.FTO.

[138] Rönnegård, L. and Valdar, W. Recent developments in statistical methods for detecting genetic loci affecting phenotypic variability. *BMC Genetics*, **13**(1):63 2012. ISSN 1471-2156. doi:10.1186/1471-2156-13-63.

[139] Dumitrascu, B., Darnell, G., Ayroles, J., and Engelhardt, B. E. A Bayesian test to identify variance effects. *arXiv* 2015.

[140] Cao, Y., Wei, P., Bailey, M., Kauwe, J. S. K., and Maxwell, T. J. A versatile omnibus test for detecting mean and variance heterogeneity. *Genetic Epidemiology*, **38**(1):51–59 2014. ISSN 07410395. doi:10.1002/gepi.21778.

[141] Cao, Y., Maxwell, T. J., and Wei, P. A Family-Based Joint Test for Mean and Variance Heterogeneity for Quantitative Traits. *Annals of Human Genetics*, **79**(1):46–56 2015. ISSN 00034800. doi:10.1111/ahg.12089.

[142] Barton, N. H., Etheridge, A. M., and Veber, A. The Infinitesimal Model. *bioRxiv* 2016.

[143] Brillinger, D. R. Some data analyses using mutual information. *Brazilian Journal of Probability and Statistics*, **18**(2000):163–182 2004.

[144] Harvey, A. A. C. Estimating regression models with multiplicative heteroscedasticity. *Econometrica*, **44**(3):461–465 1976.

[145] Wolfinger, R., Tobiast, R., Sall, J., Tobias, R., and Sall, J. Computing Gaussian Likelihoods and Their Derivatives for General Linear Mixed Models. *SIAM Journal on Scientific Computing*, **15**(6):1294–1310 1994. doi:10.1137/0915079.

[146] Zhu, Z., et al. Dominance Genetic Variation Contributes Little to the Missing Heritability for Human Complex Traits. *The American Journal of Human Genetics*, pages 377–385 2015. ISSN 00029297. doi:10.1016/j.ajhg.2015.01.001.

[147] Price, A. L., et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, **38**(8):904–909 2006. ISSN 1061-4036. doi:10.1038/ng1847.

[148] Box, A. G. E. P. and Box, G. E. P. Non-Normality and Tests on Variances. *Biometrika*, **40**(3/4):318–335 1953. ISSN 00063444.

[149] https://data.broadinstitute.org/alkesgroup/LDSCORE/.

[150] Schwarz, G. Estimating the dimension of a model. *The Annals of Statistics*, **6**(2):461–464 1978. ISSN 0090-5364. doi:10.1214/aos/1176344136.

[151] Ward, E. J. A review and comparison of four commonly used Bayesian and maximum likelihood model selection tools. *Ecological Modelling*, **211**(1-2):1–10 2008. ISSN 03043800. doi:10.1016/j.ecolmodel.2007.10.030.

[152] McCarthy, M. I., et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature*, **9**(May):356–369 2008. ISSN 1471-0064. doi:10.1038/nrg2344.

[153] Yang, J., et al. FTO genotype is associated with phenotypic variability of body mass index. *Nature*, **490**(7419):267–272 2012. ISSN 0028-0836. doi:10.1038/nature11401.

[154] Manning, A. K., et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nature Genetics*, **44**(6):659–69 2012. ISSN 1546-1718. doi:10.1038/ng.2274.

[155] The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, **348**(6235):648–660 2015. ISSN 1095-9203. doi:10.1126/science.1262110.

[156] Meyre, D., et al. Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. *Nature Genetics*, **41**(2):157–159 2009. ISSN 1061-4036. doi:10.1038/ng.301.

[157] Helgason, A., et al. Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution. *Nature Genetics*, **39**(2):218–225 2007. ISSN 1061-4036. doi:10.1038/ng1960.

[158] Schwarzer, G. *meta: General Package for Meta-Analysis* 2015.

[159] Hardy, R. J. and Thompson, S. G. Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine*, **17**(8):841–856 1998. ISSN 02776715. doi:10.1002/(SICI)1097-0258(19980430)17:8⟨841::AID-SIM781⟩3.0.CO;2-D.

[160] Zhou, Y., et al. TCF7L2 is a master regulator of insulin production and processing. *Human Molecular Genetics*, **23**(24):1–34 2014. ISSN 1460-2083. doi: 10.1093/hmg/ddu359.

[161] Fusi, N., Lippert, C., Lawrence, N. D., and Stegle, O. Warped linear mixed models for the genetic analysis of transformed phenotypes. *Nature Communications*, **5**(May):4890 2014. ISSN 2041-1723. doi:10.1038/ncomms5890.

[162] Taylor, M. B. and Ehrenreich, I. M. Higher-order genetic interactions and their contribution to complex traits. *Trends in Genetics*, **31**(1):34–40 2015. ISSN 01689525. doi:10.1016/j.tig.2014.09.001.

[163] Märtens, K., Hallin, J., Warringer, J., Liti, G., and Parts, L. Predicting quantitative traits from genome and phenome with near perfect accuracy. *Nature Communications* 2016. doi:10.1038/ncomms11512.

[164] Hallin, J., et al. Powerful decomposition of complex traits in a diploid model. *Nature Communications*, **7**:13311 2016. doi:10.1038/ncomms13311.

[165] Mäki-Tanila, A. and Hill, W. G. Influence of Gene Interaction on Complex Trait Variation with Multi-Locus Models. *Genetics*, **198**(September):355–367 2014. ISSN 1943-2631. doi:10.1534/genetics.114.165282.

[166] Coyne, J. A. and Orr, H. A. *Speciation*. W.H. Freeman 2004. ISBN 9780878930890.

[167] Polderman, T. J. C., et al. Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics*, **47**(7):702–709 2015. ISSN 1061-4036. doi:10.1038/ng.3285.

[168] Derks, E. M., Dolan, C. V., and Boomsma, D. I. A test of the equal environment assumption (EEA) in multivariate twin studies. *Twin Research and Human Genetics*, **9**(3):403–411 2006. ISSN 1832-4274. doi:10.1375/twin.9.3.403.

[169] Sackton, T. B. and Hartl, D. L. Genotypic Context and Epistasis in Individuals and Populations. *Cell*, **166**(2):279–287 2016. ISSN 00928674. doi:10.1016/j.cell.2016.06.047.

[170] Finucane, H. K., et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics*, **47**(11):1228–1235 2015. ISSN 1546-1718. doi:10.1038/ng.3404.

[171] Gibson, G. and Dworkin, I. Uncovering cryptic genetic variation. *Nature Reviews Genetics*, **5**(9):681–690 2004. ISSN 1471-0056. doi:10.1038/nrg1426.

[172] Passa, P. Diabetes trends in Europe. *Diabetes. Metab. Res. Rev.*, (18):3–8 2002. doi:10.1002/dmrr.276.

[173] Carroll, M. D. Trends in Serum Lipids and Lipoproteins of Adults, 1960-2002. *JAMA: The Journal of the American Medical Association*, **294**(14):1773 2005. ISSN 0098-7484. doi:10.1001/jama.294.14.1773.

[174] Farzadfar, F., et al. National, regional, and global trends in serum total cholesterol since 1980: Systematic analysis of health examination surveys and epidemiological studies with 321 country-years and 3.0 million participants. *The Lancet*, **377**(9765):578–586 2011. ISSN 01406736. doi:10.1016/S0140-6736(10)62038-7.

[175] Goldstein, J. R., Lutz, W., and Testa, M. R. The emergence of sub-replacement family size ideals in Europe. *Population Research and Policy Review*, **22**(2001):479–496 2003. ISSN 00987921. doi:10.1111/j.1728-4457.2012.00475.x.

[176] Klerman, G. L. and Weissman, M. M. Increasing Rates of Depression. *JAMA: The Journal of the American Medical Association*, **261**(15):2229–2235 1989. ISSN 0098-7484, 1538-3598. doi:10.1001/jama.1989.03420150079041.

[177] Fletcher, J. M. and Conley, D. The challenge of causal inference in gene-environment interaction research: Leveraging research designs from the social sciences. *American Journal of Public Health*, **103**(SUPPL.1):42–45 2013. ISSN 00900036. doi:10.2105/AJPH.2013.301290.

[178] Smith, G. D. and Ebrahim, S. 'Mendelian randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology*, **32**(1):1–22 2003. ISSN 03005771. doi: 10.1093/ije/dyg070.

[179] Smith, G. D. and Ebrahim, S. Mendelian randomization: Prospects, potentials, and limitations. *International Journal of Epidemiology*, **33**(1):30–42 2004. ISSN 03005771. doi:10.1093/ije/dyh132.

[180] Pickrell, J. Fulfilling the promise of Mendelian randomization. *bioRxiv* 2015. doi:10.1101/018150.

[181] Magnus, J. R. and Neudecker, H. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley Series in Probability and Statistics: Texts and References Section. Wiley 1999. ISBN 9780471986331.