A biobank-scale test of marginal epistasis reveals genome-wide signals of polygenic epistasis

Boyang Fu *[1], Ali Pazokitoroudi *[1], Albert Xue[2], Aakarsh Anand[1], Prateek Anand[1], Noah Zaitlen[3,4], and Sriram Sankararaman[1,4,5]

[1]*Department of Computer Science, UCLA, Los Angeles, CA , USA*
[2]*Bioinformatics Interdepartmental Program, UCLA, Los Angeles, CA, USA*
[3]*Department of Neurology, UCLA, Los Angeles, CA , USA*
[4]*Department of Computational Medicine, David Geffen School of Medicine, UCLA, Los Angeles, CA , USA*
[5]*Department of Human Genetics, David Geffen School of Medicine, UCLA, Los Angeles, CA , USA*

Abstract

The contribution of epistasis (interactions among genes or genetic variants) to human complex trait variation remains poorly understood. Methods that aim to explicitly identify pairs of genetic variants, usually single nucleotide polymorphisms (SNPs), associated with a trait suffer from low power due to the large number of hypotheses tested while also having to deal with the computational problem of searching over a potentially large number of candidate pairs. An alternate approach involves testing whether a single SNP modulates variation in a trait against a polygenic background. While overcoming the limitation of low power, such tests of polygenic or marginal epistasis (ME) are infeasible on Biobank-scale data where hundreds of thousands of individuals are genotyped over millions of SNPs.

We present a method to test for ME of a SNP on a trait that is applicable to biobank-scale data. We performed extensive simulations to show that our method provides calibrated tests of ME. We applied our method to test for ME at SNPs that are associated with 53 quantitative traits across ≈ 300 K unrelated white British individuals in the UK Biobank (UKBB). Testing $15,601$ trait-loci associations that were significant in GWAS, we identified 16 trait-loci pairs across 12 traits that demonstrate strong evidence of ME signals (p-value $p < \frac{5\times10^{-8}}{53}$). We further partitioned the significant ME signals across the genome to identify 6 trait-loci pairs with evidence of local (within-chromosome) ME while 15 show evidence of distal (cross-chromosome) ME. Across the 16 trait-loci pairs, we document that the proportion of trait variance explained by ME is about 12x as large as that explained by the GWAS effects on average (range: 0.59 to 43.89). Our results show, for the first time, evidence of interaction effects between individual genetic variants and overall polygenic background modulating complex trait variation.

1 Introduction

The effect of interactions across genes or genetic variants on a trait (*epistasis*) [1] has been hypothesized to play an important role in human complex trait variation [2, 3]. Understanding the nature and contribution of epistasis is important for elucidating the genetic architecture of complex traits and disease etiology and to improve the accuracy of genetic prediction. Epistasis is one of the factors that could explain missing heritability [4, 5] although some studies suggest a limited contribution of genetic interactions to complex

*These authors contributed equally to this work

trait variation [6]. Recent studies analyzing the estimates of genetic effects across ancestral populations [7] and the transferability of genetic predictors both within [8] and across ancestries [9, 10] suggest that genetic interactions could explain why genetic effects differ across ancestral populations and the lack of transferability of genetic predictors both within and across ancestries. Epistasis has also been hypothesized to play a role in variable expressivity of complex traits [11]. Nevertheless, our understanding of the role of epistasis in human traits remains limited [12, 5].

Over the past decade, a number of methods to detect epistasis have been developed. The first class of methods explicitly search for pairs of genetic variants (usually single nucleotide polymorphisms or SNPs) that have a non-linear effect on a trait. While allowing for an unbiased search for epistasis (analogous to GWAS enabling an unbiased approach to detect associations), these methods pose serious challenges. Exhaustively searching all pairs of SNPs is computationally difficult (scaling quadratically in the number of SNPs). Further, testing such a large number of hypotheses while controlling the false positive rate requires imposing stringent significance thresholds (scaling quadratically in the number of SNPs if a Bonferroni correction were to be used) which, in turn, reduces power. Efforts to solve this problem have involved the use of statistical techniques [13, 14, 15, 16, 17, 18], algorithmic innovations [19] or hardware infrastructure [20, 21, 22, 23, 24, 25, 26, 27, 28]. Alternate strategies have attempted to reduce the set of SNPs analyzed either restricting to analysis to SNPs identified in GWAS [29, 30, 31] or that are biologically functional [32, 33]. An alternate approach to detect epistasis aims to test for the aggregate epistatic effect across SNPs [34, 35, 36, 37]. Many of these approaches rely on the framework of variance components models that have improved power to detect additive genetic effects in aggregate (in contrast to GWAS that aims to identify individual effects). In this framework, it is of interest to test if the effect of a SNP on a trait is modulated by an individual's polygenic background. Such tests of *marginal epistasis* [36, 38] can improve power on account of the reduced multiple testing burden (that now scales with the number of SNPs) and due to the aggregation of a number of weak epistatic signals.

Even with the potential improvements in power, it is likely the case that tests of marginal epistasis need to be applied to datasets with large samples to identify robust signals of epistasis [39, 3, 40]. The availability of datasets that contain genetic and phenotypic information across hundreds of thousands of individuals offer an opportunity to detect epistasis with confidence. Estimating marginal epistasis from large data sets such as the UK Biobank consisting of $\approx 500,000$ individuals genotyped at nearly one million SNPs is computationally intractable.

We study the problem of testing whether the effect of a target SNP on a trait is modulated by the genotype of the individual at the remaining SNPs. Given genotypes collected from N individuals across M SNPs, we consider a model that aims to estimate and test the marginal epistatic effect defined as the combined pairwise interaction effects between a given SNP and all other SNPs while controlling for linear, additive effects. We propose a variance components estimation algorithm to jointly estimate the additive genetic variance component and the marginal epistatic (ME) variance components. The proposed algorithm is efficient both in terms of computation and memory. As a result, our method can be applied to data set with large sample sizes N and to estimate the ME of a target SNP to SNPs measured across the genome.

We performed extensive simulations to show that FAME provides calibrated tests of ME and has adequate power to detect true ME signals. We applied FAME to test for ME at trait-associated SNPs for 53 quantitative traits in the UK Biobank ($N \approx 300K$ corresponding to unrelated white British individuals and $M \approx 500K$ SNPs on the UKBB genotyping array). We explored the robustness of our ME signals to population stratification and to the possibility that the causal variants are missing on the UKBB array. To

better characterize the ME signals, we also attempted to partition ME signals to those that fall on the same chromosome and those that fall on the remaining chromosomes as the chromosome containing the target SNP and to estimate the proportion of variance explained by the ME effects.

2 Results

2.1 Methods overview

We aim to test whether the effect of a target SNP on a phenotype is modulated by the genetic background of the individual by assessing whether the pairwise interactions of the target SNP with each of the remaining SNPs contribute, in aggregate, to variance in the phenotype. In contrast to testing for interactions at a chosen pair of SNPs, this approach of testing for marginal epistasis (ME) can be more powerful when epistatic effects are polygenic, *i.e.*, we have a substantial number of interactions, each with a weak effect, while also benefiting from the reduced multiple testing burden. To ensure that additive genetic effects are not incorrectly attributed to interactions, we jointly model the additive effects from all genome-wide SNPs (including the target SNP) in addition to the ME effects.

Our method, **FA**st **M**arginal **E**pistasis test (FAME), uses a variance components model in which the phenotypic variance is partitioned into genome-wide additive genetic variance (σ^2_g), ME variance at a target SNP t ($\sigma^2_{gxg,t}$), and the residual variance (see Figure 1 for an example and Methods for additional details). The ME variance component $\sigma^2_{gxg,t}$ captures the aggregate contribution of all pair-wise interactions between the target SNP t and the remaining SNPs in the genome. We would like to test whether the ME variance component is significantly different from zero and, if it is, to be able to estimate its value.

Given a $N \times M$ genotype matrix $\boldsymbol{X}$ and a N-vector of phenotypes $\boldsymbol{y}$, we fit the following model [36]:

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{E}_t\boldsymbol{\alpha}_t + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim \mathcal{N}(\mathbf{0}, \sigma^2_e \boldsymbol{I}_N) \\ \boldsymbol{\beta} &\sim \mathcal{N}(\mathbf{0}, \frac{\sigma^2_g}{M}\boldsymbol{I}_M) \\ \boldsymbol{\alpha}_t &\sim \mathcal{N}(\mathbf{0}, \frac{\sigma^2_{gxg,t}}{M-1}\boldsymbol{I}_{M-1}) \end{aligned} \tag{1}$$

Here $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, $\boldsymbol{E}_t$ denotes a $N \times (M-1)$ gene-by-gene interaction matrix defined as $\boldsymbol{E}_t = \boldsymbol{X}_{-t} \odot \boldsymbol{X}_{:t}$ where $\boldsymbol{X}_{:t}$ is the t-th column of $\boldsymbol{X}$ and $\boldsymbol{X}_{-t}$ is formed by excluding the column $\boldsymbol{X}_{:t}$ from $\boldsymbol{X}$. In this model, σ^2_e, σ^2_g, and $\sigma^2_{gxg,t}$ are the residual variance, genetic variance and the ME variance components respectively. $\boldsymbol{\beta}$ denotes the additive effects while $\boldsymbol{\alpha}_t$ denotes the interaction effects between target SNP t and each of the other SNPs in the genome. This model assumes that the interaction effects are independent of the main effects so that epistasis is uncoordinated [37].

Fitting this model to Biobank-scale data, containing hundreds of thousands of individuals and millions of SNPs, is computationally impractical. FAME expands on our recent work [41, 42] to be able to test and estimate ME on Biobank-scale data. Specifically, FAME utilizes a randomized Method-of-Moments (MoM) estimator that reduces the size of the input genotype and interaction matrices by multiplying each of these matrices with a pre-specified number (B) of random vectors. We show that, even with small values of $B \approx 100$, this approach results in accurate estimates of the variance components resulting in a highly scalable method.

2.2 Calibration of FAME

First, we assessed the false positive rate of FAME by applying it to simulated phenotypes with additive genetic effects but no genetic interactions. We simulated phenotypes based on genotypes from unrelated white British individuals in the UKBB ($M = 459,792$ SNPs, $N = 291,273$ individuals). We set the proportion of trait heritability explained by additive genetic effects (additive heritability) $\sigma^2_g = 0.25$ and varied the proportion of variants $p \in \{0.01, 0.10\}$ that have non-zero additive effects (causal variants).

The key parameter in applying FAME is the number of random vectors B which determines its scalability and stability (see Section 4.2 of Materials and Methods for details). We use $B = 100$ in all our analyses (we explore the impact of this choice in Section 2.5.1). To obtain unbiased estimates, we also do not constrain the estimates of the variance components (allowing for negative estimates). We assessed the calibration of FAME when applied to two sets of target SNPs. The first set consists of target SNPs chosen randomly from across SNPs on the UKBB array. The second set consists of SNPs that were identified to have a significant additive effect based on a GWAS ($p < 5 \times 10^{-8}$) and was chosen to mirror our analyses of traits in UKBB (Section 2.5).

While FAME is calibrated when the target ME SNPs were selected at random (Supplementary Figure S1), the p-values tend to be inflated when the target SNPs were selected based on a GWAS (Section 2.5, Supplementary Figure S2). To address this issue, we excluded SNPs that lie within the LD block around the target SNP when constructing the set of genetic interactions $\boldsymbol{E}_t$ while retaining these SNPs in the additive component. This approach effectively controlled the false positive rate with no significant ME signal detected across any of the null simulation settings (Figure 2a).

Prior work has shown that tests of epistasis can suffer from inflated false positive rates due to imperfect tagging of causal variants [43, 44, 45]. To examine the robustness of FAME to such imperfect tagging, we repeated the simulations using imputed genotypes ($N = 291,273, M = 4,824,392$) while still applying FAME to analyze SNPs on the UKBB array. We observed that FAME remains calibrated indicating its robustness to imperfect tagging (Supplementary Figure S3).

2.3 Power analysis

We analyzed the power of FAME by simulating phenotypes with non-zero ME variance components under the model defined in Equation 1. We fixed the variance explained by additive effects, σ^2_g to 0.3, which is roughly the median estimated additive heritability across all the traits we tested. We then varied the proportion of variance explained by ME ($\sigma^2_{gxg,t}$ at target SNP t) and analyzed the power of FAME to detect ME (see Section 4.3 of Materials and Methods for details). We observed FAME has power $\geq 90\%$ at a stringent p-value threshold ($p < 5 \times 10^{-8}$) even when the variance explained by ME is fairly low ($\sigma^2_{gxg,t} = 0.005$) (Figure 2b). Further, we observed that the ME estimates were accurate and exhibited minimal bias (Figure 2d).

2.4 Computational efficiency

We attempted to benchmark a previously proposed method for ME testing, MAPIT [36], over datasets of various sizes (see Section 4.4 of Materials and Methods for details). We observe that the computational complexity of MAPIT grows rapidly with increasing sample size even for modest sample sizes and number of SNPs: requiring more than three days to run on 20K samples with 10K SNPs. This is, partly, because MAPIT did not provide the flexibility of modifying the variant testing strategy and, by default, tests the ME effect across all the SNPs provided. Thus, it is not feasible to run MAPIT on a large-scale dataset like

UKBB (Supplementary Figure S4). Moreover, MAPIT requires loading the whole genotype matrix at once which we extrapolate would require more than 200 GB for UK Biobank size dataset. On the other hand, FAME can test ME on 500K individuals on a genome-wide dataset containing $\approx$ 500K SNPs in less than 4 hours (Figure 2c).

2.5 Application to UK Biobank phenotypes

We applied FAME to test for ME in 53 quantitative traits measured across $N = 291,273$ unrelated white British individuals with genotypes measured across common SNPs on the UK Biobank array (see Section 4.5 for details on datasets). Our target SNPs consisted of SNPs that were found to be associated with the trait in a GWAS. Specifically, we ran GWAS on each of the traits, including covariates such as sex, age, and the top 20 genetic PCs. For each trait, we selected SNPs with p-value $p < 5 \times 10^{-8}$ followed by LD pruning (using a window size of 500 SNPs, we computed r^2 between each pair and removed one of them if $r^2 > 0.1$, shifting the window by 1 SNP, and repeating the process). We tested for ME at the resulting set of $15,601$ GWAS significant SNPs in which we also accounted for the linear additive effect of genome-wide SNPs and included age, sex, and the top 20 genetic PCs as fixed effect covariates. Following our calibration experiments, SNPs in the LD block surrounding the target SNP were excluded from the set of genetic interactions. Our tests yielded 21 significant trait-loci pairs across 13 traits ($p < \frac{5\times10^{-8}}{53}$ to account for the multiple traits tested). To additionally ensure that the additive genetic effects surrounding the target SNP do not impact estimates of ME, we applied FAME to each of these 21 trait-loci pairs after regressing out all of the SNPs in the LD block of the target SNP to observe 16 trait-loci pairs that retain significant p-values for ME ($p < \frac{5\times10^{-8}}{53}$; Figure 3a; Table 1).

2.5.1 Stability of significant ME signals

We first explored the impact of the randomization underlying FAME on our results. We selected two traits: body mass index (BMI), for which we did not detect a significant ME locus, and serum urate levels (Urate), for which we detected a significant ME locus. We computed the Pearson correlation of the negative log p-value between results of FAME run with different seeds (ρ). We experimented with the number of random vectors (B) and observed that using $B = 100$ random vectors yields consistent results ($\rho = 0.99$ for Urate; $\rho = 0.98$ for BMI; Supplementary Figure S6). Second, we reran FAME for the 16 significant trait-loci pairs using five different random number seeds and found that the results are concordant across seeds (15 of the trait-loci pairs show $p < \frac{5\times10^{-8}}{53}$ across all seeds while all of the trait-loci pairs show $p < 5 \times 10^{-8}$ across all seeds; Supplementary Table S1). These results indicate that FAME yields stable estimates of ME.

2.5.2 Robustness of significant ME signals

Population stratification in GWAS is commonly accounted for by including principal components (PCs) computed from genotype data as covariates in the analysis [46, 47]. To explore the effect of population stratification, we reran our analyses on trait-loci pairs previously discovered as significant with the number of PCs included as covariates increased to 40 (from 20). We observe a high correlation in the p-values when using 40 vs 20 PCs (Table 1, Supplementary Figure S7a; Pearson correlation $\rho = 0.997$). Importantly, 15 of the 16 significant trait-loci pairs remain significant after including the top 40 PCs, indicating that our findings are robust to population stratification (with the remaining trait-locus pair continuing to exhibit a low p-value).

A second concern with our analyses arises from the fact that the UK Biobank array might miss true causal variants which could lead to the inference of spurious epistatic effects [43, 44, 45]. Our simulations in Section 2.2 show that FAME remains calibrated in this setting. To further explore the robustness of our results, we analyzed our significant ME signals on $4,824,392$ imputed SNPs (MAF $> 1\%$). We observed 13 out of the 16 significant trait-loci pairs detected on the array dataset were significant ($p \leq \frac{5\times10^{-8}}{53}$) on the imputed dataset (the remaining three loci had p-values $p \leq 10^{-6}$ on the imputed dataset; Table 1; Pearson correlation of the p-values $\rho = 0.673$; Supplementary Figure S7b).

Third, we observe that SNPs with significant ME effects were associated with lower MAF compared to GWAS significant SNPs (Supplementary Table S2). To evaluate the calibration of FAME at low MAF SNPs, we repeated our null simulations described in Section 2.2 restricting to low MAF candidate causal variants (MAF $\in [0.01, 0.05]$). We confirm that FAME remained calibrated across all tested settings (Supplementary Figure S5).

It is well-known that scale of phenotype measurement can affect approaches to test and interpret epistasis. If epistatic effects arose due to choice of scale, we would expect a genome-wide impact for the associated SNPs. However, we do not observe widespread inflation in tests of ME suggesting that the choice of scale is unlikely to impact our results. To further explore the impact of scale, we selected one of the phenotypes (height) and randomly chose 100 GWAS significant SNPs as target SNPs. We then ran FAME by changing the scale of the trait considering two possible transformations: $y^{pow3} := y^3$ and $y^{exp} := exp(y)$. We then compared the p-value of the ME test to those obtained by analyzing height on the original scale $y^{orig} := y$. We noticed that by changing the scale of the target trait, the p-values from FAME were significantly inflated (Supplementary Figure S9), thus suggesting that the ME signals discovered by FAME are unlikely to arise due to the scale on which traits are measured.

2.5.3 Localizing signals of ME

Having demonstrated evidence for genome-wide ME, we sought to understand where these interactions localize. As a first step towards answering this question, we extended FAME to test for ME of a target SNP with only a subset of SNPs while accounting for the additive effects of genome-wide SNPs. We separately tested for ME of the target SNP with other SNPs that fall on the same chromosome (local ME; denoted as gxg_{local}) and the ME of the target SNP with SNPs located on chromosomes distinct from the chromosome containing the target SNP (distal ME; denoted as gxg_{dist}). We first confirmed that tests of $\sigma^2_{gxg,dist}$ and $\sigma^2_{gxg,local}$ are well-calibrated in simulations (Supplementary Figure S8; see Section 4.3 of Materials and Methods for details). Applying the localization test to each of 16 previously identified ME loci, we found 6 and 15 loci with significant local and distal ME effects respectively ($p \leq \frac{5\times10^{-8}}{53}$; Figure 3b; Supplementary Table S3).

2.5.4 Magnitude of ME effects

We estimated the proportion of trait variance explained by ME (ME heritability) at a target SNP t ($h^2_{gxg,t}$) from the variance components estimated by FAME (Supplementary Information Section S1). Across the 16 trait-loci pairs with significant ME signal, estimates of h^2_{gxg} tend to be modest: $10^{-3} - 10^{-2}$. We compared these estimates to the heritability of the SNP based on its GWAS effect size ($h^2_{gwas,t}$ estimated as the square of the GWAS effect size for a standardized genotype). We find that the h^2_{gxg} estimates are substantially larger than the corresponding h^2_{gwas} estimates: about $12x$ larger on average with a range of 0.59 to 43.89 (Figure 3c; Table 2). Estimates of h^2_{gxg} are not strongly correlated with the h^2_{gwas} estimates ($\rho = 0.022$).

2.5.5 Interpreting loci with significant ME effects

We observe the largest ratio of h^2_{gxg} to h^2_{gwas} at SNP rs628031 (chr6:160,560,845) that shows significant ME for serum lipoprotein A levels (lipoA). This variant is a non-synonymous polymorphism that changes methionine to valine in the protein product of the organic cation transporter gene OCT1 (also known as SLC22A1). OCT1 mediates the uptake and efflux of cationic metabolites in the liver that includes as its substrates a variety of drugs including metformin that is widely used to treat type 2 diabetes [48]. Genetic variation in OCT1 has been shown to modulate the response to metformin and to other drugs [48].

SNP rs964184 (chr11:116,648,917) shows significant ME for multiple traits: Apolipoprotein B, cholesterol, and triglycerides with substantial ME effects ($\frac{h^2_{gxg}}{h^2_{gwas}}$ = 5.14, 7.83, and 0.59 respectively). This variant lies in the 3' UTR region of the ZPR1 gene (also referred to as ZNF259) that encodes a zinc finger protein that is known to play a regulatory role in cell proliferation and signal transduction [49]. The promoter region of ZPR1 is known to be bound by transcription factors that play a role in insulin sensitivity, cholesterol metabolism, and obesity. rs964184, as well as other variants in ZPR1, have been found to be associated with serum LDL-C [50], HDL-C [51], triglyceride levels [50, 52, 53] and risk for coronary artery disease (CAD) [54] in diverse populations. A regulatory role for rs964184 has been suggested based on its location in a DNaseI hypersensitive region and its overlap with an enhancer that is active in tissues relevant for lipid biology [52]. Further, rs964184 has been association with DNA methylation of a CpG site in the promoter region of the APOA5 gene [55], potentially explaining the association between DNA methylation level at this site and triglyceride levels [56]. Integrative analyses of genotype and gene expression data have shown rs964184 to play a regulatory role: being a cis-eQTL for genes PCSK7, SIDT2, TAGLN, and BUD13 while also a trans-eQTL for TMEM165, YPEL5, PPM1B, and OBFC2A [57]. Further, mediation analyses revealed that a substantial proportion of the effect of rs964184 on HDL-C and triglycerides is mediated through its trans association with PPM1B and YPEL5 [57].

3 Discussion

We have presented a new method, FAME, that can detect marginal epistasis (ME) in Biobank-scale data. FAME yields calibrated results in simulations. Applying FAME to 53 quantitative phenotypes in the UK Biobank, we found 16 trait-loci pairs with significant signals of ME, a vast majority of which remain significant after testing with additional PCs to correct for population stratification, and on imputed genotypes to reduce the impact of missing causal SNPs. To the best of our knowledge, this work is the first to show evidence of interaction effects between individual genetic variants and overall polygenic background modulating complex trait variation. While the number of loci showing ME effects is modest (in part due to the stringent p-value threshold that we impose and the GWAS selection strategy that we used to identify target SNPs), we observe that the proportion of variance explained by ME is comparable to, and sometimes substantially larger than, the proportion of variance explained by GWAS. These results show that the polygenic background can substantially modulate the effect of a genetic variant on trait and has implications for efforts to interpret genetic variant effects, to improve phenotype prediction, and to understand how genetic effects vary across populations [7].

We further partitioned the ME signal within and across chromosomes to detect both within and cross-chromosomal signals and found 6 within chromosomal signals, which is a strict subset of the 15 cross-chromosomal signals. This observation suggests that the epistatic signal that we detect is likely to be polygenic so that the approach of testing for the aggregate effects as we do here is likely to be more powerful

than an approach that aims to identify specific pairs of SNPs. While our current application of FAME has focused on genome-wide signals of ME where we test a single target SNP against a background set consisting of SNPs across the genome (excluding those in the LD block as the target), the model underlying FAME is flexible and can be applied to test for epistasis in other settings. For example, FAME can be extended to test for interactions of a target SNP or other covariates (such as polygenic scores) with a background set of SNPs where the set is defined based on functional annotation such as genes or pathways. The ideas underlying FAME allow such tests to be applied to biobank-scale data. Such an approach can improve on our understanding by attempting to localize the ME signal. Additionally, the model underlying FAME assumes that epistasis is uncoordinated, *i.e.*, the interaction effects are independent of main effects. It would be of interest to extend our method to settings where epistasis is coordinated [37].

Our work has several limitations. First, it is plausible that the impact of population structure on epistatic effects might not be well-modeled by the approaches employed here (such as the inclusion of principal components based on common genetic variants). Second, prior studies have shown that tests of epistasis can have inflated false positive rates due to imperfect tagging of causal variants that have large additive effects [43]. Our simulations show that FAME is robust to imperfect tagging of causal variants. Further, the replication of signals discovered using array SNPs on imputed SNPs, that are unlikely to miss causal variants that are common in the population, makes the issue of missing causal variants less likely. Nevertheless, it is plausible that the distributions of causal variants and the LD patterns between causal and genotyped variants could be complex which could impact the calibration of our method. Third, the scale on which phenotypes are measured can affect our results (as is true of other approaches to detect epistasis). Our simulations applying FAME to rescaled versions of phenotypes and the observation that we find SNPs with significant and non-significant ME indicate that our results are not simply driven by scale. Fourth, our estimates of ME effects are likely to be biased upwards due to winner's curse [58]. Fifth, despite its scalability, FAME is still not efficient enough to perform genome-wide scans of ME which, in turn, led us to focus on testing for ME at GWAS loci. Extending the scope and efficiency of FAME present important directions for future work.

4 Materials and Methods

4.1 Marginal epistasis model

Given a $N \times M$ genotype matrix $\boldsymbol{X}$, a N-vector of phenotypes $\boldsymbol{y}$ and a target SNP $t \in \{1, \ldots, M\}$, we aim to jointly test the additive effect of the M SNPs and the ME of the target SNP based on the following model that was originally introduced in [36]:

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{E}_t\boldsymbol{\alpha}_t + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim \mathcal{N}(\boldsymbol{0}, \sigma_e^2 \boldsymbol{I}_N) \\ \boldsymbol{\beta} &\sim \mathcal{N}(\boldsymbol{0}, \frac{\sigma_g^2}{M} \boldsymbol{I}_M) \\ \boldsymbol{\alpha}_t &\sim \mathcal{N}(\boldsymbol{0}, \frac{\sigma_{gxg,t}^2}{M-1} \boldsymbol{I}_{M-1}) \end{aligned} \tag{2}$$

Here $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, $\boldsymbol{E}_t$ denotes a $N \times (M-1)$ gene-by-gene interaction matrix defined as $\boldsymbol{E}_t = \boldsymbol{X}_{-t} \odot \boldsymbol{X}_{:t}$ where $\boldsymbol{X}_{:t}$ is the t-th column of $\boldsymbol{X}$ and $\boldsymbol{X}_{-t}$ is formed by excluding the column $\boldsymbol{X}_{:t}$ from $\boldsymbol{X}$.

In this model, σ_e^2, σ_g^2, and $\sigma_{gxg,t}^2$ are the residual variance, genetic variance and the ME variance components respectively. $\boldsymbol{\beta}$ denotes M-vector of SNPs effect sizes and $\boldsymbol{\alpha}_t$ denotes $M-1$-vector of interaction effects between target SNP t and each of the other SNPs in the genome.

We assume without loss of generality that $\boldsymbol{y}$ is centered and the columns of $\boldsymbol{X}$ are standardized. To estimate the variance components of our LMM, we use a Method-of-Moments (MoM) estimator that searches for parameter values so that the population moments are close to the sample moments. Since $\mathbb{E}[\boldsymbol{y}] = 0$, we derived the MoM estimates by equating the population covariance to the empirical covariance. The population covariance is given by:

$$\boldsymbol{\Sigma} = cov(\boldsymbol{y}) = E[\boldsymbol{y}\boldsymbol{y}^T] - E[\boldsymbol{y}]E[\boldsymbol{y}^T] = \sigma_g^2 \frac{1}{M} \boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} + \sigma_{gxg,t}^2 \frac{1}{M-1} \boldsymbol{E}_t\boldsymbol{E}_t^{\mathrm{T}} + \sigma_e^2 \boldsymbol{I} \tag{3}$$

Using $\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}$ as our estimate of the empirical covariance, we need to solve the following least squares problem to estimate the variance parameters :

$$(\tilde{\sigma}_g^2, \tilde{\sigma}_{gxg,t}^2, \tilde{\sigma}_e^2) = argmin_{(\sigma_g^2, \sigma_{gxg,t}^2, \sigma_e^2)} ||\boldsymbol{y}\boldsymbol{y}^T - \left(\sigma_g^2 \boldsymbol{K}_1 + \sigma_{gxg,t}^2 \boldsymbol{K}_{2,t} + \sigma_e^2 \boldsymbol{K}_3\right)||_F^2 \tag{4}$$

where $\boldsymbol{K}_1 = \frac{1}{M}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$, $\boldsymbol{K}_{2,t} = \frac{1}{M-1}\boldsymbol{E}_t\boldsymbol{E}_t^{\mathrm{T}}$ and $\boldsymbol{K}_3 = \boldsymbol{I}_N$.

We show that the MoM estimator satisfies the following normal equations (see Lemma 1 in Supplementary Notes):

$$\boldsymbol{T}\boldsymbol{\sigma}^{\boldsymbol{2}} = \boldsymbol{q} \tag{5}$$

where $\boldsymbol{T}$ is a 3×3 matrix with entries $T_{kl} = tr(\boldsymbol{K}_k\boldsymbol{K}_l), k, l \in \{1, 2, 3\}$, $tr()$ denotes the trace of the matrix, and $\boldsymbol{q}$ is a 3-vector with entries $q_k = \boldsymbol{y}^T\boldsymbol{K}_k\boldsymbol{y}$.

To compute the variance components $\tilde{\boldsymbol{\sigma}^{\boldsymbol{2}}} = (\tilde{\sigma}_g^2, \tilde{\sigma}_{gxg,t}^2, \tilde{\sigma}_e^2)^T$, we have:

$$\tilde{\boldsymbol{\sigma}^{\boldsymbol{2}}} = \boldsymbol{T}^{-1}\boldsymbol{q} \tag{6}$$

Beyond the point estimates, we also need to compute confidence intervals for $\tilde{\boldsymbol{\sigma}^{\boldsymbol{2}}}$ which, in turn, allow us

to test the hypothesis of no ME ($\sigma^2_{gxg,t} = 0$). To do this, we compute the covariance matrix of $\tilde{\boldsymbol{\sigma^2}}$ as (see Lemma 2 in Supplementary Notes):

$$\text{Cov}\left[\tilde{\boldsymbol{\sigma^2}}\right] = \boldsymbol{T}^{-1}\text{Cov}\left[\boldsymbol{q}\right]\boldsymbol{T}^{-1}$$

where

$$\text{Cov}\left[\boldsymbol{q}\right] = E[\boldsymbol{q}\boldsymbol{q}^T] - E[\boldsymbol{q}]E[\boldsymbol{q}]^T$$

such that

$$\begin{aligned}\text{Cov}\left[\boldsymbol{q}\right]_{ij} &= Cov(\boldsymbol{y}^T\boldsymbol{K}_i\boldsymbol{y}, \boldsymbol{y}^T\boldsymbol{K}_j\boldsymbol{y}) \\ &= 2tr(\boldsymbol{\Sigma}\boldsymbol{K}_i\boldsymbol{\Sigma}\boldsymbol{K}_j) \end{aligned} \tag{7}$$

4.2 Efficient computation of variance components

Computing the coefficients $tr(\boldsymbol{K}_k\boldsymbol{K}_l)$ of the system of linear equation 5 and $\text{Cov}\left[\boldsymbol{q}\right]_{ij}$ require $\mathcal{O}(N^2M)$ time complexity and $\mathcal{O}(NM)$ memory usage imposing challenging memory and computation requirements for Biobank-scale data (N in the hundreds of thousands, M in the millions). To test hypotheses, we need to compute p-values. This requires computing the point estimate and standard error which is appropriate when we have large sample sizes N.

To obtain an efficient estimate of $\tilde{\boldsymbol{\sigma^2}}$, we approximate each of the coefficients of the matrix $\boldsymbol{T}$ which involves computing the trace of a matrix by an unbiased trace estimator [59]. Specifically, we estimate $T_{i,j}$ as follows:

$$\begin{aligned} T_{i,j} = tr(\boldsymbol{K}_i\boldsymbol{K}_j) &= \frac{1}{M_iM_j}tr(\boldsymbol{Z}_i\boldsymbol{Z}_i^T\boldsymbol{Z}_j\boldsymbol{Z}_j^T) \\ \hat{T}_{i,j} &\approx \frac{1}{BM_iM_j}\sum_{b=1}^{B}\boldsymbol{v_b}^T\boldsymbol{Z}_i\boldsymbol{Z}_i^T\boldsymbol{Z}_j\boldsymbol{Z}_j^T\boldsymbol{v_b} \end{aligned} \tag{8}$$

where each $\boldsymbol{v_b}$ is an independent random vector with mean zero and covariance $\boldsymbol{I}_N$, B is the total number of random vectors used for the approximation, and $\boldsymbol{Z}_i = \boldsymbol{X}$ or $\boldsymbol{E}$ with M_i columns.

To estimate $\text{Cov}\left[\boldsymbol{q}\right]_{ij}$ efficiently, we replace $\boldsymbol{\Sigma}$ in Equation 7 to obtain the plug-in estimate of $\text{Cov}\left[\boldsymbol{q}\right]_{kl}$:

$$\begin{aligned}\widehat{\text{Cov}\left[\boldsymbol{q}\right]}_{kl} &= 2\boldsymbol{y}^{\mathrm{T}}\boldsymbol{K}_k\tilde{\boldsymbol{\Sigma}}\boldsymbol{K}_l\boldsymbol{y} \\ &= 2\boldsymbol{y}^{\mathrm{T}}\boldsymbol{K}_k\left(\sum_{t=1}^{3}\tilde{\sigma_t^2}\boldsymbol{K}_i\right)\boldsymbol{K}_l\boldsymbol{y} \\ &= 2\left(\sum_{i=1}^{3}\tilde{\sigma_i^2}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{K}_k\boldsymbol{K}_i\boldsymbol{K}_l\boldsymbol{y}\right) \\ &= 2\sum_{t=1}^{3}\tilde{\sigma_i^2}\left(\boldsymbol{w}_k^{\mathrm{T}}\frac{\boldsymbol{Z}_i\boldsymbol{Z}_i^{\mathrm{T}}}{M_i}\boldsymbol{w}_l\right)\end{aligned}$$

where $\boldsymbol{w}_i = \boldsymbol{K}_i\boldsymbol{y}, i \in \{1, \ldots, 3\}$.

The computation of $\widehat{\text{Cov}\left[\boldsymbol{q}\right]}_{kl}$ can be performed in $\mathcal{O}(NM)$ time. Multiplication of $\boldsymbol{w}_k = \boldsymbol{K}_k\boldsymbol{y}$ can be decomposed into $\boldsymbol{Z}_k\boldsymbol{Z}_k^T\boldsymbol{y}$ which can be computed in $\mathcal{O}(MN)$ time. Further, by leveraging the fact that

the matrices are discrete-valued genotype matrices, we can improve the time complexity of matrix-vector multiplication from $\mathcal{O}(NM)$ to $\mathcal{O}(\frac{NM}{max(\log_3 N, \log_3 M)})$ by using the Mailman algorithm [60] . Hence, $\hat{T}_{i,j}$ and $\widehat{\text{Cov}}[\boldsymbol{q}]_{kl}$ can be computed in time $\mathcal{O}(\frac{NMB}{\max(\log_3(N),\log_3(M))})$ and $\mathcal{O}(\frac{NM}{\max(\log_3(N),\log_3(M))})$ respectively. FAME uses a streaming implementation that does not require all the genotypes to be stored in memory leading to scalable memory requirements with $\mathcal{O}(NS)$ where S is the number of SNPs per each stream block . We have shown that we can compute point estimates and the corresponding standard errors in sub-linear time with respect to sample size N and number of SNPs M. Therefore, we can apply our method to data sets with high sample sizes and test for the existence of ME. Finally, we note that FAME can also account for fixed-effects covariates such as age, sex, and genetic principal components (PCs) (Supplementary Information Section S2).

4.3 Simulations

Simulations to assess power and accuracy

We designed simulations to assess the power of FAME and the accuracy of its ME variance components estimates. We used the following generative model:

$$
\begin{aligned}
\boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{E}_t\boldsymbol{\alpha}_t + \boldsymbol{\epsilon} \\
\boldsymbol{\epsilon} &\sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_N) \\
\boldsymbol{\beta}_j &\overset{i.i.d}{\sim} \begin{cases} \mathcal{N}(0, \frac{\sigma_g^2}{|M_a|}) & \text{if } j \in M_a \\ 0 & \text{otherwise} \end{cases} \\
\boldsymbol{\alpha}_{t,j} &\overset{i.i.d}{\sim} \begin{cases} \mathcal{N}(0, \frac{\sigma_{gxg,t}^2}{|M_e|}) & \text{if } j \in M_e \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

where β_j and $\alpha_{t,j}$ denotes the j^{th} element in the respective vectors of effect sizes. We set σ_g^2 to 0.3 which is approximately the median value of the additive heritability across all the traits that we analyzed in this study. We varied the value of $\sigma_{gxg,t}^2$ from 0.001 to 0.1. We randomly selected 10% of the SNPs to be causal for the additive effects (assigned to the indicator set M_a) and 10% of the SNPs to be causal for the ME effect (assigned to the indicator set M_e which do not overlap with M_a and fall outside the LD block of the target SNP). As target SNP, we selected three representative SNPs with different MAF values (MAF $\in \{1\%, 14\%, 49\%\}$ respectively). For computational convenience, we limited our analysis to chromosomes 12 and 20 of the UKBB data, which we used as our $\boldsymbol{X}$ matrix. We simulated $1,000$ replicates for each setting. In order to assess the accuracy of the ME variance component estimates obtained by FAME, we used the same simulations as above. We then assumed that the target SNPs were known and then estimated the ME effect by partitioning the SNPs in $\boldsymbol{X}$ into two bins, LD block, which contains all the SNPs within the LD region of the target SNP; LD removed block, which contains all the SNPs outside of the LD region of the target SNP. We then used FAME to jointly fit the additive effect for both regions while only fitting the ME effect on the LD-removed region. Finally, we compared the estimated ME variance components with the ground truth.

Regional simulation and estimation

To localize the ME signal, we partitioned the whole genome into the region with all the SNPs lying in the same chromosome as the target SNP but outside of the LD block (termed as *local*) and all the SNPs lying on chromosomes different from the one with the target SNP (termed as *distal*). To validate the calibration of FAME when applied to test the ME effect on a specified region, we used the simulation with a total heritability of 0.25 and ratio of causal SNPs of 1%. We applied FAME to estimate the calibration of $\sigma^2_{gxg,local}$ and $\sigma^2_{gxg,distal}$ respectively.

4.4 Runtime comparisons

All experiments used a machine equipped with AMD EPYC 7501 32-Core Processor and a runtime budget of 3 days was provided to all tested methods.

4.5 Datasets

Simulation dataset

We obtained a set of $N = 291,273$ unrelated white British individuals measured at $M = 459,792$ common SNPs genotyped on the UK Biobank Axiom array to use in simulations by extracting individuals that are $>$ 3rd-degree relatives and excluding individuals with putative sex chromosome aneuploidy. Unless otherwise specified, all simulations were conducted using this dataset.

UKBB genotypes

For analysis of real traits, we restricted our analysis to SNPs that were presented in the UK Biobank Axiom array used to genotype the UK Biobank. SNPs with greater than 1% missingness and minor allele frequency smaller than 1% were removed. Moreover, SNPs that fail the Hardy-Weinberg test at significance threshold 10^{-7} were removed. We restricted our study to self-reported British white ancestry individuals which are $> 3^{rd}$ degree relatives that is defined as pairs of individuals with kinship coefficient $< 1/2^{(9/2)}$ [61]. Furthermore, we removed individuals who are outliers for genotype heterozygosity and/or missingness and excluded SNPs that fall within the MHC region. Finally, we obtained a set of $N = 291,273$ individuals and $M = 454,207$ SNPs for real data analyses. We used this dataset in our analyses unless specified otherwise.

We also analyzed imputed genotypes across $N = 291,273$ unrelated white British individuals. We removed SNPs with greater than 1% missingness, minor allele frequency smaller than 1%, SNPs that fail the Hardy-Weinberg test at significance threshold 10^{-7} as well as SNPs that lie within the MHC region (Chr6: 25–35 Mb) to obtain $4,824,392$ SNPs.

Covariates and phenotypes

We selected 53 quantitative traits in the UKBB. The selected phenotypes span eight known categories: Anthropometry, Blood Biochemistry, Bone, Cardiovascular, Diabetes, Eye, Liver, and Renal. We included sex, age, and the top 20 genetic principal components (PCs) as covariates in our analysis for all phenotypes. Extra covariates were added for diastolic/systolic blood pressure (adjusted for cholesterol-lowering medication, blood pressure medication, insulin, hormone replacement therapy, and oral contraceptives). We used the PCs computed in the UKBB from a superset of $488,295$ individuals. Following prior studies, all traits were inverse rank normalized [62, 63].

Data availability

The UK Biobank dataset used in this study is not publicly available but can be obtained by application (https://www.ukbiobank.ac.uk/).

Code availability

FAME can be found at https://github.com/sriramlab/FAME. The simulator used in the experiments can be found at https://github.com/alipazokit/simulator. MAPIT can be found at https://github.com/lorinanthony/MAPIT.

Acknowledgments

This research was conducted using the UK Biobank Resource under application 33127. We thank the participants of UK Biobank for making this work possible. This work was supported, in part, by NIH grants GM125055 (B.F., A.P., and S.S) and HG006399 (S.S.), and NSF grant CAREER-1943497 (B.F., A.P., and S.S.). N.Z. was supported by NIH grants R01MH130581, U01MH126798, R01MH122688, and R01GM142112.

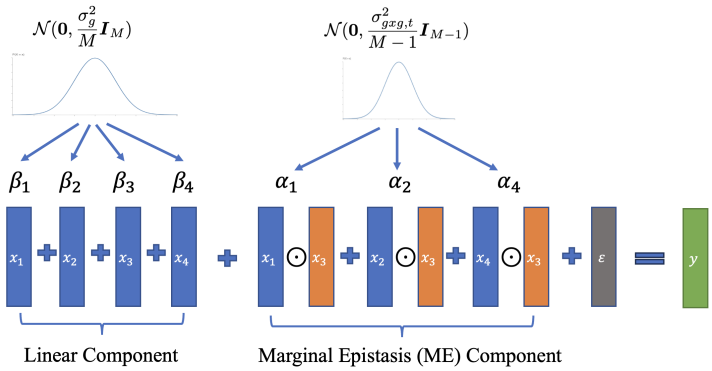
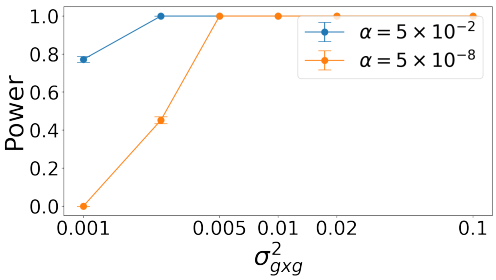
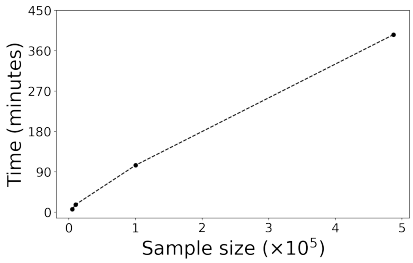


Figure 1: **The model underlying FAME**. In this example, we have genotypes at four SNPs denoted by $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$. We would like to test for marginal epistasis (ME) between SNP 3 (the target SNP) and the remaining SNPs. We model the relationship of the phenotype $\boldsymbol{y}$ to the genotypes as arising due to the additive effect of genotypes at each of the four SNPs, the pairwise interaction effects between genotypes at the target SNP ($\boldsymbol{x}_3$) and the remaining SNPs, and environmental noise $\boldsymbol{\epsilon}$. The additive effect sizes $\boldsymbol{\beta}$ are drawn from a distribution with variance parameter proportional to σ_g^2 while the ME effect sizes are drawn from a distribution with variance parameter proportional to $\sigma_{gxg,t}^2$ where $t = 3$.

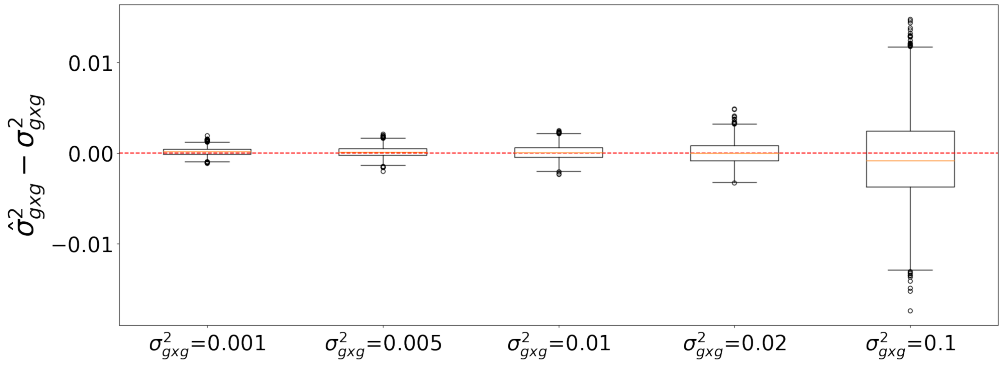
(a) QQ-plots under null simulations



(b) Power analysis of FAME

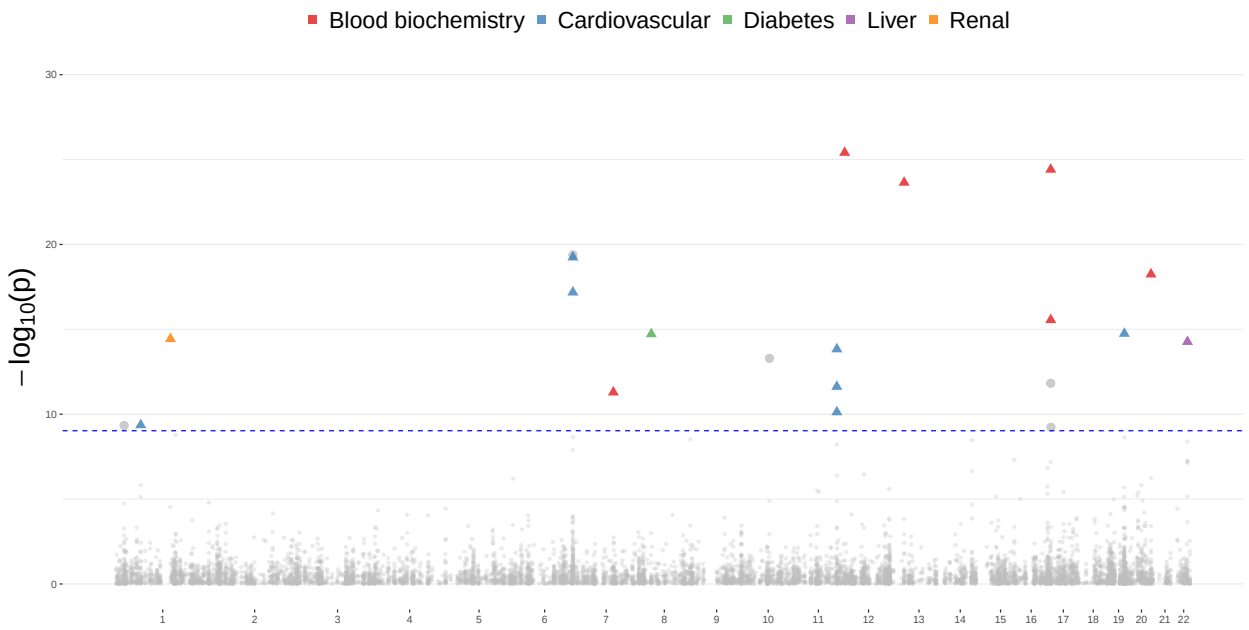


(c) Runtime of FAME with increasing sample size

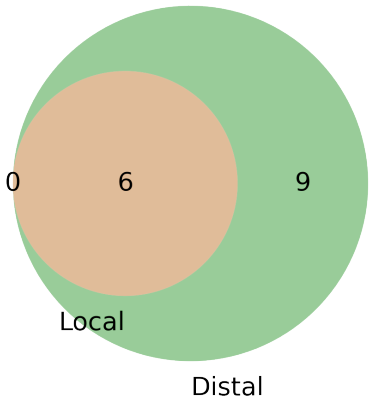


(d) Accuracy of FAME

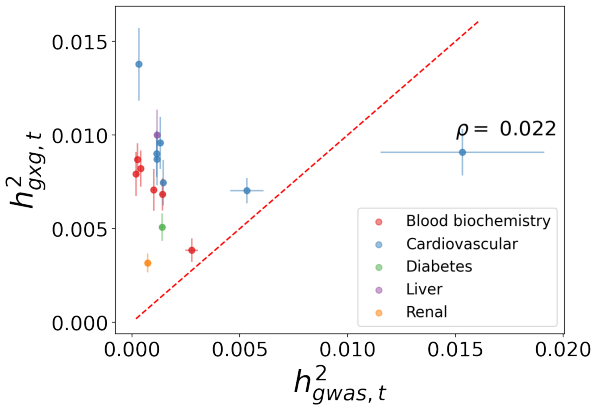
Figure 2: **Accuracy and runtime analysis of FAME.** (a) QQ-plot of FAME in simulations. We applied FAME to phenotypes simulated from genotypes with linear additive effects but no marginal epistatic (ME) effects. Phenotypes were simulated using genotypes measured on $\approx 300K$ unrelated white-British individuals in the UK Biobank, with varying ratio of causal SNPs (Causal ratio) and heritability (h^2). We first ran GWAS to identify significant SNPs which were then used as target SNPs in a test of ME. We detected no significant ME signals ($p \leq 5 \times 10^{-8}$) across all the settings. (b) Power analysis of FAME. We simulated phenotypes by fixing the additive variance component to 0.3 (roughly the median value estimated across real traits). We varied the strength of the ME variance component (σ^2_{gxg}). Each setting was simulated $1,000$ times.We plot power for detecting ME at a p-value threshold of 0.05 as well as the genome-wide threshold of 5×10^{-8} averaged across the replicates. (c). Runtime analysis of FAME. We computed the runtime of FAME applied to common SNPs on the UKBB whole-genome array data and varying sample size. We ran the experiment three times at each setting and reported the average runtime. (d). Accuracy of estimates of the ME variance component (σ^2_{gxg}) in simulations. We used exactly the same simulation as in (b). We plot the error in the parameter estimates (defined as $\hat{\sigma}^2_{gxg} - \sigma^2_{gxg}$) for each parameter setting.



(a) Manhattan plot of FAME p-value distribution



(b) Localization of ME signals



(c) $h^2_{gxg,t}$ vs. $h^2_{gwas,t}$

Figure 3: **ME signals in the UKBB.** (a) Manhattan plot of the ME loci across 53 complex traits in UKBB. Colored shapes denote trait-loci pairs that are significant at $p \leq \frac{5\times10^{-8}}{53}$; shapes with colored triangles were the loci that are statistically significant in our initial analysis and after we regressed out all SNPs within the LD block as fixed effects. (b) Localization of ME signals. For each of 16 trait-loci pairs, we tested whether the ME signals remained significant when testing against all SNPs on the same chromosome as the target SNP (after removing SNPs in the same LD block as the target SNP), which we term *local*, and against all SNPs on chromosomes different from the chromosome containing the target SNP, which we term *distal*. We then compared the overlap between the *local* and *distal* significant signals ($p \leq \frac{5\times10^{-8}}{53}$). (c) We compared the fraction of phenotypic variance explained by marginal epistatic effects ($h^2_{gxg,t}$) to the fraction of phenotypic variance explained by GWAS (denoted as the $h^2_{gwas,t}$) for trait-loci pairs that show significant ME. Vertical (horizontal) bars denote the standard error of $h^2_{gxg,t}$ ($h^2_{gwas,t}$).

Table 1: **Trait-loci pairs with evidence for significant marginal epistasis (ME)**. *Array* denotes tests of ME run on SNPs genotyped on the UKBB array; *PC40 array* denotes tests performed with the top 40 PCs regressed out instead of top 20 PCs as in the original analysis; *Imputed* represents tests on imputed SNPs. p denotes the p-value corresponding to the ME effect while σ^2_{gxg} denotes the estimate of the ME variance at the tested SNP. p-values that passed the significance threshold ($\frac{5\times10^{-8}}{53}$) for *PC40 array* and *Imputed* have been highlighted.

			Array		*PC40 Array*		*Imputed*	
			σ^2_{gxg}	p	σ^2_{gxg}	p	σ^2_{gxg}	p
Trait	CHR	SNP	$\times 0.01$	$-log_{10}$	$\times 0.01$	$-log_{10}$	$\times 0.01$	$-log_{10}$
Alanine aminotransferase	22	44,388,817	1.09	14.27	1.10	**14.53**	0.70	**10.16**
Apolipoprotein B	11	116,648,917	0.81	10.12	0.83	**10.40**	0.47	6.055
C-reactive protein	1	66,257,838	0.89	9.36	0.85	8.792	0.59	6.157
Cholesterol	11	116,648,917	0.96	13.83	0.94	**13.21**	0.56	8.080
Hemoglobin A1c	8	41,542,093	0.65	14.72	0.66	**14.97**	0.58	**15.05**
Lipoprotein-A	6	160,578,069	1.42	19.24	1.41	**19.22**	1.50	**41.33**
		160,560,845	1.77	17.18	1.79	**17.43**	1.09	**13.52**
	19	45,414,399	1.15	14.75	1.14	**14.61**	0.77	**10.04**
Mean platelet volume	20	57,597,970	0.89	18.25	0.88	**17.82**	0.73	**16.03**
Monocyte count	13	28,623,048	0.79	23.65	0.79	**23.92**	0.66	**21.49**
SHBG	17	7,145,117	1.06	24.41	1.09	**25.12**	0.91	**21.80**
		7,254,315	0.98	15.55	0.98	**15.56**	0.74	**12.82**
Testosterone	7	99,032,593	0.82	11.29	0.81	**11.13**	0.63	**9.32**
	12	2,977,954	0.93	25.41	0.91	**24.88**	0.84	**23.64**
Triglycerides	11	116,648,917	1.77	11.62	1.75	**11.74**	1.08	**16.10**
Urate	1	145,630,111	0.44	14.44	0.44	**14.58**	0.40	**14.37**

Table 2: **Analysis of the heritability at loci with significant ME effects**. For each SNP t with significant ME, we report estimates of the ME heritability ($h^2_{gxg,t}$), the heritability of the SNP based on its GWAS effect ($h^2_{gwas,t}$), the standard error (SE), and the ratio between $h^2_{gxg,t}$ and $h^2_{gwas,t}$ (Ratio:= $h^2_{gxg,t}/h^2_{gwas,t}$). The ME effects were estimated after regressing out the additive effect within the LD region of the target SNPs.

Trait	CHR	SNP	$h^2_{gxg,t}$ $\times 0.001$	$SE(h^2_{gxg,t})$ $\times 0.001$	$h^2_{gwas,t}$ $\times 0.001$	$SE(h^2_{gwas,t})$ $\times 0.001$	Ratio
Alanine aminotransferase	22	44,388,817	10.003	1.352	1.168	0.080	8.562
Apolipoprotein B	11	116,648,917	7.457	1.218	1.451	0.111	5.140
C-reactive protein	1	66,257,838	8.703	1.389	1.157	0.079	7.520
Cholesterol	11	116,648,917	9.001	1.231	1.150	0.078	7.828
Hemoglobin A1c	8	41,542,093	5.078	0.737	1.397	0.104	3.636
Lipoprotein-A	6	160,560,845	13.781	1.949	0.314	0.011	43.888
	19	45,414,399	9.587	1.389	1.315	0.095	7.291
	6	160,578,069	7.029	0.668	5.330	0.778	1.319
Mean platelet volume	20	57,597,970	6.830	0.864	1.421	0.107	4.805
Monocyte count	13	28,623,048	3.851	0.628	2.771	0.292	1.390
SHBG	17	7,145,117	8.211	0.965	0.410	0.017	20.004
		7,254,315	7.064	1.126	1.006	0.064	7.021
Testosterone	7	99,032,593	7.915	1.176	0.184	0.005	42.981
	12	2,977,954	8.687	0.865	0.260	0.008	33.473
Triglycerides	11	116,648,917	9.083	1.242	15.326	3.795	0.593
Urate	1	145,630,111	3.166	0.495	0.720	0.039	4.395

References

[1] Heather J Cordell. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human molecular genetics*, 11(20):2463–2468, 2002.

[2] Patrick C Phillips. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, 9(11):855–867, 2008.

[3] Wen-Hua Wei, Gibran Hemani, and Chris S Haley. Detecting epistasis in human complex traits. *Nature Reviews Genetics*, 15(11):722–733, 2014.

[4] Evan E Eichler, Jonathan Flint, Greg Gibson, Augustine Kong, Suzanne M Leal, Jason H Moore, and Joseph H Nadeau. Missing heritability and strategies for finding the underlying causes of complex disease. *Nature reviews genetics*, 11(6):446–450, 2010.

[5] Pankhuri Singhal, Shefali Setia Verma, and Marylyn D Ritchie. Gene interactions in human disease studies—evidence is mounting. *Annual Review of Biomedical Data Science*, 6, 2023.

[6] William G Hill, Michael E Goddard, and Peter M Visscher. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet*, 4(2):e1000008, 2008.

[7] Roshni A Patel, Shaila A Musharoff, Jeffrey P Spence, Harold Pimentel, Catherine Tcheandjieu, Hakhamanesh Mostafavi, Nasa Sinnott-Armstrong, Shoa L Clarke, Courtney J Smith, VA Million Veteran Program, et al. Genetic interactions drive heterogeneity in causal variant effect sizes for gene expression and complex traits. *The American Journal of Human Genetics*, 109(7):1286–1297, 2022.

[8] Hakhamanesh Mostafavi, Arbel Harpak, Ipsita Agarwal, Dalton Conley, Jonathan K Pritchard, and Molly Przeworski. Variable prediction accuracy of polygenic scores within an ancestry group. *Elife*, 9:e48376, 2020.

[9] Alicia R Martin, Christopher R Gignoux, Raymond K Walters, Genevieve L Wojcik, Benjamin M Neale, Simon Gravel, Mark J Daly, Carlos D Bustamante, and Eimear E Kenny. Human demographic history impacts genetic risk prediction across diverse populations. *The American Journal of Human Genetics*, 100(4):635–649, 2017.

[10] Alicia R Martin, Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M Neale, and Mark J Daly. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature genetics*, 51(4):584–591, 2019.

[11] Tricia A Thornton-Wells, Jason H Moore, and Jonathan L Haines. Dissecting trait heterogeneity: a comparison of three clustering methods applied to genotypic data. *BMC bioinformatics*, 7:1–18, 2006.

[12] Örjan Carlborg and Chris S Haley. Epistasis: too often neglected in complex trait studies? *Nature Reviews Genetics*, 5(8):618–625, 2004.

[13] Yu Zhang and Jun S Liu. Bayesian inference of epistatic interactions in case-control studies. *Nature genetics*, 39(9):1167–1173, 2007.

[14] Yu Zhang, Bo Jiang, Jun Zhu, and Jun S Liu. Bayesian models for detecting epistatic interactions from genetic data. *Annals of human genetics*, 75(1):183–193, 2011.

[15] Wanwan Tang, Xuebing Wu, Rui Jiang, and Yanda Li. Epistatic module detection for case-control studies: a bayesian model with a gibbs sampling strategy. *PLoS genetics*, 5(5):e1000464, 2009.

[16] Xiang Wan, Can Yang, Qiang Yang, Hong Xue, Xiaodan Fan, Nelson LS Tang, and Weichuan Yu. Boost: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *The American Journal of Human Genetics*, 87(3):325–340, 2010.

[17] Attila Gyenesei, Jonathan Moody, Colin AM Semple, Chris S Haley, and Wen-Hua Wei. High-throughput analysis of epistasis in genome-wide association studies with biforce. *Bioinformatics*, 28(15):1957–1964, 2012.

[18] Yu Zhang. A novel bayesian graphical model for genome-wide multi-snp association mapping. *Genetic epidemiology*, 36(1):36–47, 2012.

[19] Snehit Prabhu and Itsik Pe'er. Ultrafast genome-wide scan for snp–snp interactions in common complex disease. *Genome research*, 22(11):2230–2240, 2012.

[20] Tony Kam-Thong, Benno Pütz, Nazanin Karbalai, Bertram Müller-Myhsok, and Karsten Borgwardt. Epistasis detection on quantitative phenotypes by exhaustive enumeration using gpus. *Bioinformatics*, 27(13):i214–i221, 2011.

[21] Attila Gyenesei, Jonathan Moody, Asta Laiho, Colin AM Semple, Chris S Haley, and Wen-Hua Wei. Biforce toolbox: powerful high-throughput computational analysis of gene–gene interactions in genome-wide association studies. *Nucleic acids research*, 40(W1):W628–W632, 2012.

[22] Yang Liu, Haiming Xu, Suchao Chen, Xianfeng Chen, Zhenguo Zhang, Zhihong Zhu, Xueying Qin, Landian Hu, Jun Zhu, Guo-Ping Zhao, et al. Genome-wide interaction-based association analysis identified multiple new susceptibility loci for common diseases. *PLoS genetics*, 7(3):e1001338, 2011.

[23] Thierry Schüpbach, Ioannis Xenarios, Sven Bergmann, and Karen Kapur. Fastepistasis: a high performance computing solution for quantitative trait epistasis. *Bioinformatics*, 26(11):1468–1469, 2010.

[24] Ling Sing Yung, Can Yang, Xiang Wan, and Weichuan Yu. Gboost: a gpu-based tool for detecting gene–gene interactions in genome–wide case control studies. *Bioinformatics*, 27(9):1309–1310, 2011.

[25] Gibran Hemani, Athanasios Theocharidis, Wenhua Wei, and Chris Haley. Epigpu: exhaustive pairwise epistasis scans parallelized on consumer level graphics cards. *Bioinformatics*, 27(11):1462–1465, 2011.

[26] Zhengkui Wang, Yue Wang, Kian-Lee Tan, Limsoon Wong, and Divyakant Agrawal. eceo: an efficient cloud epistasis computing model in genome-wide association study. *Bioinformatics*, 27(8):1045–1051, 2011.

[27] Jie Kate Hu, Xianlong Wang, and Pei Wang. Testing gene–gene interactions in genome wide association studies. *Genetic epidemiology*, 38(2):123–134, 2014.

[28] Lars Wienbrandt, Jan Christian Kässens, Jorge González-Domínguez, Bertil Schmidt, David Ellinghaus, and Manfred Schimmler. Fpga-based acceleration of detecting statistical epistasis in gwas. *Procedia Computer Science*, 29:220–230, 2014.

[29] Amy Strange, Francesca Capon, Chris CA Spencer, Jo Knight, Michael E Weale, Michael H Allen, Anne Barton, Gavin Band, Celine Bellenguez, Judith GM Bergboer, et al. A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between hla-c and erap1. *Nature Genetics*, 42(11):985–990, 2010.

[30] David M Evans, Chris CA Spencer, Jennifer J Pointon, Zhan Su, David Harvey, Grazyna Kochan, Udo Oppermann, Alexander Dilthey, Matti Pirinen, Millicent A Stone, et al. Interaction between erap1 and hla-b27 in ankylosing spondylitis implicates peptide handling in the mechanism for hla-b27 in disease susceptibility. *Nature genetics*, 43(8):761–767, 2011.

[31] Juan Pablo Lewinger, John L Morrison, Duncan C Thomas, Cassandra E Murcray, David V Conti, Dalin Li, and W James Gauderman. Efficient two-step testing of gene-gene interactions in genome-wide association studies. *Genetic epidemiology*, 37(5):440–451, 2013.

[32] Li Ma, Ariel Brautbar, Eric Boerwinkle, Charles F Sing, Andrew G Clark, and Alon Keinan. Knowledge-driven analysis identifies a gene–gene interaction affecting high-density lipoprotein cholesterol levels in multi-ethnic populations. *PLoS genetics*, 8(5):e1002714, 2012.

[33] Gary K Chen and Duncan C Thomas. Using biological knowledge to discover higher order interactions in genetic association studies. *Genetic epidemiology*, 34(8):863–878, 2010.

[34] Jean-Luc Jannink. Identifying quantitative trait locus by genetic background interactions in association studies. *Genetics*, 176(1):553–561, 2007.

[35] Danny S Park, Itamar Eskin, Eun Yong Kang, Eric R Gamazon, Celeste Eng, Christopher R Gignoux, Joshua M Galanter, Esteban Burchard, Chun J Ye, Hugues Aschard, et al. An ancestry-based approach for detecting interactions. *Genetic epidemiology*, 42(1):49–63, 2018.

[36] Lorin Crawford, Ping Zeng, Sayan Mukherjee, and Xiang Zhou. Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. *PLoS genetics*, 13(7):e1006869, 2017.

[37] Brooke Sheppard, Nadav Rappoport, Po-Ru Loh, Stephan J Sanders, Noah Zaitlen, and Andy Dahl. A model and test for coordinated polygenic epistasis in complex traits. *Proceedings of the National Academy of Sciences*, 118(15):e1922305118, 2021.

[38] Lorin Crawford and Xiang Zhou. Genome-wide marginal epistatic association mapping in case-control studies. *bioRxiv*, page 374983, 2018.

[39] W James Gauderman. Sample size requirements for association studies of gene-gene interaction. *American journal of epidemiology*, 155(5):478–484, 2002.

[40] Valentin Hivert, Julia Sidorenko, Florian Rohart, Michael E Goddard, Jian Yang, Naomi R Wray, Loic Yengo, and Peter M Visscher. Estimation of non-additive genetic variance in human complex traits from a large sample of unrelated individuals. *bioRxiv*, 2020.

[41] Yue Wu and Sriram Sankararaman. A scalable estimator of snp heritability for biobank-scale data. *Bioinformatics*, 34(13):i187–i194, 2018.

[42] Ali Pazokitoroudi, Yue Wu, Kathryn S. Burch, Kangcheng Hou, Aaron Zhou, B. Pasaniuc, and S. Sankararaman. Efficient variance components analysis across millions of genomes. *Nature Communications*, 11, 2020.

[43] Gibran Hemani, Konstantin Shakhbazov, Harm-Jan Westra, Tonu Esko, Anjali K Henders, Allan F McRae, Jian Yang, Greg Gibson, Nicholas G Martin, Andres Metspalu, et al. Detection and replication of epistasis influencing transcription in humans. *Nature*, 508(7495):249–253, 2014.

[44] Frank Dudbridge and Olivia Fletcher. Gene-environment dependence creates spurious gene-environment interaction. *The American Journal of Human Genetics*, 95(3):301–307, 2014.

[45] Andrew R Wood, Marcus A Tuke, Mike A Nalls, Dena G Hernandez, Stefania Bandinelli, Andrew B Singleton, David Melzer, Luigi Ferrucci, Timothy M Frayling, and Michael N Weedon. Another explanation for apparent epistasis. *Nature*, 514(7520):E3–E5, 2014.

[46] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

[47] Alkes L Price, Noah A Zaitlen, David Reich, and Nick Patterson. New approaches to population stratification in genome-wide association studies. *Nature reviews genetics*, 11(7):459–463, 2010.

[48] Yan Shu, Steven A Sheardown, Chaline Brown, Ryan P Owen, Shuzhong Zhang, Richard A Castro, Alexandra G Ianculescu, Lin Yue, Joan C Lo, Esteban G Burchard, et al. Effect of genetic variation in the organic cation transporter 1 (oct1) on metformin action. *The Journal of clinical investigation*, 117(5):1422–1431, 2007.

[49] Zoya Galcheva-Gargova, Konstantin N Konstantinov, I-Huan Wu, F George Klier, Tamera Barrett, and Roger J Davis. Binding of zinc finger protein zpr1 to the epidermal growth factor receptor. *Science*, 272(5269):1797–1802, 1996.

[50] Tanya M Teslovich, Kiran Musunuru, Albert V Smith, Andrew C Edmondson, Ioannis M Stylianou, Masahiro Koseki, James P Pirruccello, Samuli Ripatti, Daniel I Chasman, Cristen J Willer, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466(7307):707–713, 2010.

[51] Sekar Kathiresan, Cristen J Willer, Gina M Peloso, Serkalem Demissie, Kiran Musunuru, Eric E Schadt, Lee Kaplan, Derrick Bennett, Yun Li, Toshiko Tanaka, et al. Common variants at 30 loci contribute to polygenic dyslipidemia. *Nature genetics*, 41(1):56–65, 2009.

[52] Esteban J Parra, Andrew Mazurek, Christopher R Gignoux, Alexandra Sockell, Michael Agostino, Andrew P Morris, Lauren E Petty, Craig L Hanis, Nancy J Cox, Adan Valladares-Salgado, et al. Admixture mapping in two mexican samples identifies significant associations of locus ancestry with triglyceride levels in the bud13/znf259/apoa5 region and fine mapping points to rs964184 as the main driver of the association signal. *PLoS One*, 12(2):e0172880, 2017.

[53] Robert W Read, Karen A Schlauch, Vincent C Lombardi, Elizabeth T Cirulli, Nicole L Washington, James T Lu, and Joseph J Grzymski. Genome-wide identification of rare and common variants driving triglyceride levels in a nevada population. *Frontiers in Genetics*, 12:639418, 2021.

[54] Heribert Schunkert, Inke R König, Sekar Kathiresan, Muredach P Reilly, Themistocles L Assimes, Hilma Holm, Michael Preuss, Alexandre FR Stewart, Maja Barbalic, Christian Gieger, et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature genetics*, 43(4):333–338, 2011.

[55] Erika L Moen, Xu Zhang, Wenbo Mu, Shannon M Delaney, Claudia Wing, Jennifer McQuade, Jamie Myers, Lucy A Godley, M Eileen Dolan, and Wei Zhang. Genome-wide variation of cytosine modifications between european and african populations and the implications for complex traits. *Genetics*, 194(4):987–996, 2013.

[56] L Pfeiffer, S Wahl, LC Pilling, E Reischl, JK Sandling, S Kunze, et al. Dna methylation of lipid-related genes affects blood lipid levels. *Circ Cardiovasc Genet*, 8(2):334–42, 2015.

[57] Chen Yao, Brian H Chen, Roby Joehanes, Burcak Otlu, Xiaoling Zhang, Chunyu Liu, Tianxiao Huan, Oznur Tastan, L Adrienne Cupples, James B Meigs, et al. Integromic analysis of genetic variation and gene expression identifies networks for cardiovascular disease phenotypes. *Circulation*, 131(6):536–549, 2015.

[58] Rui Xiao and Michael Boehnke. Quantifying and correcting for the winner's curse in genetic association studies. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 33(5):453–462, 2009.

[59] MF Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 18(3):1059–1076, 1989.

[60] Edo Liberty and Steven W Zucker. The mailman algorithm: A note on matrix–vector multiplication. *Information Processing Letters*, 109(3):179–182, 2009.

[61] Bycroft C et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562:203–209, 2018.

[62] Nasa Sinnott-Armstrong, Yosuke Tanigawa, David Amar, Nina Mars, Christian Benner, Matthew Aguirre, Guhan Ram Venkataraman, Michael Wainberg, Hanna M Ollila, Tuomo Kiiskinen, et al. Genetics of 35 blood and urine biomarkers in the uk biobank. *Nature genetics*, 53(2):185–194, 2021.

[63] Xinzhu Wei, Christopher R Robles, Ali Pazokitoroudi, Andrea Ganna, Alexander Gusev, Arun Durvasula, Steven Gazal, Po-Ru Loh, David Reich, and Sriram Sankararaman. The lingering effects of neanderthal introgression on human complex traits. *eLife*, 12:e80757, mar 2023.

[64] Doug Speed, Gibran Hemani, Michael R Johnson, and David J Balding. Improved heritability estimation from genome-wide snps. *The American Journal of Human Genetics*, 91(6):1011–1021, 2012.

[65] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, 81(3):559–575, 2007.