Partitioning Tagged Non-Additive Genetic Effects in Summary Statistics Provides Evidence of Pervasive Epistasis in Complex Traits

Gregory Darnell[1,2,*], Samuel Pattillo Smith[1,3,*], Dana Udwin[4], Sohini Ramachandran[1,3,5,§], and Lorin Crawford[1,4,6,§,†]

1 Center for Computational Molecular Biology, Brown University, Providence, RI, USA

2 Institute for Computational and Experimental Research in Mathematics, Brown University, Providence, RI, USA

3 Department of Ecology and Evolutionary Biology, Brown University, Providence, RI, USA

4 Department of Biostatistics, Brown University, Providence, RI, USA

5 Data Science Initiative, Brown University, Providence, RI, USA

6 Microsoft Research New England, Cambridge, MA, USA

*** Authors Contributed Equally**

§ **Authors Contributed Equally**

† **Corresponding E-mail: lcrawford@microsoft.com**

Abstract

The inflation of test statistics in genome-wide association (GWA) studies due to confounding factors such as cryptic relatedness, population stratification, and spurious non-zero genetic effects driven by linkage disequilibrium (LD) has been well characterized in the literature. The key theoretical contribution of this work is that epistasis (i.e., the interaction between multiple loci and/or genes) can also lead to misestimated GWA summary statistics. To address this challenge, we develop marginal epistatic LD score regression and the accompanying software package `MELD`: an extended framework which takes in GWA test statistics and accurately partitions true additive genetic variation from non-additive genetic variation, as well as other biases. By re-analyzing 25 well-studied quantitative phenotypes from 349,468 individuals of European ancestry in the UK Biobank and up to 159,095 individuals in BioBank Japan, we

illustrate that nonlinear effects are a significant source of signal in reported GWA summary statistics and provide evidence that epistasis is more widespread in human phenotypes than previously reported. Of the 25 complex traits we analyzed in the UK Biobank, 23 phenotypes have a significant amount of tagged epistasis captured within additive summary statistics, including height, urate level, and cholesterol levels. The `MELD` software and its application to these biobanks represent a significant step towards resolving the true contribution of epistasis to human complex traits.

Introduction

Understanding the genetic contribution to trait variation, or heritability, has been a central line of inquiry for over a century in a range of species, including our own[1,2]. Until recently, studies of genetic heritability in humans have been reliant on typically small sized family studies with known relatedness structure between individuals[3,4]. Due to advances in genomic sequencing and the steady development of novel statistical tools, it is now possible to obtain reliable heritability estimates from biobank-scale datasets of unrelated individuals[5–8]. Accurate estimation of heritability in these larger cohorts is crucial for gaining insight into the biological underpinnings of complex trait variation.

Narrow-sense heritability (denoted h^2) is defined as the true contribution of additive genetic effects in the generative model phenotypic trait variation[5,6,9]. Due to computational and privacy considerations with biobank-scale genome-wide association (GWA) studies, a recent trend has been to estimate narrow-sense heritability using GWA summary statistics (i.e., effect sizes and standard errors estimated from the GWA linear model). In the GWA linear model, additive effect sizes and standard errors for individual single nucleotide polymorphisms (SNPs) are estimated by regressing phenotype measurements onto the allele counts of each locus independently. It has become clear that many traits have a complex and polygenic basis—that is, hundreds to thousands of individual genetic loci across the genome often contribute to the variation of a single trait[10]. However, broad-sense heritability (H^2), which includes all genetic factors that contribute to trait variation, including non-additive factors such as dominance or epistatic effects, has not been a focus in these traditional studies.

Recent statistical methods have been developed to better distinguish true polygenic genetic architecture from confounding factors, such as cryptic relatedness and population stratification, when estimating narrow-sense heritability from genetic variants[5,6,11,12]. The most widely used of these approaches is

linkage disequilibrium (LD) score regression and the corresponding `LDSC` software[5], which corrects for inflation in GWA summary statistics by modeling the relationship between the variance of SNP-level effect sizes and the sum of correlation coefficients between focal SNPs and their genomic neighbors (i.e., the LD score of each variant). The main motivation behind the `LDSC` model is that, for polygenic traits, non-associated (or "null") SNPs have a higher probability to emit spurious nonzero effects. This can be simply because they are in some degree of LD with (at least) one-of-many causal variants[5] or because they have a trans-interaction effect with variants located within a gene enriched for associations with the trait of interest[13]. The goal of `LDSC` is to partition the bias in summary statistics due to this confounding and thereby provide a more precise estimate of narrow-sense heritability. As of late, there have been many efforts to build upon and improve the `LDSC` framework. For example, one limitation of the `LDSC` model is that, in practice, it only uses the diagonal elements of the squared LD matrix in its formulation. This tradeoff helps the method to scale genome-wide, but it also has been shown to lead to large standard errors for heritability estimates[12,14,15]. As a result, newer approaches have attempted to reformulate the `LDSC` model by using the eigenvalues of the LD matrix to leverage more of the information present in the correlation structure between SNPs[6,12].

While the `LDSC` model and its current extensions have improved accuracy for narrow-sense heritability estimation, none consider the need to correct possible misestimation in additive GWA summary statistics stemming from tagged nonlinear genetic effects. This is in part due to the longstanding and ongoing debate about the contribution of non-additive effects (e.g., epistasis and dominance effects) on the architecture of human complex traits[16–26]. However, despite these controversies, many association mapping studies in humans have identified candidates of epistasis that notably contribute to trait variation[27–30], and some have recently shown that gene-by-gene interactions can drive heterogeneity of causal variant effect sizes across diverse human populations[31]. Epistasis is a well-known contributor to trait architecture in several model organisms[32–43]. Importantly, non-additive genetic variation has been proposed as one of the main factors that explains missing heritability—the proportion of heritability not explained by the top associated variants in GWA studies[44]. Lastly, and particularly relevant to this work, studies have hypothesized that nonlinear genetic effects can confound heritability estimation in pedigree studies and cause misestimation of heritability statistics, creating so-called "phantom heritability" [22,45,46].

The key theoretical insight we highlight in this manuscript is that, in addition to polygenicity and other biases, SNP-level GWA summary statistics can provide evidence of epistasis if there is a nonzero

correlation between individual-level genotypes and their nonlinear genetic interactions in the generative model of complex traits. Here, we limit our demonstration to second-order (or pairwise) epistasis but this general concept can easily be extended to other sources of nonlinear genetic variation (e.g., dominance). To that end, we present the "marginal epistatic LD score" regression model or `MELD`: a simple extension of the `LDSC` framework which takes SNP-level effect sizes as input and aims to uniquely partition true additive genetic variation from non-additive genetic variation and other uncontrolled confounding factors. The main difference between `MELD` and `LDSC` is that the `MELD` model includes an additional set of "marginal epistatic" LD scores in its regression. These scores measure the amount of higher-order genetic variation that is tagged by each SNP in the GWA dataset. In practice, these additional scores are computationally efficient to compute and require nothing more than access to an ancestry-matched set of samples if genotype data are not available to the user, equivalent to the necessary data for performing LD score regression.

Through extensive simulations, we show that `MELD` improves upon the estimation of narrow-sense heritability when genetic interactions are indeed present in the generative model for complex traits. More importantly, `MELD` has a calibrated type I error rate and does not overestimate non-additive genetic contribution to trait variation in simulated data when only additive effects are present. In real data analyses of 25 complex, continuous traits in the UK Biobank and BioBank Japan, we illustrate that pairwise interactions are a significant source of bias in reported additive GWA summary statistics—suggesting that epistasis is more pervasive in human phenotypes than previously reported. We believe that `MELD` represents a significant step towards resolving the true contribution of epistasis to human complex traits.

Results

Overview of marginal epistatic LD score regression

Marginal epistatic LD score regression is a statistical framework which seeks to accurately partition true additive genetic effects from both tagged non-additive genetic variation and confounding factors such as polygenicity, cryptic relatedness, and population stratification. As an overview of the method and our corresponding software `MELD`, we will assume that we are analyzing a GWA dataset $\mathcal{D} = \{\mathbf{X}, \mathbf{y}\}$ where $\mathbf{X}$ is an $N \times J$ matrix of genotypes with J denoting the number of SNPs (each of which is encoded as

$\{0, 1, 2\}$ copies of a reference allele at each locus j) and $\mathbf{y}$ is an N-dimensional vector of measurements of a quantitative trait. `MELD` only requires summary statistics of individual-level data: namely, marginal effect size estimates for each SNP $\widehat{\boldsymbol{\beta}}$ and an empirical LD matrix $\mathbf{R}$ (which can be provided via reference panel data). In this study, we focus on pairwise statistical epistasis but this framework can easily be adapted to distinguish higher-order nonlinear interactions as well.

We begin by assuming the following generative linear model for complex traits

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, (1 - H^2)\mathbf{I}), \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)$ is a J-dimensional vector containing the true additive effect sizes for an additional copy of the reference allele at each locus on $\mathbf{y}$; $\mathbf{W}$ is an $N \times M$ matrix of (pairwise) epistatic interactions between some subset of causal SNPs, where columns of this matrix are assumed to be the Hadamard (element-wise) product between genotypic vectors of the form $\mathbf{x}_j \circ \mathbf{x}_k$ for the j-th and k-th variants; $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_M)$ is an M-dimensional vector containing the interaction effect sizes; $\boldsymbol{\varepsilon}$ is a normally distributed error term with mean zero and variance scaled according to the proportion of phenotypic variance not explained by the broad-sense heritability of the trait [47], where the broad-sense heritability of the trait is denoted by H^2. $\mathbf{I}$ denotes an $N \times N$ identity matrix. For convenience, we will assume that the genotype matrix (column-wise) and the trait of interest have been mean-centered and standardized. Lastly, we let each individual effect size follow a normal distribution with variances proportional to their individual contributions to the broad-sense heritability of the trait of interest [47–51]

$$\beta_j \sim \mathcal{N}(0, H^2\rho/J), \qquad \theta_m \sim \mathcal{N}(0, H^2(1 - \rho)/M) \tag{2}$$

where ρ measures the proportion of total genetic effects that is contributed by the additive effects. Effectively, we say $\mathbb{V}[\mathbf{X}\boldsymbol{\beta}] = H^2\rho = h^2$ is the narrow-sense heritability for a trait, while $\mathbb{V}[\mathbf{W}\boldsymbol{\theta}] = H^2(1-\rho)$ makes up the remaining proportion of the broad-sense heritability.

A central goal in GWA studies is to infer the true additive effects for each SNP. This is usually done by assuming two conditions: (i) non-additive genetic effects play a negligible role on the overall architecture of complex traits [24,25], and (ii) that the genotype and interaction matrices $\mathbf{X}$ and $\mathbf{W}$ do not share the same column space (i.e., such that $\mathbf{X}^\intercal\mathbf{W} = \mathbf{0}$). However, if we relax these assumptions, then the following relationship between the moment matrix $\mathbf{X}^\intercal\mathbf{y}$, the observed marginal GWA summary statistics $\widehat{\boldsymbol{\beta}}$, and

the true coefficient values $\boldsymbol{\beta}$ holds in expectation (see Materials and Methods)

$$\mathbf{X}^\intercal\mathbf{y} = (\mathbf{X}^\intercal\mathbf{X})\boldsymbol{\beta} + (\mathbf{X}^\intercal\mathbf{W})\boldsymbol{\theta} \qquad \overset{\approx}{\Longleftrightarrow} \qquad \widehat{\boldsymbol{\beta}} = \mathbf{R}\boldsymbol{\beta} + \mathbf{V}\boldsymbol{\theta} \tag{3}$$

where $\mathbf{R}$ is an empirical estimate of the LD matrix and $\mathbf{V}$ represents an empirical estimate of the correlation between the individual-level genotypes $\mathbf{X}$ and the span of genetic interactions between causal SNPs in $\mathbf{W}$. Intuitively, the term $\mathbf{V}\boldsymbol{\theta}$ can be interpreted as "bias" in the additive effect estimate that stem from tagged interaction effects. Here, we use "bias" in the statistical sense to mean any systematic difference between the expected value of an estimator and true value of the parameter being estimated (i.e., $\mathbb{E}[\widehat{\boldsymbol{\beta}}] - \boldsymbol{\beta} \neq \mathbf{0}$). Note that when either conditions (i) or (ii) are indeed met such that $\mathbf{V}\boldsymbol{\theta} = \mathbf{0}$, the equation above simplifies to a relationship between LD and summary statistics that is assumed in many common GWA studies[13,52–57].

Recall that the goal of `MELD` is to identify the proportion of bias that stems from epistatic effects within additive GWA summary statistics. To do this, we build upon the LD score regression framework and the corresponding `LDSC` software[47]. Here, we note that, according to Eq. (3), $\widehat{\boldsymbol{\beta}} \sim \mathcal{N}(\mathbf{R}\boldsymbol{\beta} + \mathbf{V}\boldsymbol{\theta}, \lambda\mathbf{R})$ where λ is a misestimation factor (i.e., inflation or deflation) due to uncontrolled confounding effects[12,58]. Next, we condition on $\boldsymbol{\Theta} = (\boldsymbol{\beta}, \boldsymbol{\theta})$ and take the expectation of chi-square statistics $\boldsymbol{\chi}^2 = N\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^\intercal$ to yield

$$\begin{aligned}
\mathbb{E}[\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^\intercal] = \mathbb{E}\left[\mathbb{E}\left[\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^\intercal \,|\, \boldsymbol{\Theta}\right]\right] &= \mathbb{E}\left[\mathbb{V}\left[\widehat{\boldsymbol{\beta}} \,|\, \boldsymbol{\Theta}\right] + \mathbb{E}\left[\widehat{\boldsymbol{\beta}} \,|\, \boldsymbol{\Theta}\right]\mathbb{E}\left[\widehat{\boldsymbol{\beta}} \,|\, \boldsymbol{\Theta}\right]^\intercal\right] \\
&= \mathbb{E}\left[\lambda\mathbf{R} + (\mathbf{R}\boldsymbol{\beta} + \mathbf{V}\boldsymbol{\theta})(\mathbf{R}\boldsymbol{\beta} + \mathbf{V}\boldsymbol{\theta})^\intercal\right] \\
&= \mathbb{E}\left[\lambda\mathbf{R} + \mathbf{R}\boldsymbol{\beta}\boldsymbol{\beta}^\intercal\mathbf{R} + 2\mathbf{R}\boldsymbol{\beta}\boldsymbol{\theta}^\intercal\mathbf{V} + \mathbf{V}\boldsymbol{\theta}\boldsymbol{\theta}\mathbf{V}^\intercal\right] \\
&= \lambda\mathbf{R} + \left(\frac{H^2\rho}{J}\right)\mathbf{R}^2 + \left(\frac{H^2(1-\rho)}{M}\right)\mathbf{V}^2.
\end{aligned} \tag{4}$$

We define $\ell_j = \sum_k r_{jk}^2$ as the LD score for the additive effect of the j-th variant[47], and $f_j = \sum_m v_{jm}^2$ represents the "marginal epistatic" LD score which encodes the interaction between the j-th variant and all other variants in the data set[51], respectively. By considering only the diagonal elements of LD matrix in the first term, similar the original `LDSC` approach[12,47], we get the following simplified regression

$$\mathbb{E}[\boldsymbol{\chi}^2] \propto \mathbf{1} + \boldsymbol{\ell}\tau + \boldsymbol{f}\sigma \tag{5}$$

where $\boldsymbol{\chi}^2 = (\chi_1^2, \ldots, \chi_J^2)$ is a J-dimensional vector of chi-square summary statistics, and $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_J)$ and $\boldsymbol{f} = (f_1, \ldots, f_J)$ are J-dimensional vectors of additive and marginal epistatic LD scores, respectively. Furthermore, we define the variance components $\tau = NH^2\rho/J$ and $\sigma = NH^2(1-\rho)/M$ as the additive and epistatic regression coefficients of the model, and $\mathbf{1}$ is the intercept meant to model the misestimation factor due to uncontrolled confounding effects (e.g., cryptic relatedness structure). In practice, we efficiently compute the marginal epistatic LD scores by considering only a subset of interactions between each j-th focal SNP and SNPs within a *cis*-proximal window around the j-th SNP. This is based on the observation that LD decays outside of a window of 1 centimorgan (cM); therefore, SNPs outside the 1cM window centered on the j-th SNP will not significantly contribute to its LD scores. The `MELD` software package combines weighted least squares with a model averaging strategy (over different genomic window values) to estimate regression parameters. It then derives P-values for identifying summary statistics with significant bias stemming from epistatic signal by testing the null hypothesis $H_0 : \sigma = 0$. Importantly, under the null of a trait being generated by only additive effects, the `MELD` model in Eq. (5) is equivalent to the original `LDSC` framework.

Lastly, we want to note the empirical observation that the additive ($\boldsymbol{\ell}$) and marginal epistatic ($\boldsymbol{f}$) LD scores are lowly correlated. This is important because that means that the presence of marginal epistatic LD scores in the model specified in Eq. (5) has little-to-no influence over the estimate for the additive coefficient τ. Instead, the inclusion of $\boldsymbol{f}$ re-partitions the proportion of summary statistics biased by non-additive genetic variation (which would usually be included in the intercept) and places it within σ. In other words, we can interpret σ as the misestimation factor due to tagged epistasis. As a result, we use the difference between coefficient estimates $\tau - \sigma$ to construct unbiased estimates of narrow-sense heritability. A full theoretical derivation of the marginal epistatic LD regression framework and details about its corresponding implementation in our software `MELD` can be found in Materials and Methods.

Detection of significant tagged epistasis using `MELD` in simulations

We test the utility of `MELD` across different genetic trait architectures via an extensive simulation study (Materials and Methods). Here, we generate synthetic phenotypes using real genome-wide genotype data from individuals of self-identified European ancestry in the UK Biobank. To do so, we first assume that traits have a polygenic architecture where all SNPs have a non-zero additive effect. Next, we randomly select a set of causal epistatic variants and divide them into two interacting groups (Materials and

Methods). One may interpret the SNPs in group #1 as being the "hubs" in an interaction map [51]; while, SNPs in group #2 are selected to be variants within some kilobase (kb) window around each SNP in group #1. We assume a wide-range of simulation scenarios by varying the following parameters:

- broad-sense heritability: $H^2 = 0.3$ and 0.6;
- proportion of phenotypic variation that is explained by additive effects: $\rho = 0.5$, 0.8, and 1;
- percentage of SNPs selected to be in group #1: 1%, 5%, and 10%;
- genomic window used to assign SNPs to group #2: $\pm$10 and $\pm$100 kb.

We also varied the correlation between SNP effect size and minor allele frequency (MAF) (as discussed in Schoech et al. [59]). All results presented in this section are based on 100 different simulated phenotypes for each parameter combination.

Overall, results show that `MELD` robustly detects significant tagged epistatic effects, regardless of the total number of causal interactions genome-wide (Figure 1). Instead, the power of `MELD` depends on the proportion of phenotypic variation that is explained by additive versus non-additive effects, and its power tends to scale with the window size used to compute the marginal epistatic LD scores (again see Materials and Methods). `MELD` shows similar ability to detect tagged epistatic effects even in the presence of MAF-dependent effect sizes and when we vary the number of SNPs assigned to be in group #2 (Figures S1-S5).

Importantly, `MELD` does not falsely identify putative epistatic effects in GWA summary statistics when the synthetic phenotype they were derived from was generated only by additive effects. Figure 2 illustrates the performance of `MELD` under the null hypothesis, with the type I error rates for different estimation window sizes of the marginal epistatic LD scores highlighted in panel A. Here, we also show that, when no epistasis is present, `MELD` unbiasedly estimates the epistatic coefficient in the regression model $\sigma = 0$ (Figure 2B), robustly estimates the narrow-sense heritability of traits correctly (Figure 2C), and provides well-calibrated P-values when assessed over many traits (Figure 2D). This behavior is consistent across different MAF-dependent effect size distributions, and `MELD` is not sensitive to misspecification of the estimation windows used to generate the marginal epistatic LD scores (Figures S6-S7).

Lastly, one of the most important innovations that `MELD` offers over the traditional `LDSC` framework is the correction of narrow-sense heritability estimates after detecting bias from non-additive genetic

variation. Here, we applied both methods to the same set of simulations in order to understand how `LDSC` behaves for traits that were generated with epistatic effects. Figures 3 and S8 depict boxplots of the narrow-sense heritability estimates for each approach and shows that, across an array of different synthetic phenotype architectures, `LDSC` routinely overestimates the truth in our simulations that include nonzero epistatic effects. In contrast, `MELD` more accurately partitions the total genetic variance explained, which in turn leads to more precise estimation. The mean absolute error between the true h^2 value and the estimates produced by `MELD` and `LDSC` are shown in Table S1 and S2, respectively. Generally, the error in narrow-sense heritability estimates is higher for `LDSC` than it is for `MELD` across each of the scenarios that we consider.

Application of `MELD` to the UK Biobank and BioBank Japan

To assess whether nonlinear genetic interactions are significantly biasing GWA summary statistics in empirical biobank data, we applied `MELD` to 25 continuous quantitative traits from the UK Biobank and BioBank Japan (Table S3). Protocols for computing GWA summary statistics for the UK Biobank are described in the Materials and Methods; while pre-computed summary statistics for BioBank Japan were downloaded directly from the consortium website (see URLs). We release marginal epistatic LD scores on the `MELD` GitHub page from two reference populations in the 1000 Genomes: 489 individuals from the European superpopulation (EUR) and 504 individuals from the East Asian (EAS) superpopulation (see also Table S4).

In 23 of the 25 traits we analyzed in the UK Biobank, we detected significant bias stemming from pairwise epistasis (Table 1). This includes many canonical traits of interest in heritability analyses: height, cholesterol levels, urate levels, and both systolic and diastolic blood pressure. Our findings in Table 1 are supported by multiple published studies identifying epistasis in a given trait of interest. For example, Li et al.[60] found statistical evidence for epistatic interactions that contributed to the pathogenesis of coronary artery disease. It was also recently shown that non-additive variation plays a significant role in body mass index[12]. Generally, we find that the traditional `LDSC` underestimates trait narrow-sense heritability when it does not consider this additional source of genetic signal as opposed to `MELD` (Table S6). In BioBank Japan, the only trait with a significant nonlinear component was triglyceride levels. We believe that this, in part, may be due to the discrepancy in sample sizes between the UK Biobank ($N = 349,469$ for all traits) and BioBank Japan (Table S5).

For each of the 25 traits that we analyzed, we found that the MELD narrow-sense heritability estimates are generally correlated with that of the LDSC in both the UK Biobank ($r^2 = 0.591$, $P = 1.13 \times 10^{-5}$) and BioBank Japan ($r^2 = 0.815$, $P = 6.95 \times 10^{-10}$). Additionally, we found that the narrow-sense heritability estimates for the same traits between the two biobanks are highly correlated according to both LDSC ($r^2 = 0.664$, $P = 1.26 \times 10^{-6}$) and MELD ($r^2 = 0.734$, $P = 4.69 \times 10^{-8}$) analysis. These results are shown in Figure 4A and B, respectively.

After comparing the MELD narrow-sense heritability estimates to LDSC, we then assessed whether there was significant difference in the amount of bias in the GWA summary statistics derived from the the UK Biobank and BioBank Japan (i.e., comparing the estimates of σ; see Figure 4C). We show that, while heterogeneous between traits, the bias introduced by nonlinear interactions is relatively of the same magnitude for both biobanks ($r^2 = 0.239$, $P = 0.013$). Notably, the trait with the most significant evidence of epistatic bias in GWA summary statistics is height which is known to have a highly polygenic architecture. Across the 25 traits studied, the estimated additive coefficients between UK Biobank and BioBank Japan are also highly correlated ($r^2 = 0.748$, $P = 2.49 \times 10^{-10}$).

Finally, we show that the intercepts estimated by LDSC and MELD are highly correlated in both the UK Biobank and the BioBank Japan. Recall that these intercept estimates represent the confounding factor due to uncontrolled effects. For LDSC this does include bias from pairwise genetic interactions, while MELD intercept estimates do not include bias due to these types of nonlinear effects. The MELD intercept estimates tend to be correlated but generally different than those computed with LDSC — empirically indicating that non-additive genetic variation is partitioned away from other types of biases when marginal epistatic scores included in the LD score framework (Figure S9). This result shows similar patterns of bias both the UK Biobank and BioBank Japan, and it confirms that nonlinear effects can be a source of bias in heritability estimation.

Discussion

In this paper, we present MELD, an extension of the LD score regression framework that partitions true additive genetic variation from biases introduced by non-additive genetic effects using GWA summary statistics. The key insight underlying MELD is that SNP-level GWA summary statistics can be biased if there is a nonzero correlation between individual-level genotypes and their nonlinear genetic interactions;

this is in addition to other biases well-known to affect GWA results such as polygenic trait architecture. `MELD` builds upon the original `LDSC` model through the inclusion of "marginal epistatic" LD scores which capture sources of epistasis that are tagged by each SNP in the data (Figures 1 and S1-S5). Through extensive simulations, we show that `MELD` is well-calibrated under the null model when traits are generated only by additive effects (Figures 2 and S6-S7), and it provides improved narrow-sense heritability estimates over `LDSC` when traits are generated with interaction effects (Figures 3 and S8, and Tables S1 and S2). Lastly, in real data, we show examples of many traits with estimated GWA summary statistics that are biased by epistatic effects in the UK Biobank and BioBank Japan (Figures 4 and S9, and Tables 1 and S6). We have made `MELD` a publicly available command line tool that requires minimal updates to the environment used to run the original implementation of LD score regression. In addition, we provide pre-computed marginal epistatic LD scores calculated from the European (EUR) and East Asian (EAS) reference populations in the 1000 Genomes phase 3 data (see Data and Software Availability under Materials and Methods).

The current implementation of the `MELD` framework offers many directions for future development and applications. First, we note that in this study we did not incorporate additional variant annotations (e.g., based on epigenetic information, regulatory genomic units) during our computation of LD scores[61–63]. The inclusion of additional annotations has been shown to provide more refined narrow-sense heritability estimates from GWA summary statistics while accounting for linkage[64]. A key part of our future work is to explore whether considering annotation groups would also improve our ability to identify tagged epistasis. Second, in its current form, the `MELD` software only considers non-additive genetic variation and ignores unobserved environmental or population-specific covariates that could also cause biases in GWA summary statistics. In the future, we plan to expand the `MELD` framework to also study confounding stemming from factors such as gene-by-environment (G×E) or gene-by-sex (G×Sex) interactions. We can do this by computing a new set of scores which encode how loci interact with one or more environmental instruments[65–67]. Lastly, we have only focused on analyzing one phenotype at a time in this study. However, many previous studies have extensively shown that modeling multiple phenotypes can often dramatically increase power[68]. Therefore, it would be interesting to extend the `MELD` framework to multiple traits to study nonlinear genetic correlations in the same way that `LDSC` was recently extended to uncover additive genetic correlation maps across traits[69].

URLs

`MELD` software package for implementing marginal epistatic LD score regression, `https://github.com/lcrawlab/MELD`; `LDSC` software package for implementing LD score regression, `https://github.com/bulik/ldsc/`; UK Biobank, `https://www.ukbiobank.ac.uk`; BioBank Japan, `http://jenger.riken.jp/en/result`; 1000 Genomes Project genetic map and haplotypes, `http://mathgen.stats.ox.ac.uk/impute/data_download_1000G_phase1_integrated.html`; Database of Genotypes and Phenotypes (dbGaP), `https://www.ncbi.nlm.nih.gov/gap`; NHGRI-EBI GWAS Catalog, `https://www.ebi.ac.uk/gwas/`; GRM-MAF-LD package, `https://github.com/arminschoech/GRM-MAF-LD`; GCTA toolkit, `https://yanglab.westlake.edu.cn/software/gcta/`.

Acknowledgements

We thank Jeffrey P. Spence, Roshni Patel, Matthew Aguirre, and Mineto Ota for insightful comments on an earlier version of this manuscript as well as the Ramachandran and Crawford Labs for helpful discussions. This research was conducted in part using computational resources and services at the Center for Computation and Visualization at Brown University. This research was also conducted using the UK Biobank Resource under Application Numbers 14649 (LC) and 22419 (SR). This research was supported in part by an Alfred P. Sloan Research Fellowship, and a David & Lucile Packard Fellowship for Science and Engineering awarded to L. Crawford. This research was also partly supported by US National Institutes of Health (NIH) grant R01 GM118652, NIH grant R35 GM139628, and National Science Foundation (NSF) CAREER award DBI-1452622 to S. Ramachandran. G. Darnell was supported by NSF Grant No. DMS-1439786 while in residence at the Institute for Computational and Experimental Research in Mathematics (ICERM) in Providence, RI. S.P. Smith and D. Udwin were trainees supported under the Brown University Predoctoral Training Program in Biological Data Science (NIH T32 GM128596). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of any of the funders.

Author Contributions

GD, SR, and LC conceived the study and developed the methods. GD, SPS, and LC developed the algorithms and software. GD, SPS, and DU performed the analyses. All authors wrote and revised the manuscript.

Competing Interests

The authors declare no competing interests.

Materials and Methods

Generative statistical model for complex traits

Our goal in this study is to re-analyze summary statistics from genome-wide association (GWA) studies and distinguish true additive genetic associations from bias stemming from tagged epistatic interactions. We begin by assuming the following generative linear model for complex traits and phenotypes

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, (1 - H^2)\mathbf{I}), \tag{6}$$

where $\mathbf{y}$ denotes an N-dimensional vector of phenotypic states for a quantitative trait of interest measured in N individuals; $\mathbf{X}$ is an $N \times J$ matrix of genotypes, with J denoting the number of single nucleotide polymorphism (SNPs) encoded as $\{0, 1, 2\}$ copies of a reference allele at each locus; $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)$ is a J-dimensional vector containing the true additive effect sizes for an additional copy of the reference allele at each locus on $\mathbf{y}$; $\mathbf{W}$ is an $N \times M$ matrix of (pairwise) epistatic interactions between some subset of causal SNPs, where columns of this matrix are assumed to be the Hadamard (element-wise) product between genotypic vectors of the form $\mathbf{x}_j \circ \mathbf{x}_k$ for the j-th and k-th variants; $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_M)$ is an M-dimensional vector containing the interaction effect sizes; $\boldsymbol{\varepsilon}$ is a normally distributed error term with mean zero and variance scaled according to the proportion of phenotypic variance not explained by the broad-sense heritability of the trait, denoted by H^2; and $\mathbf{I}$ denotes an $N \times N$ identity matrix.

For convenience, we further assume that the genotype matrix (column-wise) and trait of interest have been mean-centered and standardized. Furthermore, we want to point out that the generative formulation of Eq. (6) can also be easily extended to accommodate other fixed effects (e.g., age, sex, or genotype principal components), as well as other random effects terms that can be used to account for sample non-independence due to other environmental factors. In addition, we choose to assume that $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are fixed effects here, but modeling these coefficients as a random effect is straightforward. Lastly, in this work, we only consider second order (or pairwise) epistatic relationships between SNPs. However, the generalization of the proposed framework to detect bias from higher-order interactions is also straightforward and only involves manipulating the epistatic matrix $\mathbf{W}$ [51,70].

GWA summary statistics and tagged epistatic effects

As previously mentioned, the key theoretical insight of this work is that, in addition to polygenicity and other sources of confounding such as cryptic relatedness and population stratification, SNP-level GWA summary statistics can also be biased if there is a nonzero correlation between individual-level genotypes and their interactions (as defined in Eq. (6)). Here, we use the term "bias" in the statistical sense to mean any systematic difference between the expected value of an estimator and true value of the parameter being estimated (i.e., $\mathbb{E}[\widehat{\boldsymbol{\beta}}] - \boldsymbol{\beta} \neq \mathbf{0}$). We now formally derive this concept. Throughout this section, we will use $\mathbf{X}^\intercal\mathbf{X}/N$ to denote the linkage disequilibrium (LD) or pairwise correlation matrix between SNPs. We will then let $\mathbf{R}$ represent an LD matrix empirically estimated from external data (e.g., directly from GWA study data, or using an LD map from a population with similar genomic ancestry to that of the samples analyzed in the GWA study). The important property here is that

$$\mathbb{E}[\mathbf{X}^\intercal\mathbf{X}] \approx N\mathbf{R}, \qquad \mathbb{E}[\mathbf{x}_j^\intercal\mathbf{x}_j] \approx N, \qquad \mathbb{E}[\mathbf{x}_j^\intercal\mathbf{x}_k] \approx Nr_{jk} \tag{7}$$

where the term r_{jk} is defined as the Pearson correlation coefficient between the j-th and k-th SNPs, respectively, and $\mathbf{x}_j$ denotes the j-th column of the individual-level genotype matrix $\mathbf{X}$.

A central goal in GWA studies is to jointly infer the true additive effects $\boldsymbol{\beta} = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}$ for each SNP, given both genotypic and phenotypic measurements for each assayed individual. However, since the generative model in Eq. (6) is an underdetermined linear system (i.e., $J > N$) for many GWA applications, we need to make additional modeling assumptions on the regression coefficients to make the generative model identifiable. To do so, we follow standard linear modeling approaches[47–51] and assume that each individual effect size follows a normal distribution with variances proportional to their individual contributions to the broad-sense heritability of the trait of interest. Namely, we assume that

$$\beta_j \sim \mathcal{N}(0, H^2\rho/J), \qquad \theta_m \sim \mathcal{N}(0, H^2(1-\rho)/M), \qquad j = 1, \ldots, J \qquad m = 1, \ldots, M \tag{8}$$

where ρ measures the proportion of total genetic effects that is contributed by the additive effects. Alternatively, we say that $\mathbb{V}[\mathbf{X}\boldsymbol{\beta}] = H^2\rho = h^2$ is said to be the narrow-sense heritability of the trait, while the set of nonlinear interactions involving some subset of causal SNPs contribute the remaining $\mathbb{V}[\mathbf{W}\boldsymbol{\theta}] = H^2(1-\rho)$ to the overall broad-sense heritability.

Additive GWA summary statistics assuming no epistasis

In traditional GWA studies, genetic interactions are assumed to play a negligible role on the overall architecture of complex traits (i.e., $\rho \approx 1$ or $\boldsymbol{\theta} = \mathbf{0}$)[23,24,26]; therefore, summary statistics of the true additive effects $\boldsymbol{\beta}$ in Eq. (6) are typically derived by computing a marginal least squares estimate with the observed data

$$\widehat{\beta}_j = (\mathbf{x}_j^\intercal \mathbf{x}_j)^{-1}\mathbf{x}_j^\intercal \mathbf{y} \qquad \Longleftrightarrow \qquad \widehat{\boldsymbol{\beta}} = \text{diag}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}. \tag{9}$$

There are two key identities that may be taken from Eq. (9). The first uses Eq. (7) and is the approximate relationship (in expectation) between the moment matrix $\mathbf{X}^\intercal\mathbf{y}$ and the additive effect size estimates $\widehat{\boldsymbol{\beta}}$:

$$\mathbf{X}^\intercal\mathbf{y} = \text{diag}(\mathbf{X}^\intercal\mathbf{X})\widehat{\boldsymbol{\beta}} \approx N\widehat{\boldsymbol{\beta}}. \tag{10}$$

The second key point combines Eqs. (7) and (10) and describes the asymptotic relationship between the observed marginal GWA summary statistics $\widehat{\boldsymbol{\beta}}$ and the true coefficient values $\boldsymbol{\beta}$ where

$$\boldsymbol{\beta} = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y} \approx (N\mathbf{R})^{-1}N\widehat{\boldsymbol{\beta}} = \mathbf{R}^{-1}\widehat{\boldsymbol{\beta}}. \tag{11}$$

After some algebra, the above mirrors a high-dimensional regression model where $\widehat{\boldsymbol{\beta}} = \mathbf{R}\boldsymbol{\beta}$ with the estimated summary statistics as the response variables and the empirically estimated LD matrix acting as the design matrix[13,52,53,55,57]. Theoretically, the resulting output coefficients from this high-dimensional model are the desired true effect size estimates used to generate the phenotype of interest.

Additive GWA summary statistics with tagged epistasis

When genetic interactions do significantly contribute to the architecture of complex traits (i.e., $\rho < 1$ or $\boldsymbol{\theta} \neq \mathbf{0}$), the marginal GWA summary statistics derived using least squares in Eq. (9) can be confounded if there is a nonzero correlation between genotypes and and their epistatic interactions. To see this, we take the joint solution for the true regression coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ from the generative model in Eq. (6)

$$\begin{bmatrix}\boldsymbol{\beta} \\ \boldsymbol{\theta}\end{bmatrix} = \begin{bmatrix}\mathbf{X}^\intercal\mathbf{X} & \mathbf{X}^\intercal\mathbf{W} \\ \mathbf{W}^\intercal\mathbf{X} & \mathbf{W}^\intercal\mathbf{W}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{X}^\intercal \\ \mathbf{W}^\intercal\end{bmatrix}\mathbf{y}, \tag{12}$$

where the matrix $\mathbf{X}^\intercal\mathbf{W}$ can be interpreted as the sample correlation between individual-level genotypes and the epistatic interactions between causal SNPs. By solving for the additive genetic effects (again in expectation using Eqs. (7) and (10)), we get the following alternative relationship between the moment matrix $\mathbf{X}^\intercal\mathbf{y}$, the observed marginal GWA summary statistics $\widehat{\boldsymbol{\beta}}$, and the true coefficient values $\boldsymbol{\beta}$ where

$$\mathbf{X}^\intercal\mathbf{y} = (\mathbf{X}^\intercal\mathbf{X})\boldsymbol{\beta} + (\mathbf{X}^\intercal\mathbf{W})\boldsymbol{\theta} \qquad \overset{\approx}{\Longleftrightarrow} \qquad \widehat{\boldsymbol{\beta}} = \mathbf{R}\boldsymbol{\beta} + \mathbf{V}\boldsymbol{\theta}. \tag{13}$$

Here, we define $\mathbf{V}$ to represent an empirical estimate of the correlation between the individual-level genotypes and the non-additive genetic interaction matrix such that $\mathbb{E}[\mathbf{X}^\intercal\mathbf{W}] \approx N\mathbf{V}$. Similar to the LD matrix $\mathbf{R}$, the correlation matrix $\mathbf{V}$ is also assumed to be computed from reference panel data. Intuitively, when $\mathbf{V}\boldsymbol{\theta} \neq \mathbf{0}$ there is additional bias in the effect size estimates, and when $\mathbf{V}\boldsymbol{\theta} = \mathbf{0}$ then the relationship in Eq. (13) converges onto the conventional asymptotic assumption between GWA summary statistics and the true SNP additive effects in Eq. (11) [13,52,53,55,57].

Full derivation of marginal epistatic LD score regression

In order to derive the marginal epistatic LD score regression framework, recall that our goal is to identify evidence of tagged epistatic effects within misestimated GWA summary statistics. To do this, we build upon the LD score regression framework and the `LDSC` software [47]. Much of the derivation in this section will be done mirroring this previous work. Here, we assume nonzero contributions from epistatic effects in the generative model of complex traits as in Eq. (13), and we use the observed least squares estimates from Eq. (9) to compute chi-square statistics $\chi_j^2 = N\widehat{\beta}_j^2$ for every $j = 1, \ldots, J$ SNP in the data. Taking the expectation of these chi-square statistics yields

$$\mathbb{E}[\chi_j^2] = N\mathbb{E}[\widehat{\beta}_j^2] = N\left[\mathbb{V}[\widehat{\beta}_j] + \left(\mathbb{E}[\widehat{\beta}_j]\right)^2\right]. \tag{14}$$

We can simplify Eq. (14) in two steps. First, by combining the prior assumption in Eq. (8) and the asymptotic approximation in Eq. (13), we can show that marginal expectation (i.e., when not conditioning on the true coefficients) $\mathbb{E}[\widehat{\beta}_j] = 0$ for all variants. Second, by conditioning on the generative model from

Eq. (6), we can use the law of total variance to simplify $\mathbb{V}[\widehat{\beta}_j]$ where

$$\begin{aligned}\mathbb{V}[\widehat{\beta}_j] = \mathbb{E}[\mathbb{V}[\widehat{\beta}_j \mid \mathbf{X}]] + \mathbb{V}[\mathbb{E}[\widehat{\beta}_j \mid \mathbf{X}]] &\approx \mathbb{E}[\mathbb{V}[\mathbf{x}_j^\intercal \mathbf{y}/N \mid \mathbf{X}]] + 0 \\ &= \mathbb{E}\left[\frac{1}{N^2}\mathbf{x}_j^\intercal \left\{\mathbb{V}[\mathbf{y} \mid \mathbf{X}]\right\} \mathbf{x}_j\right] \\ &= \mathbb{E}\left[\frac{1}{N^2}\mathbf{x}_j^\intercal \left\{\frac{H^2\rho}{J}\mathbf{X}\mathbf{X}^\intercal + \frac{H^2(1-\rho)}{M}\mathbf{W}\mathbf{W}^\intercal + (1-H^2)\right\} \mathbf{x}_j\right] \\ &= \mathbb{E}\left[\frac{1}{N^2}\left\{\frac{H^2\rho}{J}\mathbf{x}_j^\intercal\mathbf{X}\mathbf{X}^\intercal\mathbf{x}_j + \frac{H^2(1-\rho)}{M}\mathbf{x}_j^\intercal\mathbf{W}\mathbf{W}^\intercal\mathbf{x}_j + N(1-H^2)\right\}\right].\end{aligned}$$

Using the same logic from the original `LDSC` regression framework[47], we can use Isserlis' theorem[71] to write the above in terms of more familiar quantities based on sample correlations

$$\frac{1}{N^2}\mathbf{x}_j^\intercal\mathbf{X}\mathbf{X}^\intercal\mathbf{x}_j = \sum_{k=1}^{J}\widetilde{r}_{jk}^2, \qquad \frac{1}{N^2}\mathbf{x}_j^\intercal\mathbf{W}\mathbf{W}^\intercal\mathbf{x}_j = \sum_{m=1}^{M}\widetilde{v}_{jm}^2 \tag{15}$$

where $\widetilde{r}_{jk}$ is used to denote the sample correlation between additively-coded genotypes at the j-th and k-th variants, and $\widetilde{v}_{jm}$ is used to denote the sample correlation between the genotype of the j-th variant and the m-th epistatic interaction on the phenotype of interest (again see Eq. (13)). Furthermore, we can use the delta method (only displaying terms up to $\mathcal{O}(1/N^2)$) to show that (in expectation)

$$\mathbb{E}[\widetilde{r}_{jk}^2] \approx r_{jk}^2 + (1 - r_{jk}^2)/N, \qquad \mathbb{E}\left[\widetilde{v}_{jm}^2\right] \approx v_{jm}^2 + \left(1 - v_{jm}^2\right)/N. \tag{16}$$

Next, we can then approximate the quantities in Eq. (15) via the following

$$\mathbb{E}\left[\sum_{k=1}^{J}\widetilde{r}_{jk}^2\right] \approx \ell_j + (J - \ell_j)/N, \qquad \mathbb{E}\left[\sum_{m=1}^{M}\widetilde{v}_{jm}^2\right] \approx f_j + (M - f_j)/N \tag{17}$$

where ℓ_j is the corresponding LD score for the additive effect of the j-th variant and f_j represents the "marginal epistatic" LD score between the j-th SNP and all other variants in the data set[51], respectively. Altogether, this leads to the specification of the univariate framework with the j-th SNP

$$\mathbb{E}[\chi_j^2] \approx N\left[\left(\frac{H^2\rho}{J}\right)\ell_j + \left(\frac{H^2(1-\rho)}{M}\right)f_j + \frac{1}{N}(1-H^2)\right] = \ell_j\tau + f_j\sigma + 1 \tag{18}$$

where we define $\tau = NH^2\rho/J$ as estimates of the true additive genetic signal, the coefficient $\sigma = NH^2(1-\rho)/M$ as an inflation or deflation factor due to tagged epistasis, and $\mathbf{1}$ is the intercept meant to model the misestimation due to uncontrolled confounding effects. Similar to the original `LDSC` formulation, an intercept greater than one means significant bias from sources other than polygenicity. Note that the simplification for many of the terms above such as $(1-H^2)/N \approx 1/N$ results from our assumption that the number of individuals in our study is large. For example, the sample sizes for each biobank-scale study considered in the analyses of this manuscript are at least on the order of $N \geq 10^4$ observations (see Table S5). Altogether, we can jointly express Eq. (18) in multivariate form as the following

$$\mathbb{E}[\boldsymbol{\chi}^2] \approx \boldsymbol{\ell}\tau + \boldsymbol{f}\sigma + \mathbf{1} \tag{19}$$

where $\boldsymbol{\chi}^2 = (\chi_1^2, \ldots, \chi_J^2)$ is a J-dimensional vector of chi-square summary statistics, and $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_J)$ and $\boldsymbol{f} = (f_1, \ldots, f_J)$ are J-dimensional vectors of additive and marginal epistatic LD scores, respectively. It is important to note that, while $\boldsymbol{\chi}^2$ must be recomputed for each trait of interest, both vectors $\boldsymbol{\ell}$ and $\boldsymbol{f}$ only need to be constructed once per reference panel or individual-level genotypes (see next section for efficient computational strategies).

To identify summary statistics that have significant tagged epistatic signal, we test the null hypothesis $H_0 : \sigma = 0$. The `MELD` software package implements the same model fitting strategy as `LDSC`. Here, we use weighted least squares to fit the joint regression in Eq. (19) such that

$$\widehat{\sigma} = (\boldsymbol{f}^\intercal \boldsymbol{\Psi} \boldsymbol{f})^{-1} \boldsymbol{f}^\intercal \boldsymbol{\Psi} \boldsymbol{\chi}^2, \qquad \psi_{jj} = [\ell_j \widehat{\tau} + f_j \widehat{\sigma} + 1]^{-2} \tag{20}$$

where $\boldsymbol{\Psi}$ is a $J \times J$ diagonal weight matrix with nonzero elements set to values inversely proportional to the conditional variance $\mathbb{V}[\chi_j^2 \,|\, \ell_j, f_j] = \psi_{jj}^{-1}$ to adjust for both heteroscedasticity and over-estimation of the summary statistics for each SNP[47]. Standard errors for each coefficient estimate are derived via a delete-one jackknife over blocks of SNPs in the data[64], and we then use those standard errors to derive P-values with a two-sided test (i.e., testing the alternative hypothesis $H_A : \sigma \neq 0$). For all analyses in this paper, we estimate narrow-sense heritability using a de-biased coefficient which is computed by taking the difference between $\widehat{\tau} - \widehat{\sigma}$ (i.e., the estimated additive component minus the inflation or deflation that stems from tagged pairwise genetic effects).

Efficient computation of marginal epistatic LD scores

In practice, marginal epistatic LD scores in `MELD` can be computed efficiently through realizing two key opportunities for optimization. First, given J SNPs, the full matrix of genome-wide interaction effects $\mathbf{W}$ contains on the order of $J(J-1)/2$ total pairwise interactions. However, the correlation between the genotype of the j-th SNP and the interactions where its involved (i.e., $\mathbf{x}_j^\intercal(\mathbf{x}_j \circ \mathbf{x}_l)$ for $l \neq j$) is bound to be much larger than the correlation between the genotype of the j-th SNP $\mathbf{x}_j$ and interactions involving some other SNP (e.g., $\mathbf{x}_j^\intercal(\mathbf{x}_k \circ \mathbf{x}_l)$ for $k \neq j$ and $l \neq j$). To that end, we can compute the `MELD` score for each SNP by replacing the full $\mathbf{W}$ matrix with $\mathbf{W}_j$ which includes only interactions involving the j-th SNP. Analogous to the original `LDSC` formulation[47], we consider only interactive SNPs within a *cis*-window proximal to the focal j-th SNP for which we are computing the `MELD` score. In the original `LDSC` methodology, this is based on the observation that LD decays outside of a window of 1 centimorgan (cM); therefore, SNPs outside the 1cM window centered on the j-th SNP j will not significantly contribute to its LD score.

The second opportunity for optimization comes from the fact that the matrix of interaction effects, $\mathbf{W}_j$, does not ever need to be explicitly generated. Referencing Eq. (15), the `MELD` scores are defined as $\mathbf{x}_j^\intercal \mathbf{W}_j \mathbf{W}_j^\intercal \mathbf{x}_j / N^2$. This can be re-written as $\mathbf{x}_j^\intercal(\mathbf{D}_j\mathbf{X}^{(j)})(\mathbf{D}_j\mathbf{X}^{(j)})^\intercal\mathbf{x}_j$, where $\mathbf{D}_j = \text{diag}(\mathbf{x}_j)$ is a diagonal matrix with the j-th genotype as its nonzero elements[51] and $\mathbf{X}^{(j)}$ denotes the subset SNPs within a *cis*-window proximal to the focal j-th SNP. This means that the `MELD` score for the j-th SNP can be simply computed as the following

$$f_j \approx \frac{1}{N^2}(\mathbf{x}_j^\intercal)^2 \mathbf{X}^{(j)} \mathbf{X}^{(j)\intercal} (\mathbf{x}_j)^2. \tag{21}$$

With these simplifications, the computational complexity of generating `MELD` scores reduces to that of computing LD scores — modulo a vector-by-vector Hadamard product which, for each SNP, is constant factor of N (i.e., the number of genotyped individuals).

Model averaged coefficient estimates

When computing the marginal epistatic LD scores, the most important decision is choosing the number of interacting SNPs to include in $\mathbf{X}^{(j)}$ (or equivalently $\mathbf{W}_j$ for each j-th focal SNP in the calculation of f_j in Eq. (21). The `MELD` framework considers different estimating windows to account for our lack of *a*

priori knowledge about the "correct" non-additive genetic architecture of traits. Here, we follow previous work[50,54,56–58,72] by considering an L-valued grid of possible SNP interaction window sizes. After fitting a series of `MELD` regressions with marginal epistatic LD scores $\boldsymbol{f}^{(l)}$ generated under the L-different window sizes, we compute normalized importance weights using their maximized likelihoods via the following

$$\pi^{(l)} = \frac{\mathcal{L}\left(\boldsymbol{\ell}, \boldsymbol{f}^{(l)}; \widehat{\boldsymbol{\beta}}\right)}{\sum_{l'} \mathcal{L}\left(\boldsymbol{\ell}, \boldsymbol{f}^{(l')}; \widehat{\boldsymbol{\beta}}\right)}, \qquad \sum_{l=1}^{L} \pi^{(l)} = 1. \tag{22}$$

As a final step in the model fitting procedure, we empirically compute averaged estimates of the coefficients τ and σ by marginalizing (or averaging) over the L-different grid combinations of estimating windows

$$\widehat{\tau} = \sum_{l=1}^{L} \pi^{(l)} \widehat{\tau}^{(l)}, \qquad \widehat{\sigma} = \sum_{l=1}^{L} \pi^{(l)} \widehat{\sigma}^{(l)}. \tag{23}$$

This final step can be viewed as an analogy to model averaging where marginal estimates are computed via a weighted average using the importance weights[73]. In the current study, we average over estimated marginal epistatic LD scores generated using different windows of ± 5, ± 10, ± 25, and ± 50 SNPs around each j-th focal SNP.

Relationship between minor allele frequency and effect size

The `LDSC` software computes LD scores using annotations over equally spaced minor allele frequency (MAF) bins. These annotations enable the per trait relationship between the MAF and the effect size of each variant in the genome to vary based on the discrete category (or MAF bin) it is placed into. This additional flexibility is intended to help `LDSC` be more robust when estimating narrow-sense heritability. The relationship between MAF and effect size is already implicitly encoded in the `LDSC` formulation since we assume genotypes are normalized. When normalizing by the variance of each SNP (or equivalently its MAF), we make the assumption that rare variants inherently have larger effect sizes. There exists a true functional relationship between MAF and effect size which is likely to be somewhere between the two extremes of (i) normalizing each SNP by its MAF and (ii) allowing the variance per SNP to be dictated by its MAF.

Recent approaches have proposed using a single parameter α to better represent the nonlinear relationship between MAF and variant effect size. The main idea is that this α not only provides the same

additional flexibility to LDSC as the MAF-based discrete annotations, but it also empirically yields even more precise narrow-sense heritability estimates [74]. Namely, we use

$$\ell_j(c) := \sum_k L_{jk}(\alpha) a_c(k), \qquad L_{jk}(\alpha) = r_{jk}^2 \mathbb{V}[\mathbf{x}_k]^{1-\alpha} \tag{24}$$

where $a_c(k)$ is the annotation value for the c-th categorical bin. The α parameter is unknown in practice and needs to be estimated for any given trait. While standard ranges for α can be used for heritability estimates, we use a restricted maximum likelihood (REML) based method which was recently developed [59].

In the MELD software, we use this α construction to handle the relationship between MAF and variant effect size for two specific reasons. First, by constructing the LD scores using α, we more accurately capture the variation in chi-square test statistics due to additive effects [74]. Second, we note that there is correlation between MAF and (*i*) LD scores, (*ii*) marginal epistatic LD scores, and (*iii*) trait architecture. To that end, if we do not properly condition on MAF, there becomes additional bias, and we may falsely attribute some amount of variation in the chi-square test statistics to LD or the tagged epistasis. Therefore, in our formulation, we include an α term on the LD scores to condition on this effect. We demonstrate in simulation that this removes the bias introduced by the relationship between MAF and trait architecture, and it mitigates potential inflation of type I error rates in the MELD test.

Estimation of allele frequency parameters

In the main text, we analyzed 25 complex traits in both the UK Biobank and BioBank Japan data sets. In order to account for minor allele frequency (MAF) dependent trait architecture, we calculated α values for each trait that had not been analyzed by previous studies [59]. The α estimates for each of the 25 traits analyzed in this study are shown in Table S4. Intuitively, α parameterizes the weighting of the effects of each individual variant given its frequency in the study cohort and can take on values in the range of [-1,0]. More negative values of α indicate that lower frequency variants contribute more to the observed variation in a trait of interest, whereas values of α closer to zero indicate that common variants contribute a greater amount of variation to observed trait values.

We took α values for 11 traits (again see Table S4) that had previously been calculated from Schoech et al. [59]. For the remaining 14 traits analyzed in this study, we followed the estimation protocol described in the same manuscript. Specifically, using the variants passing the quality control step in our pipeline for

25,000 randomly selected individuals in the UK Biobank cohort, we constructed MAF-dependent genetic relatedness matrices for values of $\alpha = \{-1, -0.95, -0.9, \ldots, 0\}$ using the `GRM-MAF-LD` software, `https://github.com/arminschoech/GRM-MAF-LD`. We then used the `GCTA` software[75] to obtain heritability and likelihood estimates using REML for each α-trait pairing. We then fit a trait-specific profile likelihood across the range of α values and estimate the maximum likelihood value of α using a natural cubic spline.

Simulation studies

We used a simulation scheme to generate synthetic quantitative traits and SNP-level summary statistics under multiple genetic architectures using real genome-wide data from individuals of self-identified European ancestry in the UK Biobank. First, we assume that every SNP in the genome has at least a small additive effect on the traits of interest. Next, we randomly select a subset of SNPs to have nonzero epistatic effects and assume that complex traits are generated via the following general linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \kappa^2\mathbf{I}), \tag{25}$$

where $\mathbf{y}$ is an N-dimensional vector containing all the phenotypes; $\mathbf{X}$ is an $N \times J$ matrix of genotypes encoded as 0, 1, or 2 copies of a reference allele; $\boldsymbol{\beta}$ is a J-dimensional vector of additive effect sizes for each SNP; $\mathbf{W}$ is an $N \times M$ matrix which holds all pairwise interactions between the randomly selected subset of the interacting SNPs with corresponding effects $\boldsymbol{\theta}$; and $\boldsymbol{\varepsilon}$ is an N-dimensional vector of environmental noise. The phenotypic variance is assumed to be $\mathbb{V}[\mathbf{y}] = 1$. The additive and interaction effect sizes for SNPs are randomly drawn from independent standard normal distributions and then rescaled so that they explain a fixed proportion of the broad-sense heritability $\mathbb{V}[\mathbf{X}\boldsymbol{\beta}] + \mathbb{V}[\mathbf{W}\boldsymbol{\theta}] = H^2$. Note that we do not assume any specific correlation structure between the effect sizes $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. We then rescale the random error term such that $\mathbb{V}[\boldsymbol{\varepsilon}] = (1 - H^2)$. In the main text, we compare the traditional `LDSC` to its direct extension in `MELD`. For each method, GWA summary statistics are computed by fitting a single-SNP univariate linear model via least squares where $\widehat{\beta}_j = (\mathbf{x}_j^\intercal \mathbf{x}_j)^{-1}\mathbf{x}_j^\intercal \mathbf{y}$ for every $j = 1, \ldots, J$ SNP in the data. These effect size estimates are used to derive the chi-square test statistics $\chi_j^2 = N\widehat{\beta}_j^2$. We implement both `LDSC` and `MELD` with the LD matrix $\mathbf{R} = \mathbf{X}^\intercal\mathbf{X}/N$ and the additive-epistatic correlation matrix $\mathbf{V} = \mathbf{X}^\intercal\mathbf{W}/N$ being computed using a reference panel of 489 individuals from the European superpopulation (EUR) of the 1000 Genomes Project. The resulting matrices $\mathbf{R}$ and $\mathbf{V}$ are used to

compute the LD scores and marginal epistatic LD scores, respectively.

When generating synthetic traits, we assume that the additive effects make up $\rho\%$ of the broad-sense heritability while the pairwise interactions make up the remaining $(1-\rho)\%$. Alternatively, the proportion of the heritability explained by additivity is said to be $\mathbb{V}[\mathbf{X}\boldsymbol{\beta}] = \rho H^2$, while the proportion detailed by genetic interactions is given as $\mathbb{V}[\mathbf{W}\boldsymbol{\theta}] = (1-\rho)H^2$. The setting of $\rho = 1$ represents the limiting null case for `MELD` where the variation of a trait is driven by solely additive effects. Here, we use the same simulation strategy used in Crawford et al.[51] where we divide the causal epistatic variants into two groups. One may view the SNPs in group #1 as being the "hubs" of an interaction map. SNPs in group #2 are selected to be variants within some kilobase (kb) window around each SNP in group #1. Given different parameters for the generative model in Eq. (25), we simulate data mirroring a wide range of genetic architectures by toggling the following parameters:

- broad-sense heritability: $H^2 = 0.3$ and 0.6;
- proportion of phenotypic variation that is explained by additive effects: $\rho = 0.5$, 0.8, and 1;
- percentage of SNPs selected to be in group #1: 1% (sparse), 5%, and 10% (polygenic);
- genomic window used to assign SNPs to group #2: $\pm$10 and $\pm$100 kilobase (kb);
- allele frequency parameter: α = -1, -0.5, and 0.

All figures and tables show the mean performances (and standard errors) across 100 simulated replicates.

Preprocessing for the UK Biobank and BioBank Japan

In order to apply the the `MELD` framework to 25 continuous traits the UK Biobank[76], we first downloaded genotype data for 488,377 individuals in the UK Biobank using the `ukbgene` tool (`https://biobank.ctsu.ox.ac.uk/crystal/download.cgi`) and converted the genotypes using the provided `ukbconv` tool (`https://biobank.ctsu.ox.ac.uk/crystal/refer.cgi?id=149660`). Phenotype data for the 25 continuous traits were also downloaded for those same individuals using the `ukbgene` tool. Individuals identified by the UK Biobank as having high heterozygosity, excessive relatedness, or aneuploidy were removed (1,550 individuals). After then separating individuals into self-identified ancestral cohorts using data field `21000`, unrelated individuals were selected by randomly choosing an individual from each pair of related individuals. This resulted in N =349,469 white British individuals to be

included in our analysis. We downloaded imputed SNP data from the UK Biobank for all remaining individuals and removed SNPs with an information score below 0.8. Information scores for each SNP are provided by the UK Biobank (`http://biobank.ctsu.ox.ac.uk/crystal/refer.cgi?id=1967`).

Quality control for the remaining genotyped and imputed variants was then performed on each cohort separately using the following steps. All structural variants were first removed, leaving only single nucleotide polymorphisms (SNPs) in the genotype data. Next, all AT/CG SNPs were removed to avoid possible confounding due to sequencing errors. Then, SNPs with minor allele frequency less than 1% were removed using the `PLINK 2.0`[77] command `--maf 0.01`. We then removed all SNPs found to be in Hardy-Weinberg equilibrium, using the `PLINK` `--hwe 0.000001` flag to remove all SNPs with a Fisher's exact test P-value $> 10^{-6}$. Finally, all SNPs with missingness greater than 1% were removed using the `PLINK` `--mind 0.01` flag.

We then performed a genome-wide association (GWA) study for each trait in the UK Biobank on the remaining 8,981,412 SNPs. SNP-level GWA effect sizes were calculated using `PLINK` and the `--glm` flag[77]. Age, sex, and the first twenty principal components were included as covariates for all traits analyzed[78]. Principal component analysis was performed using `FlashPCA 2.0`[79] on a set of independent markers derived separately for each ancestry cohort using the `PLINK` command `--indep-pairwise 100 10 0.1`. Using the parameters `--indep-pairwise` removes all SNPs that have a pairwise correlation above 0.1 within a 100 SNP window, then slides forward in increments of ten SNPs genome-wide.

In order to analyze data from BioBank Japan, we downloaded publicly available GWA summary statistics for the 25 traits listed in Table S5 from `http://jenger.riken.jp/en/result`. Summary statistics used age, sex, and the first ten principal components as confounders in the initial GWA study. We then used individuals from the East Asian (EAS) superpopulation from the 1000 Genomes Project Phase 3 to calculate paired `LDSC` and `MELD` scores from a reference panel. We pruned the reference panel using the `PLINK` command `--indep-pairwise 100 10 0.5` to limit the computational time of calculating scores[77]. This resulted in reference scores for 1,164,666 SNPs that are included on the `MELD` GitHub repository (see URLs). Using summary statistics from BioBank Japan, with scores calculated from the EAS population in the 1000 Genomes, we obtained `MELD` narrow-sense heritability estimates for each of the 25 traits.

Data and software availability

Source code and tutorials for implementing marginal epistatic LD score regression via the `MELD` package are written in Python and are publicly available online at `https://github.com/lcrawlab/MELD`. All software for the traditional LD score regression framework with `LDSC` were fit using the default settings, unless otherwise stated in the main text. Source code for `LDSC` was downloaded from `https://github.com/bulik/ldsc`. Data from the UK Biobank Resource[76] (`https://www.ukbiobank.ac.uk`) was made available under Application Numbers 14649 and 22419. Data can be accessed by direct application to the UK Biobank.

Figures

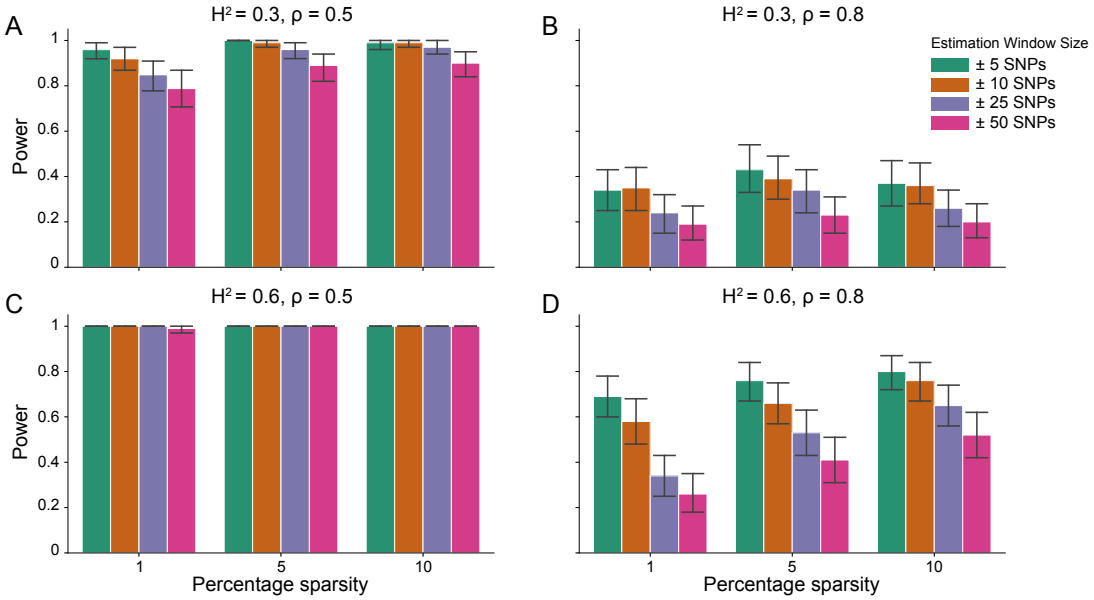

Figure 1. Power calculations for the MELD framework on simulated data. Synthetic trait architecture was simulated using real genotype data from individuals of self-identified European ancestry in the UK Biobank. All SNPs were considered to have at least an additive effect (i.e., creating a polygenic trait architecture). Next, we randomly select two groups of interacting variants and divide them into two interacting groups. The group #1 SNPs are chosen to be 1%, 5%, and 10% of the total number of SNPs genome-wide (see the x-axis in each panel). These interact with the group #2 SNPs which are selected to be variants within a ±10 kilobase (kb) window around each SNP in group #1. Coefficients for additive and interaction effects were simulated with no minor allele frequency dependency $\alpha = 0$ (see Materials and Methods). Panels **(A)** and **(B)** are results with simulations using a broad-sense heritability $H^2 = 0.3$, while panels **(C)** and **(D)** were generated with $H^2 = 0.6$. We also varied the proportion of broad-sense heritability contributed by additive effects to **(A, C)** $\rho = 0.5$ and **(B, D)** $\rho = 0.8$, respectively. Here, we are blind to the parameter settings used in generative model and run MELD while computing the marginal epistatic LD scores using different estimating windows of ±5 (green), ±10 (orange), ±25 (purple), and ±50 (pink) SNPs. Results are based on 100 simulations per parameter combination and the horizontal bars represent standard errors. Generally, the performance of MELD increases with larger broad-sense heritability and lower proportions of additive variation. Note that LDSC is not shown here because it does not search for tagged epistatic effects in summary statistics. Similar plots for a range of α values and generative interacting SNP window sizes are shown in Figures S1-S5.

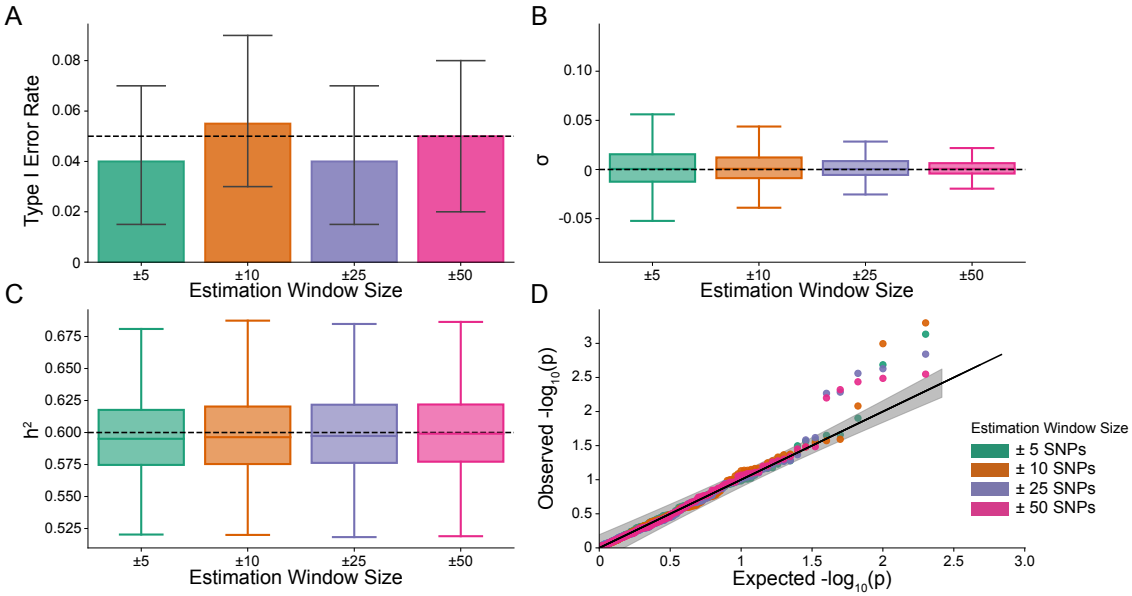

Figure 2. The MELD framework is well-calibrated and does not identify evidence of tagged epistasis when traits are generated by only additive effects. In these simulations, synthetic trait architecture is made up of only additive genetic variation (i.e., $\rho = 1$). Coefficients for additive and interaction effects were simulated with no minor allele frequency dependency $\alpha = 0$ (see Materials and Methods). Here, we are blind to the parameter settings used in generative model and run MELD while computing the marginal epistatic LD scores using different estimating windows of ± 5 (green), ± 10 (orange), ± 25 (purple), and ± 50 (pink) SNPs. **(A)** Mean type I error rate using the MELD framework across an array of estimation window sizes for the marginal epistatic scores. This is determined by assessing the P-value of the epistatic coefficient (σ) in the MELD regression model and checking whether $P < 0.05$. **(B)** Estimates of the epistatic coefficient (σ). Since traits were simulated with only additive effects, these estimates should be centered around zero. **(C)** Narrow-sense heritability (h^2) estimates where the true value is $H^2\rho = h^2 = 0.6$. **(D)** QQ-plot of the P-values for the epistatic coefficient (σ) in MELD. Results are based on 100 simulations per parameter combination and the horizontal bars represent standard errors. Similar plots for a range of α values and generative interacting SNP window sizes are shown in Figures S6-S7.

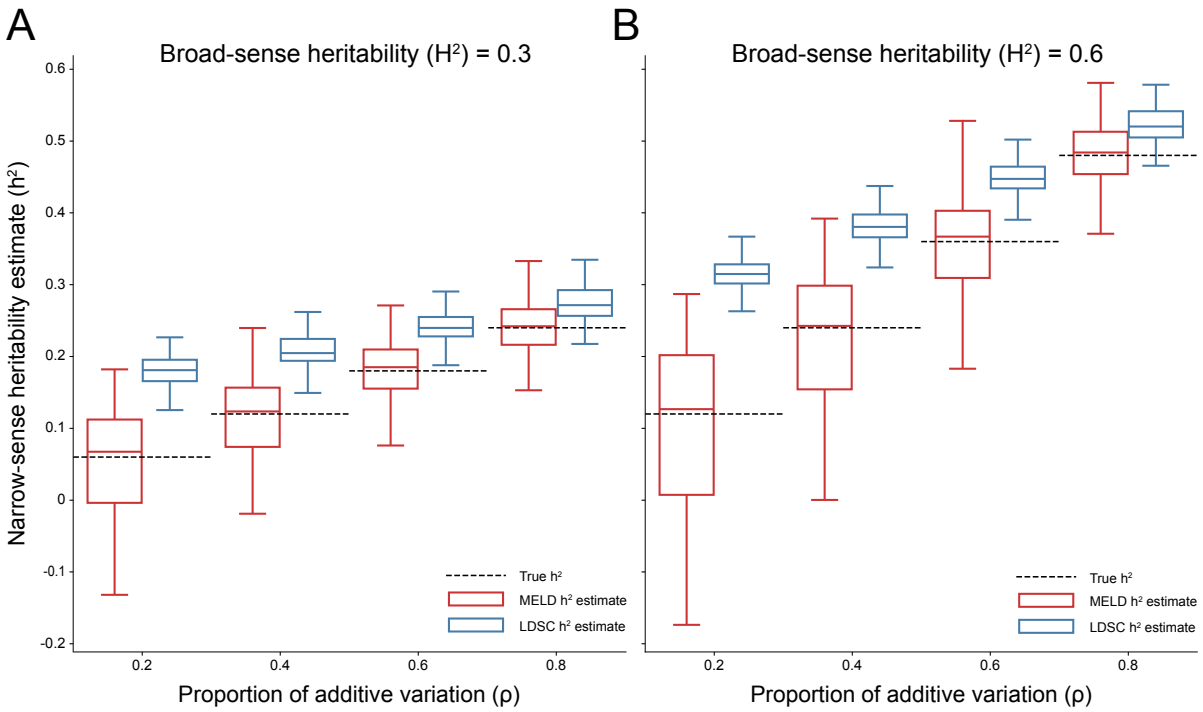

Figure 3. `MELD` robustly and accurately estimates narrow-sense heritability in simulations, compared to `LDSC`, due to our accounting for epistatic signals in additive GWA summary statistics. Synthetic trait architecture was simulated using real genotype data from individuals of self-identified European ancestry in the UK Biobank (Materials and Methods). All SNPs were considered to have at least an additive effect (i.e., creating a polygenic trait architecture). Next, we randomly select two groups of interacting variants and divide them into two interacting groups. The group #1 SNPs are chosen to be 10% of the total number of SNPs genome-wide. These interact with the group #2 SNPs which are selected to be variants within a ±100 kilobase (kb) window around each SNP in group #1. Coefficients for additive and interaction effects were simulated with no minor allele frequency dependency $\alpha = 0$ (see Materials and Methods). Here, we assume a broad-sense heritability **(A)** $H^2 = 0.3$ or **(B)** $H^2 = 0.6$, and we vary the proportion contributed by additive effects with $\rho = \{0.2, 0.4, 0.6, 0.8\}$. The true narrow-sense heritability is set as $H^2\rho = h^2$. `MELD` outperforms `LDSC` in each scenario. Results are based on 100 simulations per parameter combination. `MELD` estimates of narrow-sense heritability partitioned by estimation window are shown in Figure S8. The mean absolute error between the true h^2 value and the estimates produced by `MELD` and `LDSC` are shown in Table S1 and S2, respectively.

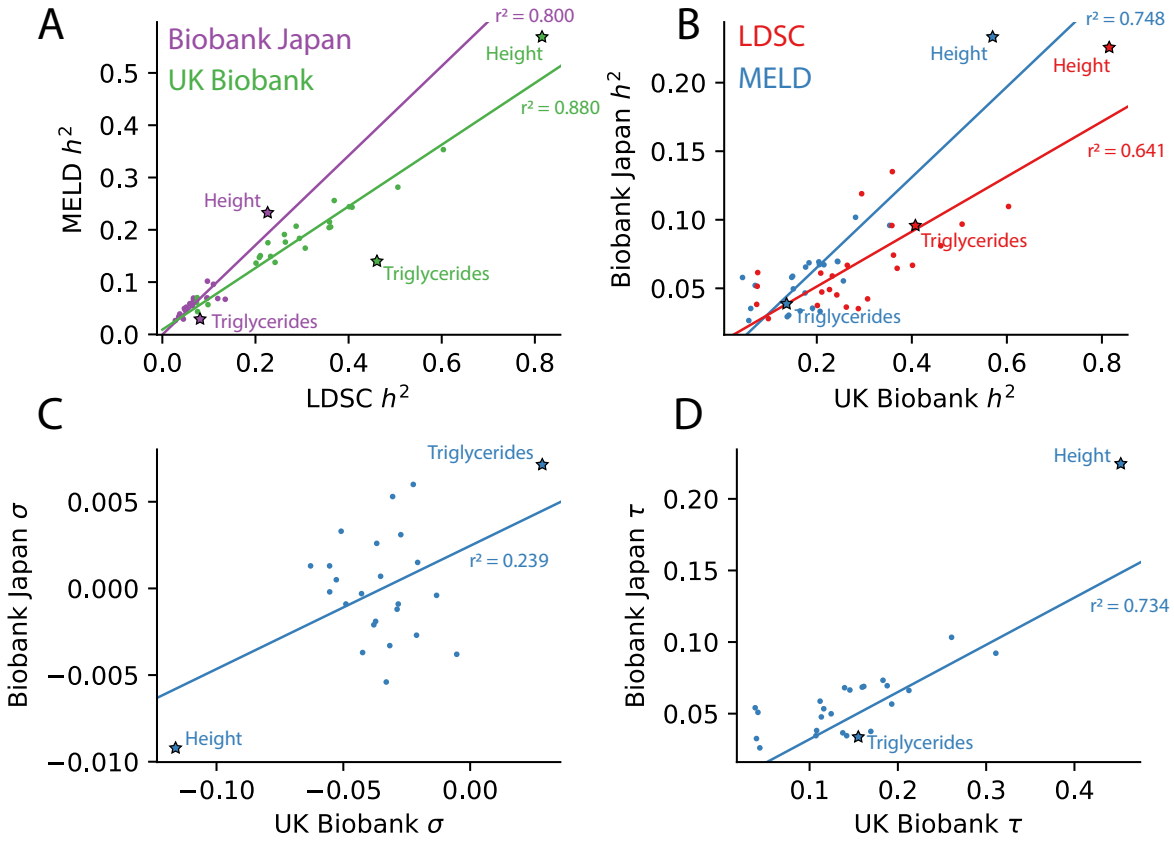

Figure 4. The `MELD` framework recovers narrow-sense heritability and provides estimates of tagged epistasis in GWA summary statistics (σ) for 25 quantitiative traits in the UK Biobank and BioBank Japan. (A) In both the UK Biobank (green) and BioBank Japan (purple), narrow-sense heritability estimates from `MELD` and `LDSC` are highly correlated for 25 different complex traits. The Spearman correlation coefficient between h^2 estimates for the UK Biobank and BioBank Japan is $r^2 = 0.880$ and $r^2 = 0.800$, respectively. **(B)** Narrow-sense heritability estimates from the UK Biobank are correlated with those from the BioBank Japan across 25 traits using both `LDSC` and `MELD`. Estimates from `MELD` are more agreeable (Spearman $r^2 = 0.748$) between biobanks than those from the original LD score regression model (Spearman $r^2 = 0.641$). **(C)** `MELD` estimates of the inflation or deflation due to tagged epistasis (i.e., estimates of σ) between traits in the UK Biobank and BioBank Japan. **(D)** `MELD` estimates of the additive coefficient τ. Note that the narrow-sense heritability estimates displayed in panels **(A)** and **(B)** are also given in Table S6.

Tables

Trait	UK Biobank	BioBank Japan
BMI	0.008	0.611
Basophil	0.290	0.301
CRP	0.005	0.928
Cholesterol	1.52×10^{-4}	0.262
DBP	5.76×10^{-6}	0.743
EGFR	3.41×10^{-4}	0.189
Eosinophil	4.21×10^{-10}	0.506
HBA1C	1.37×10^{-8}	0.925
HDL	7.00×10^{-11}	0.832
Height	1×10^{-22}	0.197
Hematocrit	1.51×10^{-8}	0.798
Hemoglobin	1.89×10^{-8}	0.883
LDL	5.37×10^{-5}	0.250
Lymphocyte	2.19×10^{-8}	0.830
MCH	3.66×10^{-5}	0.953
MCHC	4.91×10^{-4}	0.358
MCV	7.50×10^{-9}	0.961
Monocyte	2.84×10^{-7}	0.246
Neutrophil	0.002	0.121
Platelet	5.81×10^{-4}	0.253
RBC	2.99×10^{-10}	0.686
SBP	7.79×10^{-10}	0.558
Triglyceride	0.530	0.003
Urate	4.41×10^{-6}	0.582
WBC	1.33×10^{-7}	0.418

Table 1. `MELD` P-values for the estimated bias stemming from non-additive variation for 25 traits in the UK Biobank and BioBank Japan. Note that 23 of the 25 traits in the UK Biobank had a significant amount of uncorrected bias ($P < 0.05$), while one trait (Triglyceride) had significant tagged epistasis in the BioBank Japan. The two traits without significant tagged epistasis in the UK Biobank were Basophil ($P = 0.290$) and Triglyceride ($P = 0.530$).

Supplementary Figures

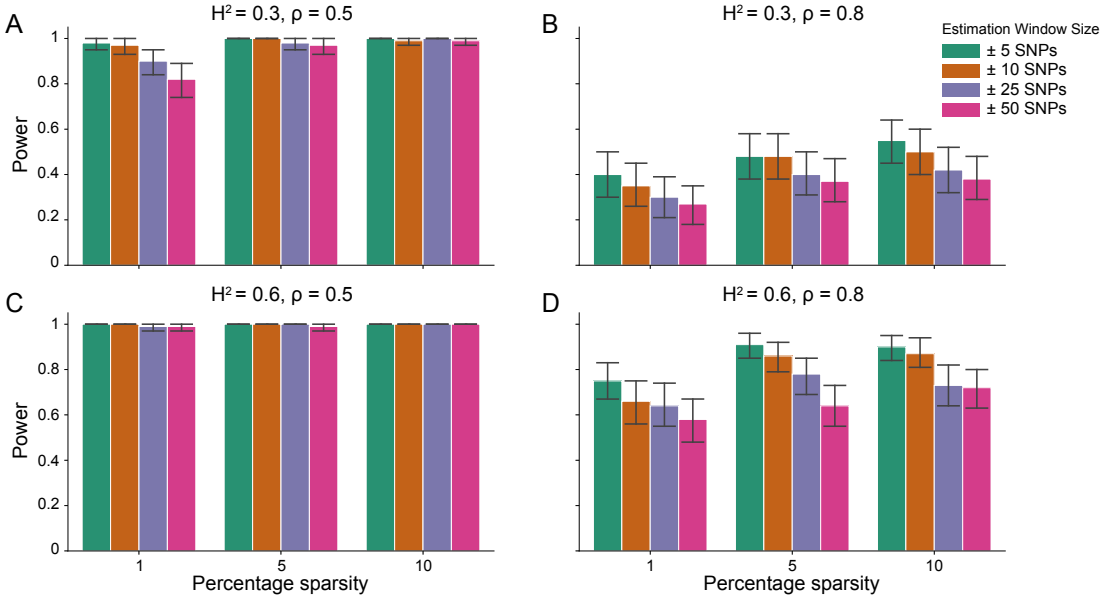

Figure S1. Power calculations for the MELD framework on simulated data using a ±10 kilobase (kb) window to generate pairwise interactions between causal SNPs and a moderate minor allele frequency dependency $\alpha = -0.5$ for effect sizes. Synthetic trait architecture was simulated using real genotype data from individuals of self-identified European ancestry in the UK Biobank. All SNPs were considered to have at least an additive effect (i.e., creating a polygenic trait architecture). Next, we randomly select two groups of interacting variants and divide them into two interacting groups. The group #1 SNPs are chosen to be 1%, 5%, and 10% of the total number of SNPs genome-wide (see the x-axis in each panel). These interact with the group #2 SNPs which are selected to be variants within a ±10 kilobase (kb) window around each SNP in group #1. Coefficients for additive and interaction effects were simulated with minor allele frequency dependency $\alpha = -0.5$ (see Materials and Methods). Panels **(A)** and **(B)** are results with simulations using a broad-sense heritability $H^2 = 0.3$, while panels **(C)** and **(D)** were generated with $H^2 = 0.6$. We also varied the proportion of broad-sense heritability contributed by additive effects to **(A, C)** $\rho = 0.5$ and **(B, D)** $\rho = 0.8$, respectively. Here, we are blind to the parameter settings used in generative model and run MELD while computing the marginal epistatic LD scores using different estimating windows of ±5 (green), ±10 (orange), ±25 (purple), and ±50 (pink) SNPs. Results are based on 100 simulations per parameter combination and the horizontal bars represent standard errors.

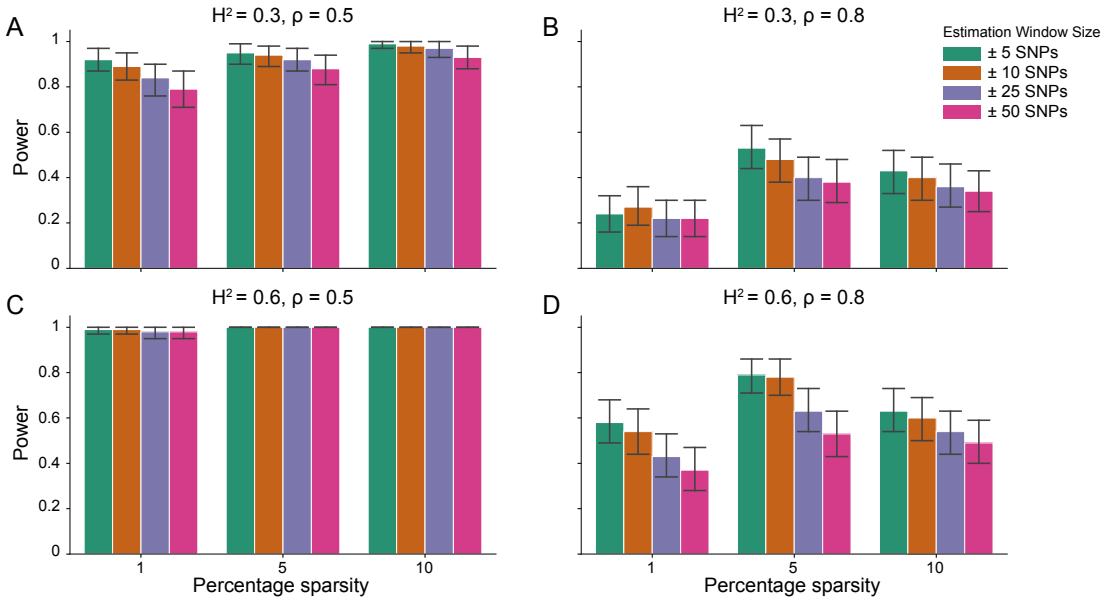

Figure S2. Power calculations for the `MELD` framework on simulated data using a ±10 kilobase (kb) window to generate pairwise interactions between causal SNPs and a strong minor allele frequency dependency $\alpha = -1$ for effect sizes. Synthetic trait architecture was simulated using real genotype data from individuals of self-identified European ancestry in the UK Biobank. All SNPs were considered to have at least an additive effect (i.e., creating a polygenic trait architecture). Next, we randomly select two groups of interacting variants and divide them into two interacting groups. The group #1 SNPs are chosen to be 1%, 5%, and 10% of the total number of SNPs genome-wide (see the x-axis in each panel). These interact with the group #2 SNPs which are selected to be variants within a ±10 kilobase (kb) window around each SNP in group #1. Coefficients for additive and interaction effects were simulated with minor allele frequency dependency $\alpha = -1$ (see Materials and Methods). Panels **(A)** and **(B)** are results with simulations using a broad-sense heritability $H^2 = 0.3$, while panels **(C)** and **(D)** were generated with $H^2 = 0.6$. We also varied the proportion of broad-sense heritability contributed by additive effects to **(A, C)** $\rho = 0.5$ and **(B, D)** $\rho = 0.8$, respectively. Here, we are blind to the parameter settings used in generative model and run `MELD` while computing the marginal epistatic LD scores using different estimating windows of ±5 (green), ±10 (orange), ±25 (purple), and ±50 (pink) SNPs. Results are based on 100 simulations per parameter combination and the horizontal bars represent standard errors.

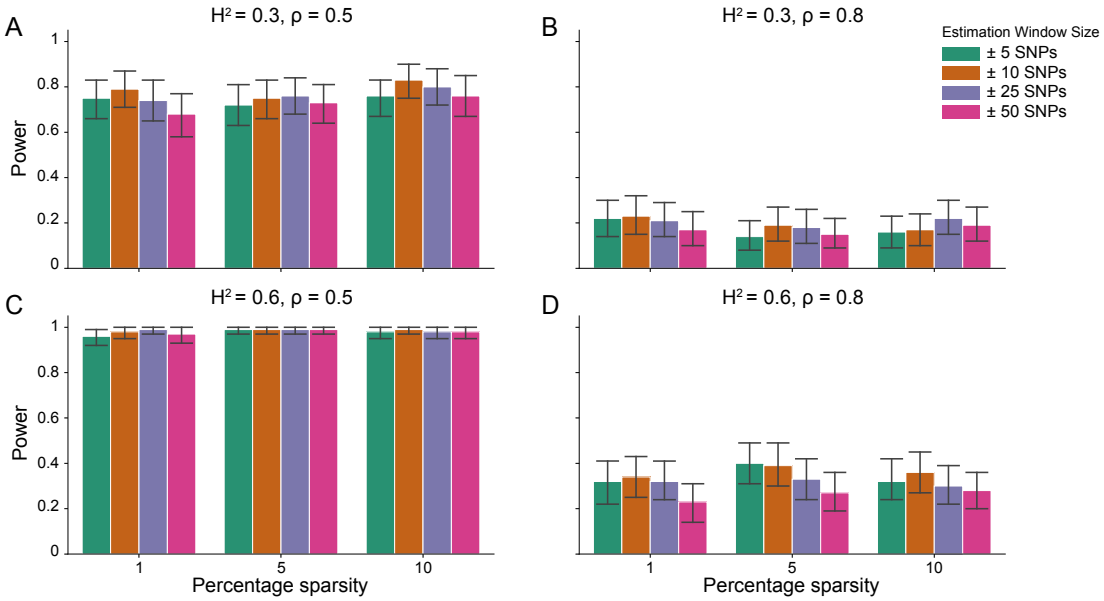

Figure S3. Power calculations for the MELD framework on simulated data using a ±100 kilobase (kb) window to generate pairwise interactions between causal SNPs and no minor allele frequency dependency $\alpha = 0$ for effect sizes. Synthetic trait architecture was simulated using real genotype data from individuals of self-identified European ancestry in the UK Biobank. All SNPs were considered to have at least an additive effect (i.e., creating a polygenic trait architecture). Next, we randomly select two groups of interacting variants and divide them into two interacting groups. The group #1 SNPs are chosen to be 1%, 5%, and 10% of the total number of SNPs genome-wide (see the x-axis in each panel). These interact with the group #2 SNPs which are selected to be variants within a ±100 kilobase (kb) window around each SNP in group #1. Coefficients for additive and interaction effects were simulated with no minor allele frequency dependency $\alpha = 0$ (see Materials and Methods). Panels **(A)** and **(B)** are results with simulations using a broad-sense heritability $H^2 = 0.3$, while panels **(C)** and **(D)** were generated with $H^2 = 0.6$. We also varied the proportion of broad-sense heritability contributed by additive effects to **(A, C)** $\rho = 0.5$ and **(B, D)** $\rho = 0.8$, respectively. Here, we are blind to the parameter settings used in generative model and run MELD while computing the marginal epistatic LD scores using different estimating windows of ±5 (green), ±10 (orange), ±25 (purple), and ±50 (pink) SNPs. Results are based on 100 simulations per parameter combination and the horizontal bars represent standard errors.

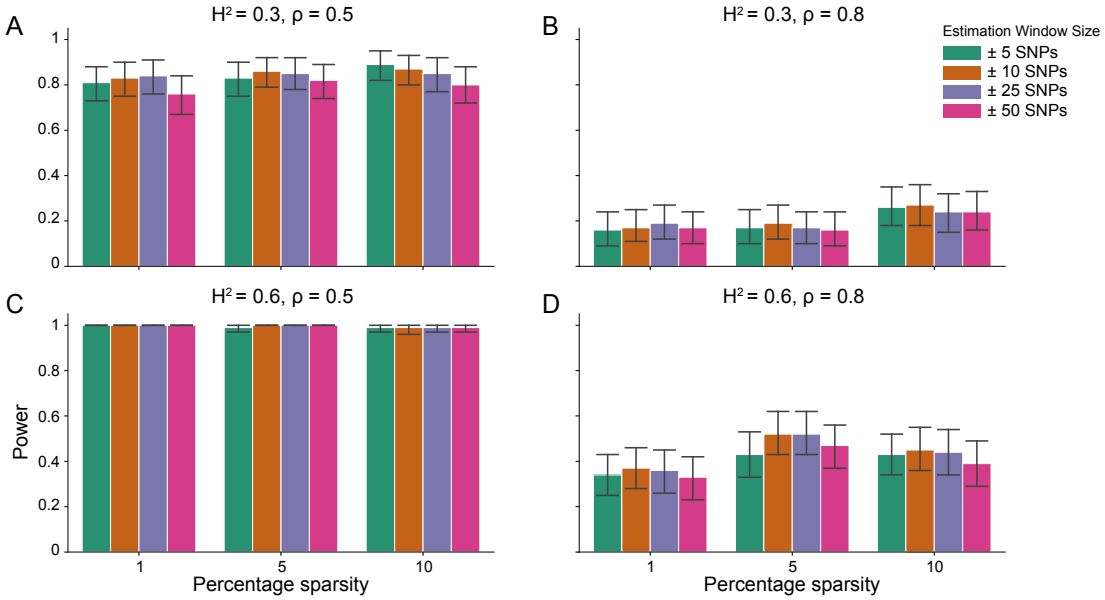

Figure S4. Power calculations for the `MELD` framework on simulated data using a ±100 kilobase (kb) window to generate pairwise interactions between causal SNPs and a moderate minor allele frequency dependency $\alpha = -0.5$ for effect sizes. Synthetic trait architecture was simulated using real genotype data from individuals of self-identified European ancestry in the UK Biobank. All SNPs were considered to have at least an additive effect (i.e., creating a polygenic trait architecture). Next, we randomly select two groups of interacting variants and divide them into two interacting groups. The group #1 SNPs are chosen to be 1%, 5%, and 10% of the total number of SNPs genome-wide (see the x-axis in each panel). These interact with the group #2 SNPs which are selected to be variants within a ±100 kilobase (kb) window around each SNP in group #1. Coefficients for additive and interaction effects were simulated with minor allele frequency dependency $\alpha = -0.5$ (see Materials and Methods). Panels **(A)** and **(B)** are results with simulations using a broad-sense heritability $H^2 = 0.3$, while panels **(C)** and **(D)** were generated with $H^2 = 0.6$. We also varied the proportion of broad-sense heritability contributed by additive effects to **(A, C)** $\rho = 0.5$ and **(B, D)** $\rho = 0.8$, respectively. Here, we are blind to the parameter settings used in generative model and run `MELD` while computing the marginal epistatic LD scores using different estimating windows of ±5 (green), ±10 (orange), ±25 (purple), and ±50 (pink) SNPs. Results are based on 100 simulations per parameter combination and the horizontal bars represent standard errors.

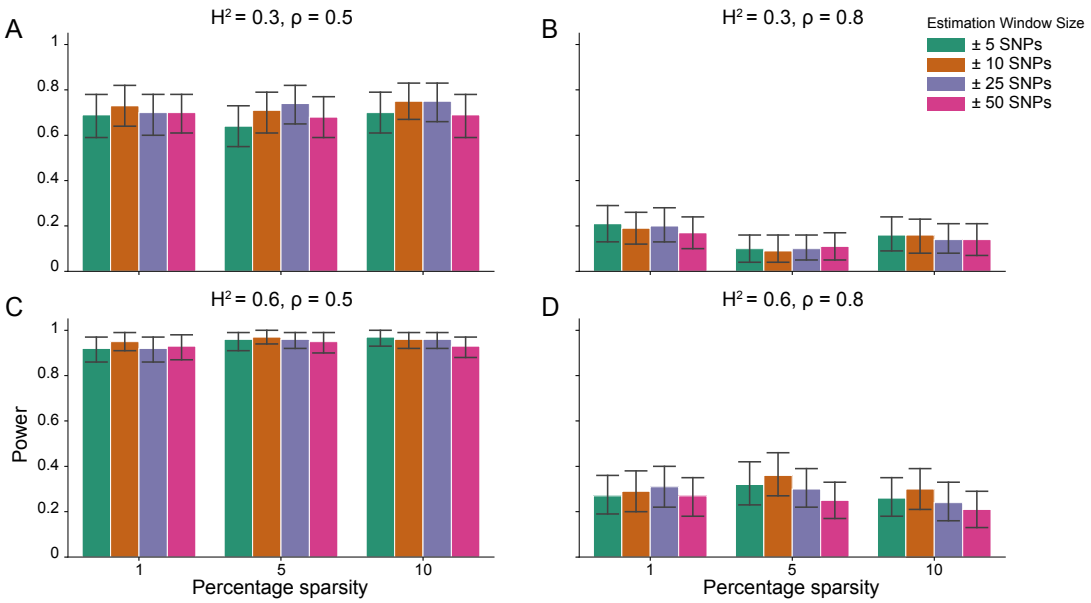

Figure S5. Power calculations for the `MELD` framework on simulated data using a ±100 kilobase (kb) window to generate pairwise interactions between causal SNPs and a strong minor allele frequency dependency $\alpha = -1$ for effect sizes. Synthetic trait architecture was simulated using real genotype data from individuals of self-identified European ancestry in the UK Biobank. All SNPs were considered to have at least an additive effect (i.e., creating a polygenic trait architecture). Next, we randomly select two groups of interacting variants and divide them into two interacting groups. The group #1 SNPs are chosen to be 1%, 5%, and 10% of the total number of SNPs genome-wide (see the x-axis in each panel). These interact with the group #2 SNPs which are selected to be variants within a ±100 kilobase (kb) window around each SNP in group #1. Coefficients for additive and interaction effects were simulated with minor allele frequency dependency $\alpha = -1$ (see Materials and Methods). Panels **(A)** and **(B)** are results with simulations using a broad-sense heritability $H^2 = 0.3$, while panels **(C)** and **(D)** were generated with $H^2 = 0.6$. We also varied the proportion of broad-sense heritability contributed by additive effects to **(A, C)** $\rho = 0.5$ and **(B, D)** $\rho = 0.8$, respectively. Here, we are blind to the parameter settings used in generative model and run `MELD` while computing the marginal epistatic LD scores using different estimating windows of ±5 (green), ±10 (orange), ±25 (purple), and ±50 (pink) SNPs. Results are based on 100 simulations per parameter combination and the horizontal bars represent standard errors.

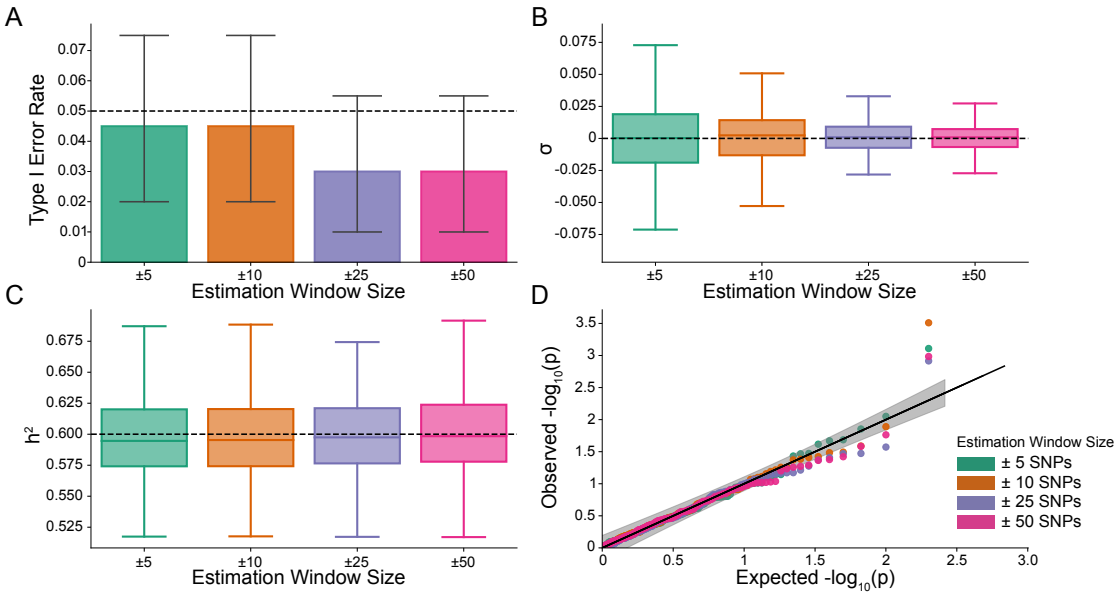

Figure S6. The MELD framework is well-calibrated and does not overestimate bias stemming from tagged epistasis when traits are generated by only additive effects and a moderate minor allele frequency dependency $\alpha = -0.5$ for effect sizes. In these simulations, synthetic trait architecture is made up of only additive genetic variation (i.e., $\rho = 1$). Coefficients for additive and interaction effects were simulated with minor allele frequency dependency $\alpha = -0.5$ (see Materials and Methods). Here, we are blind to the parameter settings used in generative model and run MELD while computing the marginal epistatic LD scores using different estimating windows of ± 5 (green), ± 10 (orange), ± 25 (purple), and ± 50 (pink) SNPs. **(A)** Mean type I error rate using the MELD framework across an array of estimation window sizes for the marginal epistatic scores. This is determined by assessing the P-value of the epistatic coefficient (σ) in the MELD regression model and checking whether $P < 0.05$. **(B)** Estimates of the epistatic coefficient (σ). Since traits were simulated with only additive effects, these estimates should be centered around zero. **(C)** Narrow-sense heritability (h^2) estimates where the true value is $H^2\rho = h^2 = 0.6$. **(D)** QQ-plot of the P-values for the epistatic coefficient (σ) in MELD. Results are based on 100 simulations per parameter combination and the horizontal bars represent standard errors.

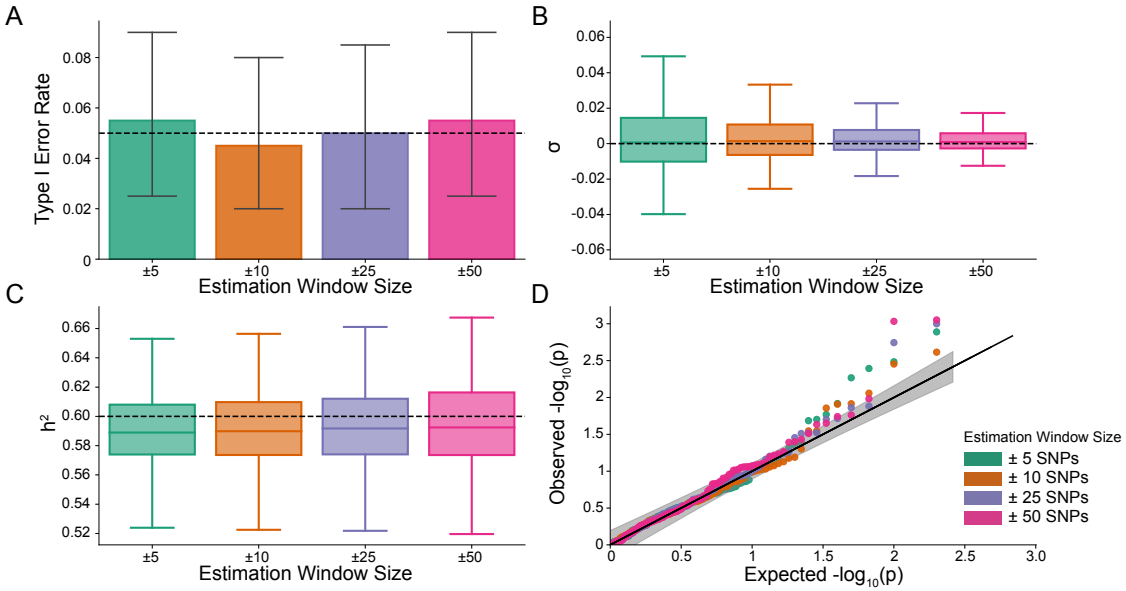

Figure S7. The MELD framework is well-calibrated and does not overestimate bias stemming from tagged epistasis when traits are generated by only additive effects and a strong minor allele frequency dependency $\alpha = -1$ for effect sizes. In these simulations, synthetic trait architecture is made up of only additive genetic variation (i.e., $\rho = 1$). Coefficients for additive and interaction effects were simulated with minor allele frequency dependency $\alpha = -1$ (see Materials and Methods). Here, we are blind to the parameter settings used in generative model and run MELD while computing the marginal epistatic LD scores using different estimating windows of ±5 (green), ±10 (orange), ±25 (purple), and ±50 (pink) SNPs. **(A)** Mean type I error rate using the MELD framework across an array of estimation window sizes for the marginal epistatic scores. This is determined by assessing the P-value of the epistatic coefficient (σ) in the MELD regression model and checking whether $P < 0.05$. **(B)** Estimates of the epistatic coefficient (σ). Since traits were simulated with only additive effects, these estimates should be centered around zero. **(C)** Narrow-sense heritability (h^2) estimates where the true value is $H^2\rho = h^2 = 0.6$. **(D)** QQ-plot of the P-values for the epistatic coefficient (σ) in MELD. Results are based on 100 simulations per parameter combination and the horizontal bars represent standard errors.

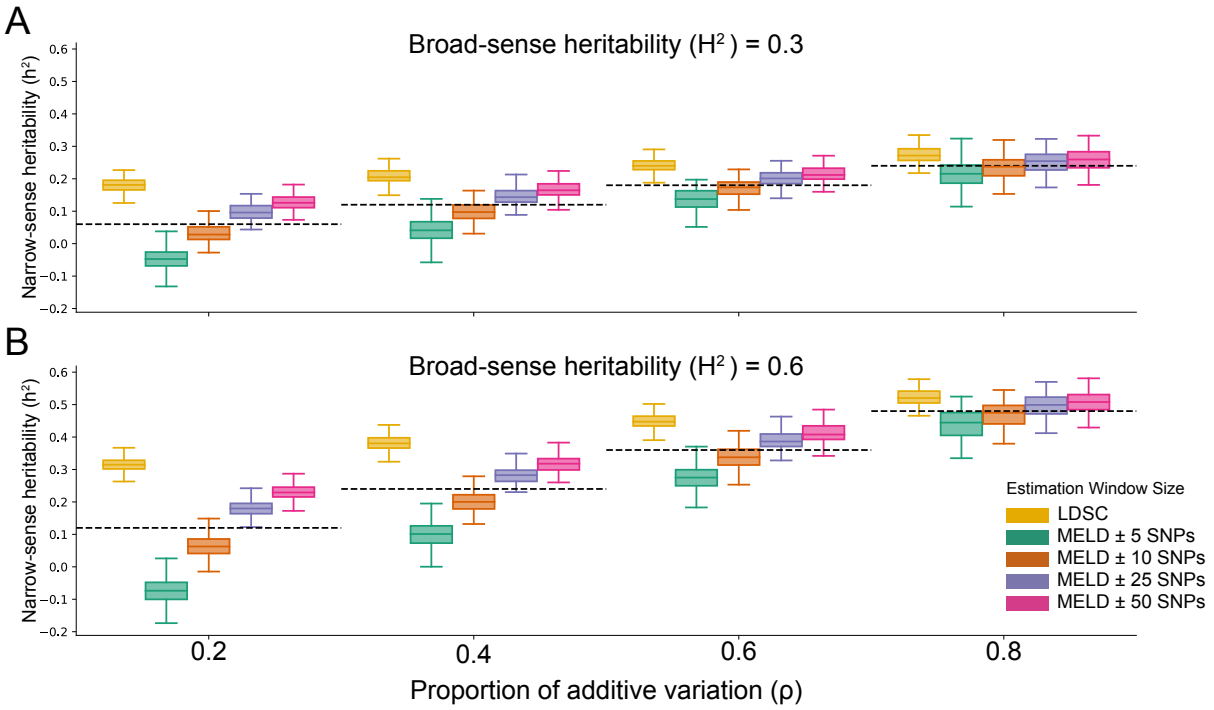

Figure S8. `MELD` robustly and accurately estimates narrow-sense heritability in simulations by controlling for epistatic bias in GWA summary statistics. Synthetic trait architecture was simulated using real genotype data from individuals of self-identified European ancestry in the UK Biobank. All SNPs were considered to have at least an additive effect (i.e., creating a polygenic trait architecture). Next, we randomly select two groups of interacting variants and divide them into two interacting groups. The group #1 SNPs are chosen to be 10% of the total number of SNPs genome-wide. These interact with the group #2 SNPs which are selected to be variants within a ±100 kilobase (kb) window around each SNP in group #1. Coefficients for additive and interaction effects were simulated with no minor allele frequency dependency $\alpha = 0$ (see Materials and Methods). Here, we assume a broad-sense heritability **(A)** $H^2 = 0.3$ or **(B)** $H^2 = 0.6$, and we vary the proportion contributed by additive effects with $\rho = \{0.2, 0.4, 0.6, 0.8\}$. The true narrow-sense heritability is set as $H^2\rho = h^2$. We run `MELD` while computing the marginal epistatic LD scores using different estimating windows of ±5, ±10, ±25, and ±50 SNPs, respectively. These results help motivate the model averaging strategy over the different estimation window sizes for the marginal epistatic LD scores in `MELD`. Results are based on 100 simulations per parameter combination.

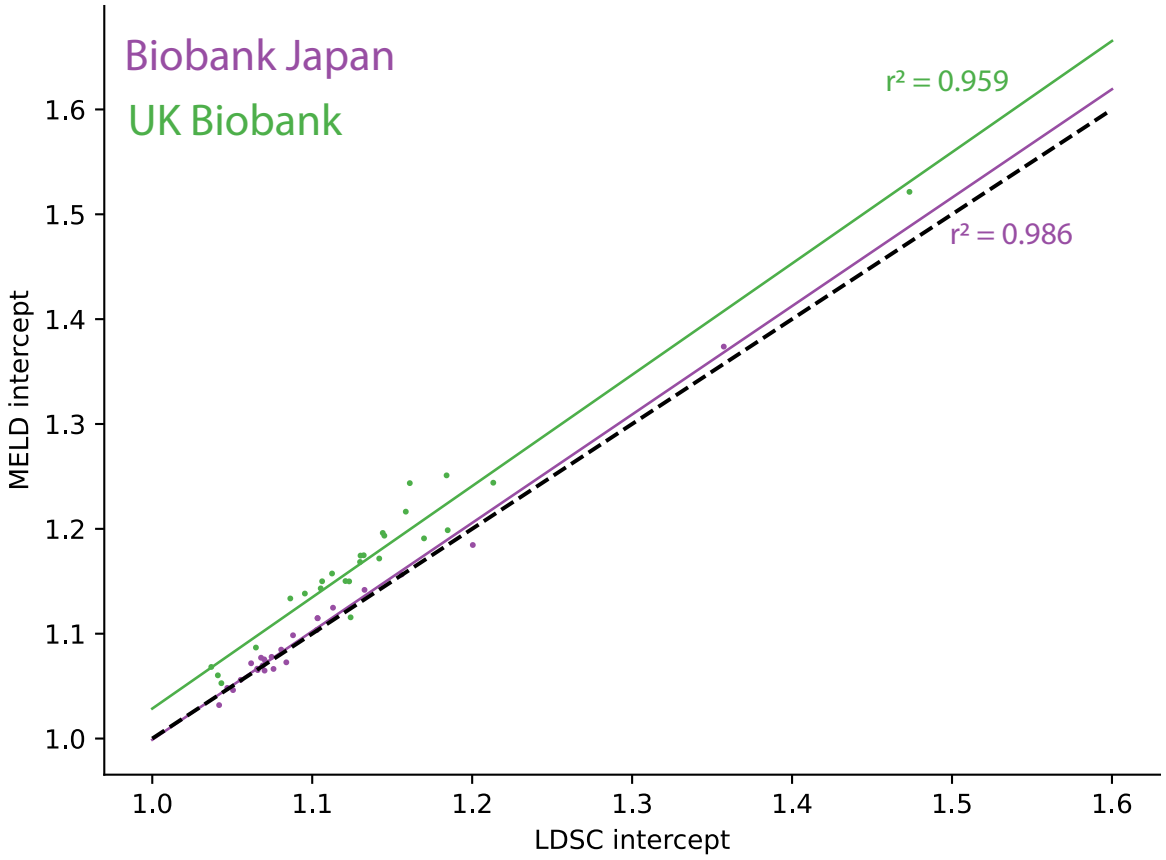

Figure S9. The `MELD` framework recovers narrow-sense heritability and provides estimates of bias in the UK Biobank and BioBank Japan. In both the UK Biobank (green) and BioBank Japan (purple), the intercepts for narrow-sense heritability estimation from `MELD` and `LDSC` are highly correlated for 25 different complex traits. Note that these intercept estimates represent the confounding factor due to uncontrolled effects. For `LDSC` this does include bias from pairwise genetic interactions, while `MELD` intercept estimates do not include bias due to these types of effects (i.e., they have been partitioned out). The Spearman correlation coefficients between estimates of the intercept for traits in the UK Biobank and BioBank Japan are $r^2 = 0.959$ and $r^2 = 0.986$, respectively. The dotted $x = y$ line represents points for when the two sets of estimates are equal.

Supplementary Tables

	LDSC		MELD (± 5 SNPs)	
True h^2	Estimated h^2	MAE	Estimated h^2	MAE
0.06	0.181 (0.002)	0.121 (0.002)	-0.047 (0.003)	0.107 (0.003)
0.12	0.208 (0.002)	0.086 (0.002)	0.043 (0.003)	0.078 (0.004)
0.18	0.241 (0.002)	0.061 (0.002)	0.135 (0.003)	0.046 (0.003)
0.24	0.272 (0.002)	0.036 (0.002)	0.214 (0.004)	0.037 (0.003)
	MELD (± 10 SNPs)		MELD (± 25 SNPs)	
True h^2	Estimated h^2	MAE	Estimated h^2	MAE
0.06	0.031 (0.003)	0.033 (0.002)	0.098 (0.002)	0.039 (0.002)
0.12	0.098 (0.003)	0.030 (0.002)	0.146 (0.003)	0.029 (0.002)
0.18	0.170 (0.003)	0.024 (0.002)	0.200 (0.002)	0.026 (0.002)
0.24	0.234 (0.003)	0.027 (0.002)	0.252 (0.003)	0.028 (0.002)
	MELD (± 50 SNPs)		MELD (Average)	
True h^2	Estimated h^2	MAE	Estimated h^2	MAE
0.06	0.128 (0.002)	0.068 (0.002)	0.052 (0.004)	0.062 (0.002)
0.12	0.167 (0.002)	0.047 (0.002)	0.113 (0.003)	0.046 (0.002)
0.18	0.214 (0.002)	0.035 (0.002)	0.180 (0.002)	0.033 (0.001)
0.24	0.259 (0.003)	0.030 (0.002)	0.240 (0.002)	0.030 (0.001)

Table S1. Comparison of LDSC and MELD estimates of narrow sense heritability when $H^2 = 0.3$. Synthetic trait architecture was simulated using real genotype data from individuals of self-identified European ancestry in the UK Biobank. All SNPs were considered to have at least an additive effect (i.e., creating a polygenic trait architecture). Next, we randomly select two groups of interacting variants and divide them into two interacting groups. The group #1 SNPs are chosen to be 10% of the total number of SNPs genome-wide. These interact with the group #2 SNPs which are selected to be variants within a ±100 kilobase (kb) window around each SNP in group #1. Coefficients for additive and interaction effects were simulated with no minor allele frequency dependency $\alpha = 0$ (see Materials and Methods). Here, we assume a broad-sense heritability $H^2 = 0.6$ and vary the proportion contributed by additive effects with $\rho = \{0.2, 0.4, 0.6, 0.8\}$. The true narrow-sense heritability is set as $H^2\rho = h^2$. We run MELD while computing the marginal epistatic LD scores using different estimating windows of ±5, ±10, ±25, and ±50 SNPs. The "average" column represents results using model averaging over the different estimating windows (see Materials and Methods). We report the mean estimates of h^2 (with standard errors in the parentheses) and use mean absolute error (MAE) to quantify the difference between the two methods. Results are based on 100 simulations per parameter combination. As shown in Figures 3 and S8, LDSC consistently overestimates narrow-sense heritability when there is non-additive trait variation.

	LDSC		MELD (± 5 SNPs)	
True h^2	Estimated h^2	MAE	Estimated h^2	MAE
0.12	0.315 (0.002)	0.194 (0.002)	-0.072 (0.004)	0.193 (0.004)
0.24	0.382 (0.002)	0.142 (0.002)	0.100 (0.004)	0.140 (0.004)
0.36	0.450 (0.003)	0.090 (0.003)	0.277 (0.004)	0.083 (0.004)
0.48	0.523 (0.003)	0.044 (0.002)	0.440 (0.004)	0.047 (0.004)
	MELD (± 10 SNPs)		MELD (± 25 SNPs)	
True h^2	Estimated h^2	MAE	Estimated h^2	MAE
0.12	0.064 (0.003)	0.058 (0.003)	0.180 (0.003)	0.060 (0.003)
0.24	0.200 (0.004)	0.045 (0.003)	0.283 (0.003)	0.045 (0.003)
0.36	0.340 (0.004)	0.045 (0.003)	0.393 (0.003)	0.036 (0.003)
0.48	0.471 (0.004)	0.031 (0.002)	0.498 (0.003)	0.030 (0.002)
	MELD (± 50 SNPs)		MELD (Averaged)	
True h^2	Estimated h^2	MAE	Estimated h^2	MAE
0.12	0.229 (0.002)	0.109 (0.002)	0.100 (0.006)	0.105 (0.003)
0.24	0.318 (0.003)	0.078 (0.003)	0.225 (0.005)	0.077 (0.003)
0.36	0.414 (0.003)	0.055 (0.003)	0.360 (0.003)	0.053 (0.002)
0.48	0.509 (0.003)	0.034 (0.002)	0.479 (0.002)	0.036 (0.001)

Table S2. Comparison of LDSC and MELD estimates of narrow sense heritability when $H^2 = 0.6$. Synthetic trait architecture was simulated using real genotype data from individuals of self-identified European ancestry in the UK Biobank. All SNPs were considered to have at least an additive effect (i.e., creating a polygenic trait architecture). Next, we randomly select two groups of interacting variants and divide them into two interacting groups. The group #1 SNPs are chosen to be 10% of the total number of SNPs genome-wide. These interact with the group #2 SNPs which are selected to be variants within a ±100 kilobase (kb) window around each SNP in group #1. Coefficients for additive and interaction effects were simulated with no minor allele frequency dependency $\alpha = 0$ (see Materials and Methods). Here, we assume a broad-sense heritability $H^2 = 0.6$ and vary the proportion contributed by additive effects with $\rho = \{0.2, 0.4, 0.6, 0.8\}$. The true narrow-sense heritability is set as $H^2\rho = h^2$. We run MELD while computing the marginal epistatic LD scores using different estimating windows of ±5, ±10, ±25, and ±50 SNPs. The "average" column represents results using model averaging over the different estimating windows (see Materials and Methods). We report the mean estimates of h^2 (with standard errors in the parentheses) and use mean absolute error (MAE) to quantify the difference between the two methods. Results are based on 100 simulations per parameter combination. As shown in Figures 3 and S8, LDSC consistently overestimates narrow-sense heritability when there is non-additive trait variation.

Trait Name	Code
Body mass index	BMI
High density lipoprotein	HDL
Low density lipoprotein	LDL
Hemoglobin A1c	HBA1C
Estimated glomerular filtration rate	EGFR
C-reactive protein	CRP
Systolic blood pressure	SBP
Diastolic blood pressure	DBP
Platelet count	PLC
Mean corpuscular hemoglobin concentration	MCHC
Mean corpuscular hemoglobin	MCH
Mean corpuscular volume	MCV
Red blood cell count	RBC
White blood cell count	WBC

Table S3. Abbreviations used throughout this study for 14 quantitative traits analyzed in this study. The remaining 11 traits analyzed were Basophil count, Cholesterol, Eosinophil count, Height, Hematocrit, Hemoglobin, Lymphocyte count, Monocyte count, Neutrophil count, and Triglyceride levels, respectively. These are not abbreviated in the main text.

Trait	$\boldsymbol{\alpha}$
Basophil count	-0.13
BMI*	-0.24
CRP	-0.39
Total cholesterol	-0.11
DBP*	-0.39
EGFR	-0.25
Eosinophil count*	-0.40
MCV*	-0.39
MCH*	-0.42
MCHC*	-0.42
Lymphocyte count*	-0.52
LDL	-0.20
Monocyte count*	-0.19
Neutrophil count	-0.09
Platelet count*	-0.19
HBA1C	-0.37
HDL	-0.41
Height*	-0.45
Hemoglobin	-0.37
Hematocrit	-0.37
RBC*	-0.39
SBP*	-0.38
Triglyceride	-0.07
Urate	-0.45
WBC*	-0.25

Table S4. Trait-specific α parameters for each of the 25 traits analyzed. Here, α values are used to weight each variant based on its minor allele frequency to account for frequency dependent architectures in each trait. The $*$ indicates α parameters that were taken directly from Schoech et al. [59]. The α parameters for other traits were calculated using the protocol used in that paper. Expansion of trait abbreviations are given in Table S3.

Trait Name or Code	Sample Size	Total SNPs	Citations
Basophil count	62,076	5,653,566	[80]
BMI	158,284	5,653,566	[81]
CRP	75,391	5,608,701	[80]
DBP	136,615	5,653,566	[80]
eGFR	143,658	5,608,701	[80]
Eosinophil count	62,076	5,653,566	[80]
HDL	70,657	5,608,701	[80]
Height	159,095	6,296,332	[82]
Hematocrit	108,757	5,653,566	[80]
Hemoglobin	108,769	5,653,566	[80]
HbA1c	75,391	5,608,701	[80]
LDL	72,866	5,608,701	[80]
Lymphocyte count	62,076	5,653,566	[80]
MCH	108,054	5,653,566	[80]
MCHC	108,738	5,653,566	[80]
MCV	108,526	5,653,566	[80]
Monocyte count	62,076	5,653,566	[80]
Neutrophil count	62,076	5,653,566	[80]
PLC	108,208	5,653,566	[80]
RBC	108,794	5,653,566	[80]
SBP	136,597	5,653,566	[80]
Cholesterol	128,305	5,608,701	[80]
Triglyceride	105,597	5,608,701	[80]
Urate	109,029	5,608,701	[80]
WBC	107,694	5,653,566	[80]

Table S5. Number of individuals and total SNPs included in the analysis of each trait in BioBank Japan.

Trait	UKB (LDSC)	UKB (MELD)	BBJ (LDSC)	BBJ (MELD)
BMI	0.506	0.282*	0.097	0.102
Basophil	0.076	0.044	0.062	0.058
CRP	0.098	0.057*	0.028	0.027
Cholesterol	0.307	0.165*	0.042	0.034
DBP	0.201	0.136*	0.038	0.039
EGFR	0.401	0.244*	0.067	0.070
Eosinophil	0.227	0.175*	0.049	0.047
HBA1C	0.208	0.147*	0.061	0.058
HDL	0.359	0.215*	0.135	0.067
Height	0.815	0.57*	0.226	0.234
Hematocrit	0.262	0.191*	0.036	0.036
Hemoglobin	0.287	0.207*	0.035	0.033
LDL	0.242	0.138*	0.045	0.029
Lymphocyte	0.075	0.07*	0.052	0.052
MCH	0.358	0.204*	0.096	0.069
MCHC	0.074	0.061*	0.038	0.035
MCV	0.408	0.243*	0.096	0.07
Monocyte	0.233	0.149*	0.059	0.059
Neutrophil	0.361	0.206*	0.074	0.067
Platelet	0.604	0.353*	0.11	0.096
RBC	0.369	0.256*	0.065	0.055
SBP	0.211	0.151*	0.047	0.05
Triglyceride	0.461	0.141	0.081	0.03*
Urate	0.294	0.184*	0.119	0.069
WBC	0.264	0.176*	0.067	0.065

Table S6. Comparison of LDSC and MELD estimates of narrow-sense heritability for 25 complex traits in the UK Biobank and BioBank Japan. MELD heritability estimates are corrected for bias from non-additive variation. Corrections can increase or decrease the total heritability estimates on a trait-by-trait basis. * denotes traits that have significant tagged epistasis as determined by the P-value for the σ coefficient in MELD. See Table 1 for trait-specific P-values. Note that 23 traits in the UK Biobank have a significant amount of uncorrected bias introduced by non-additive variance, while only one trait (Triglycerides) has significant bias in BioBank Japan. Note that these estimates are also displayed in the first two panels of Figure 4.

References

1. Ronald A Fisher. The correlation between relatives on the supposition of mendelian inheritance. Earth and Environmental Science Transactions of the Royal Society of Edinburgh, 52(2):399–433, 1919.

2. Sewall Wright. The analysis of variance and the correlations between relatives with respect to deviations from an optimum. Journal of Genetics, 30(2):243–256, 1935.

3. Noah Zaitlen, Peter Kraft, Nick Patterson, Bogdan Pasaniuc, Gaurav Bhatia, Samuela Pollack, and Alkes L. Price. Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. PLOS Genetics, 9(5):1–11, 2013. doi: 10.1371/journal.pgen.1003520. URL https://doi.org/10.1371/journal.pgen.1003520.

4. Tinca J C Polderman, Beben Benyamin, Christiaan A de Leeuw, Patrick F Sullivan, Arjen van Bochoven, Peter M Visscher, and Danielle Posthuma. Meta-analysis of the heritability of human traits based on fifty years of twin studies. Nature Genetics, 47(7):702–709, 2015. doi: 10.1038/ng.3285. URL https://doi.org/10.1038/ng.3285.

5. Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J Daly, Alkes L Price, and Benjamin M Neale. Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature genetics, 47(3):291–295, 2015.

6. Huwenbo Shi, Gleb Kichaev, and Bogdan Pasaniuc. Contrasting the genetic architecture of 30 complex traits from summary association data. The American Journal of Human Genetics, 99(1): 139–153, 2016.

7. Kangcheng Hou, Kathryn S. Burch, Arunabha Majumdar, Huwenbo Shi, Nicholas Mancuso, Yue Wu, Sriram Sankararaman, and Bogdan Pasaniuc. Accurate estimation of snp-heritability from biobank-scale data irrespective of genetic architecture. Nature Genetics, 51(8):1244–1251, 2019. doi: 10.1038/s41588-019-0465-0. URL https://doi.org/10.1038/s41588-019-0465-0.

8. Ali Pazokitoroudi, Yue Wu, Kathryn S. Burch, Kangcheng Hou, Aaron Zhou, Bogdan Pasaniuc, and Sriram Sankararaman. Efficient variance components analysis across millions of genomes.

673 Nature Communications, 11(1):4020, 2020. doi: 10.1038/s41467-020-17576-9. URL https://doi.
674 org/10.1038/s41467-020-17576-9.

675 9. Brendan Bulik-Sullivan, Hilary K Finucane, Verneri Anttila, Alexander Gusev, Felix R Day, Po-
676 Ru Loh, Laramie Duncan, John R B Perry, Nick Patterson, Elise B Robinson, Mark J Daly,
677 Alkes L Price, Benjamin M Neale, ReproGen Consortium, Psychiatric Genomics Consortium, and
678 Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3. An
679 atlas of genetic correlations across human diseases and traits. Nature Genetics, 47(11):1236–1241,
680 2015. doi: 10.1038/ng.3406. URL https://doi.org/10.1038/ng.3406.

681 10. Loic Yengo, Julia Sidorenko, Kathryn E Kemper, Zhili Zheng, Andrew R Wood, Michael N Weedon,
682 Timothy M Frayling, Joel Hirschhorn, Jian Yang, Peter M Visscher, et al. Meta-analysis of genome-
683 wide association studies for height and body mass index in 700000 individuals of european ancestry.
684 Human molecular genetics, 27(20):3641–3649, 2018.

685 11. Doug Speed and David J Balding. Sumher better estimates the snp heritability of complex traits
686 from summary statistics. Nature genetics, 51(2):277–284, 2019.

687 12. Shuang Song, Wei Jiang, Yiliang Zhang, Lin Hou, and Hongyu Zhao. Leveraging ld eigenvalue
688 regression to improve the estimation of snp heritability and confounding inflation. The American
689 Journal of Human Genetics, 2022.

690 13. Wei Cheng, Sohini Ramachandran, and Lorin Crawford. Estimation of non-null snp effect size
691 distributions enables the detection of enriched genes underlying complex traits. PLoS Genet, 16
692 (6):e1008855, 2020. URL https://doi.org/10.1371/journal.pgen.1008855.

693 14. Zheng Ning, Yudi Pawitan, and Xia Shen. High-definition likelihood inference of genetic cor-
694 relations across human complex traits. Nature Genetics, 52(8):859–864, 2020. doi: 10.1038/
695 s41588-020-0653-y. URL https://doi.org/10.1038/s41588-020-0653-y.

696 15. Yiliang Zhang, Qiongshi Lu, Yixuan Ye, Kunling Huang, Wei Liu, Yuchang Wu, Xiaoyuan Zhong,
697 Boyang Li, Zhaolong Yu, Brittany G Travers, Donna M Werling, James J Li, and Hongyu Zhao.
698 Supergnova: local genetic correlation analysis reveals heterogeneous etiologic sharing of complex
699 traits. Genome Biol, 22(1):262, 2021. ISSN 1474-760X (Electronic); 1474-7596 (Print); 1474-7596
700 (Linking). doi: 10.1186/s13059-021-02478-w.

16. William K Scott, Margaret A Pericak-Vance, and Jonathan L Haines. Genetic analysis of complex diseases. Science, 275(5304):1327–1330, 1997.

17. Anthony D Long and Charles H Langley. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. Genome Research, 9(8):720–731, 1999.

18. Jonathan Marchini, Peter Donnelly, and Lon R Cardon. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nature genetics, 37(4):413–417, 2005.

19. William G. Hill, Michael E. Goddard, and Peter M. Visscher. Data and theory point to mainly additive genetic variance for complex traits. PLOS Genetics, 4(2):1–10, 2008. doi: 10.1371/journal.pgen.1000008. URL https://doi.org/10.1371/journal.pgen.1000008.

20. Haifeng Shao, Lindsay C Burrage, David S Sinasac, Annie E Hill, Sheila R Ernest, William O'Brien, Hayden-William Courtland, Karl J Jepsen, Andrew Kirby, and EJ Kulbokas. Genetic architecture of complex traits: large phenotypic effects and pervasive epistasis. Proceedings of the National Academy of Sciences, 105(50):19910–19914, 2008.

21. James F Crow. On epistasis: why it is unimportant in polygenic directional selection. Philosophical Transactions of the Royal Society of London. Series B, Biological sciences, 365(1544):1241–1244, 2010. doi: 10.1098/rstb.2009.0275. URL https://pubmed.ncbi.nlm.nih.gov/20308099.

22. Or Zuk, Eliana Hechter, Shamil R. Sunyaev, and Eric S. Lander. The mystery of missing heritability: Genetic interactions create phantom heritability. Proceedings of the National Academy of Sciences, 109(4):1193–1198, 2012. doi: 10.1073/pnas.1119675109. URL https://www.pnas.org/doi/abs/10.1073/pnas.1119675109.

23. 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.

24. 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. The American Journal of Human Genetics, 108 (5):786–798, 2021.

25. Ali Pazokitoroudi, Alec M. Chiu, Kathryn S. Burch, Bogdan Pasaniuc, and Sriram Sankararaman. Quantifying the contribution of dominance deviation effects to complex trait variation in biobank-scale data. The American Journal of Human Genetics, 108(5):799–808, 2021. ISSN 0002-9297. doi: https://doi.org/10.1016/j.ajhg.2021.03.018. URL https://www.sciencedirect.com/science/article/pii/S0002929721001026.

26. Pierrick Wainschtein, Deepti Jain, Zhili Zheng, L Adrienne Cupples, Aladdin H Shadyab, Barbara McKnight, Benjamin M Shoemaker, Braxton D Mitchell, Bruce M Psaty, Charles Kooperberg, et al. Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. Nature Genetics, 54(3):263–273, 2022.

27. Andrew Anand Brown, Alfonso Buil, Ana Viñuela, Tuuli Lappalainen, Hou-Feng Zheng, J Brent Richards, Kerrin S Small, Timothy D Spector, Emmanouil T Dermitzakis, Richard Durbin, and Philipp Khaitovich. Genetic interactions affecting human gene expression identified by variance association mapping. eLife, 3:e01381, 2014. doi: 10.7554/eLife.01381. URL https://dx.doi.org/10.7554/eLife.01381.

28. Gang Fang, Wen Wang, Vanja Paunic, Hamed Heydari, Michael Costanzo, Xiaoye Liu, Xiaotong Liu, Benjamin VanderSluis, Benjamin Oately, Michael Steinbach, Brian Van Ness, Eric E. Schadt, Nathan D. Pankratz, Charles Boone, Vipin Kumar, and Chad L. Myers. Discovering genetic interactions bridging pathways in genome-wide association studies. Nature Communications, 10(1):4274, 2019. doi: 10.1038/s41467-019-12131-7. URL https://doi.org/10.1038/s41467-019-12131-7.

29. Joris van de Haar, Sander Canisius, Michael K. Yu, Emile E. Voest, Lodewyk F. A. Wessels, and Trey Ideker. Identifying epistasis in cancer genomes: A delicate affair. Cell, 177(6):1375–1383, 2019. doi: https://doi.org/10.1016/j.cell.2019.05.005. URL https://www.sciencedirect.com/science/article/pii/S0092867419305033.

30. 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. doi: 10.1073/pnas.1922305118. URL https://www.pnas.org/doi/abs/10.1073/pnas.1922305118.

31. Roshni A. Patel, Shaila A. Musharoff, Jeffrey P. Spence, Harold Pimentel, Catherine Tcheandjieu,

Hakhamanesh Mostafavi, Nasa Sinnott-Armstrong, Shoa L. Clarke, Courtney J. Smith, Peter P. Durda, Kent D. Taylor, Russell Tracy, Yongmei Liu, W. Craig Johnson, Francois Aguet, Kristin G. Ardlie, Stacey Gabriel, Josh Smith, Deborah A. Nickerson, Stephen S. Rich, Jerome I. Rotter, Philip S. Tsao, Themistocles L. Assimes, and Jonathan K. Pritchard. Genetic interactions drive heterogeneity in causal variant effect sizes for gene expression and complex traits. The American Journal of Human Genetics, 2022. doi: 10.1016/j.ajhg.2022.05.014. URL https://doi.org/10.1016/j.ajhg.2022.05.014.

32. Amy Hin Yan Tong, Guillaume Lesage, Gary D Bader, Huiming Ding, Hong Xu, Xiaofeng Xin, James Young, Gabriel F Berriz, Renee L Brost, Michael Chang, YiQun Chen, Xin Cheng, Gordon Chua, Helena Friesen, Debra S Goldberg, Jennifer Haynes, Christine Humphries, Grace He, Shamiza Hussein, Lizhu Ke, Nevan Krogan, Zhijian Li, Joshua N Levinson, Hong Lu, Patrice Ménard, Christella Munyana, Ainslie B Parsons, Owen Ryan, Raffi Tonikian, Tania Roberts, Anne-Marie Sdicu, Jesse Shapiro, Bilal Sheikh, Bernhard Suter, Sharyl L Wong, Lan V Zhang, Hongwei Zhu, Christopher G Burd, Sean Munro, Chris Sander, Jasper Rine, Jack Greenblatt, Matthias Peter, Anthony Bretscher, Graham Bell, Frederick P Roth, Grant W Brown, Brenda Andrews, Howard Bussey, and Charles Boone. Global mapping of the yeast genetic interaction network. Science, 303(5659):808–813, 2004. ISSN 1095-9203 (Electronic); 0036-8075 (Linking). doi: 10.1126/science.1091317.

33. Sean R. Collins, Maya Schuldiner, Nevan J. Krogan, and Jonathan S. Weissman. A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. Genome Biology, 7(7): R63, 2006. doi: 10.1186/gb-2006-7-7-r63. URL https://doi.org/10.1186/gb-2006-7-7-r63.

34. Jonathan Flint and Trudy F C Mackay. Genetic architecture of quantitative traits in mice, flies, and humans. Genome Res, 19(5):723–733, 2009. ISSN 1088-9051 (Print); 1549-5469 (Electronic); 1088-9051 (Linking). doi: 10.1101/gr.086660.108.

35. Xionglei He, Wenfeng Qian, Zhi Wang, Ying Li, and Jianzhi Zhang. Prevalent positive epistasis in escherichia coli and saccharomyces cerevisiae metabolic networks. Nature Genetics, 42(3):272–276, 2010. doi: 10.1038/ng.524. URL https://doi.org/10.1038/ng.524.

36. Balázs Szappanos, Károly Kovács, Béla Szamecz, Frantisek Honti, Michael Costanzo, Anastasia

Baryshnikova, Gabriel Gelius-Dietrich, Martin J Lercher, Márk Jelasity, Chad L Myers, Brenda J Andrews, Charles Boone, Stephen G Oliver, Csaba Pál, and Balázs Papp. An integrated approach to characterize genetic interaction networks in yeast metabolism. Nature Genetics, 43(7):656–662, 2011. doi: 10.1038/ng.846. URL https://doi.org/10.1038/ng.846.

37. Mats Pettersson, Francois Besnier, Paul B Siegel, and Orjan Carlborg. Replication and explorations of high-order epistasis using a large advanced intercross line pedigree. PLoS Genet, 7(7):e1002180, 2011. ISSN 1553-7404 (Electronic); 1553-7390 (Print); 1553-7390 (Linking). doi: 10.1371/journal.pgen.1002180.

38. Sudarshan Chari and Ian Dworkin. The conditional nature of genetic interactions: the consequences of wild-type backgrounds on mutational interactions in a genome-wide modifier screen. PLoS Genet, 9(8):e1003661, 2013. ISSN 1553-7404 (Electronic); 1553-7390 (Print); 1553-7390 (Linking). doi: 10.1371/journal.pgen.1003661.

39. Patrick J. Monnahan and John K. Kelly. Epistasis is a major determinant of the additive genetic variance in mimulus guttatus. PLOS Genetics, 11(5):e1005201, 2015. URL https://doi.org/10.1371/journal.pgen.1005201.

40. Elodie Marchadier, Mathieu Hanemian, Sebastien Tisne, Lien Bach, Christos Bazakos, Elodie Gilbault, Parham Haddadi, Laetitia Virlouvet, and Olivier Loudet. The complex genetic architecture of shoot growth natural variation in arabidopsis thaliana. PLoS genetics, 15(4):e1007954, 2019.

41. Gloria Yang, Dave W Anderson, Florian Baier, Elias Dohmen, Nansook Hong, Paul D Carr, Shina Caroline Lynn Kamerlin, Colin J Jackson, Erich Bornberg-Bauer, and Nobuhiko Tokuriki. Higher-order epistasis shapes the fitness landscape of a xenobiotic-degrading enzyme. Nature Chemical Biology, 15(11):1120–1128, 2019.

42. Patrik Nosil, Romain Villoutreix, Clarissa F. de Carvalho, Jeffrey L. Feder, Thomas L. Parchman, and Zach Gompert. Ecology shapes epistasis in a genotype–phenotype–fitness map for stick insect colour. Nature Ecology & Evolution, 4(12):1673–1684, 2020. doi: 10.1038/s41559-020-01305-y. URL https://doi.org/10.1038/s41559-020-01305-y.

811 43. Adam N Spierer, Jim A Mossman, Samuel Pattillo Smith, Lorin Crawford, Sohini Ramachandran,
812 and David M Rand. Natural variation in the regulation of neurodevelopmental genes modifies flight
813 performance in drosophila. Plos Genetics, 17(3):e1008887, 2021.

814 44. Evan E. Eichler, Jonathan Flint, Greg Gibson, Augustine Kong, Suzanne M. Leal, Jason H. Moore,
815 and Joseph H. Nadeau. Missing heritability and strategies for finding the underlying causes of
816 complex disease. Nature Reviews Genetics, 11(6):446–450, 2010. doi: 10.1038/nrg2809. URL
817 https://doi.org/10.1038/nrg2809.

818 45. Joshua S. Bloom, Ian M. Ehrenreich, Wesley T. Loo, Thúy-Lan Võ Lite, and Leonid Kruglyak.
819 Finding the sources of missing heritability in a yeast cross. Nature, 494(7436):234–237, 2013. doi:
820 10.1038/nature11867. URL https://doi.org/10.1038/nature11867.

821 46. Gibran Hemani, Sara Knott, and Chris Haley. An evolutionary perspective on epistasis and the
822 missing heritability. PLOS Genetics, 9(2):1–11, 2013. doi: 10.1371/journal.pgen.1003295. URL
823 https://doi.org/10.1371/journal.pgen.1003295.

824 47. Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Schizophre-
825 nia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J Daly,
826 Alkes L Price, and Benjamin M Neale. Ld score regression distinguishes confounding from
827 polygenicity in genome-wide association studies. Nat Genet, 47:291–295, 2015. URL http:
828 //dx.doi.org/10.1038/ng.3211.

829 48. Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt,
830 Pamela A Madden, Andrew C Heath, Nicholas G Martin, and Grant W Montgomery. Common
831 snps explain a large proportion of the heritability for human height. Nat Genet, 42(7):565, 2010.

832 49. Michael C Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. Rare-
833 variant association testing for sequencing data with the sequence kernel association test. Am J
834 Hum Genet, 89(1):82–93, 2011. ISSN 1537-6605 (Electronic); 0002-9297 (Linking). doi: 10.1016/
835 j.ajhg.2011.05.029.

836 50. Xiang Zhou, Peter Carbonetto, and Matthew Stephens. Polygenic modeling with Bayesian sparse
837 linear mixed models. PLoS Genet, 9(2):e1003264, 2013.

838 51. Lorin Crawford, Ping Zeng, Sayan Mukherjee, and Xiang Zhou. Detecting epistasis with the
839 marginal epistasis test in genetic mapping studies of quantitative traits. PLoS Genet, 13(7):
840 e1006869, 2017. URL https://doi.org/10.1371/journal.pgen.1006869.

841 52. Farhad Hormozdiari, Emrah Kostem, Eun Yong Kang, Bogdan Pasaniuc, and Eleazar Eskin. Iden-
842 tifying causal variants at loci with multiple signals of association. Genetics, 198(2):497–508, 2014.
843 doi: 10.1534/genetics.114.167908. URL https://pubmed.ncbi.nlm.nih.gov/25104515.

844 53. Farhad Hormozdiari, Martijn van de Bunt, Ayellet V. Segrè, Xiao Li, Jong Wha J. Joo, Michael
845 Bilow, Jae Hoon Sul, Sriram Sankararaman, Bogdan Pasaniuc, and Eleazar Eskin. Colocalization
846 of GWAS and eQTL signals detects target genes. Am J Hum Genet, 99(6):1245–1260, 2016. doi:
847 10.1016/j.ajhg.2016.10.003. URL https://doi.org/10.1016/j.ajhg.2016.10.003.

848 54. Xiang Zhu and Matthew Stephens. Bayesian large-scale multiple regression with summary statistics
849 from genome-wide association studies. Ann Appl Stat, 11(3):1561–1592, 2017. doi: 10.1214/
850 17-AOAS1046. URL https://projecteuclid.org:443/euclid.aoas/1507168840.

851 55. Yan Zhang, Guanghao Qi, Ju-Hyun Park, and Nilanjan Chatterjee. Estimation of complex effect-
852 size distributions using summary-level statistics from genome-wide association studies across 32
853 complex traits. Nat Genet, 50(9):1318–1326, 2018.

854 56. Xiang Zhu and Matthew Stephens. Large-scale genome-wide enrichment analyses identify new
855 trait-associated genes and pathways across 31 human phenotypes. Nat Comm, 9(1):4361, 2018.

856 57. Pinar Demetci, Wei Cheng, Gregory Darnell, Xiang Zhou, Sohini Ramachandran, and Lorin
857 Crawford. Multi-scale inference of genetic trait architecture using biologically annotated neu-
858 ral networks. PLOS Genetics, 17(8):1–53, 2021. doi: 10.1371/journal.pgen.1009754. URL
859 https://doi.org/10.1371/journal.pgen.1009754.

860 58. Yongtao Guan and Matthew Stephens. Bayesian variable selection regression for genome-wide
861 association studies and other large-scale problems. Ann Appl Stat, 5(3):1780–1815, 2011. doi:
862 10.1214/11-AOAS455. URL https://projecteuclid.org:443/euclid.aoas/1318514285.

863 59. Armin P Schoech, Daniel M Jordan, Po-Ru Loh, Steven Gazal, Luke J O'Connor, Daniel J Bal-
864 ick, Pier F Palamara, Hilary K Finucane, Shamil R Sunyaev, and Alkes L Price. Quantification

of frequency-dependent genetic architectures in 25 uk biobank traits reveals action of negative selection. Nature communications, 10(1):1–10, 2019.

60. Yabo Li, Hyosuk Cho, Fan Wang, Oriol Canela-Xandri, Chunyan Luo, Konrad Rawlik, Stephen Archacki, Chengqi Xu, Albert Tenesa, Qiuyun Chen, et al. Statistical and functional studies identify epistasis of cardiovascular risk genomic variants from genome-wide association studies. Journal of the American Heart Association, 9(7):e014146, 2020.

61. Steven Gazal, Hilary K Finucane, Nicholas A Furlotte, Po-Ru Loh, Pier Francesco Palamara, Xuanyao Liu, Armin Schoech, Brendan Bulik-Sullivan, Benjamin M Neale, Alexander Gusev, et al. Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. Nature genetics, 49(10):1421–1427, 2017.

62. Steven Gazal, Omer Weissbrod, Farhad Hormozdiari, Kushal K. Dey, Joseph Nasser, Karthik A. Jagadeesh, Daniel J. Weiner, Huwenbo Shi, Charles P. Fulco, Luke J. O'Connor, Bogdan Pasaniuc, Jesse M. Engreitz, and Alkes L. Price. Combining snp-to-gene linking strategies to identify disease genes and assess disease omnigenicity. Nature Genetics, 54(6):827–836, 2022. doi: 10.1038/s41588-022-01087-y. URL `https://doi.org/10.1038/s41588-022-01087-y`.

63. Kushal K. Dey, Steven Gazal, Bryce van de Geijn, Samuel Sungil Kim, Joseph Nasser, Jesse M. Engreitz, and Alkes L. Price. Snp-to-gene linking strategies reveal contributions of enhancer-related and candidate master-regulator genes to autoimmune disease. Cell Genomics, 2(7):100145, 2022. doi: https://doi.org/10.1016/j.xgen.2022.100145. URL `https://www.sciencedirect.com/science/article/pii/S2666979X22000787`.

64. Hilary K Finucane, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verneri Anttila, Han Xu, Chongzhi Zang, Kyle Farh, Stephan Ripke, Felix R Day, Shaun Purcell, Eli Stahl, Sara Lindstrom, John R B Perry, Yukinori Okada, Soumya Raychaudhuri, Mark J Daly, Nick Patterson, Benjamin M Neale, and Alkes L Price. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet, 47(11): 1228–1235, 2015. ISSN 1546-1718 (Electronic); 1061-4036 (Print); 1061-4036 (Linking). doi: 10.1038/ng.3404.

892 65. Rachel Moore, Francesco Paolo Casale, Marc Jan Bonder, Danilo Horta, Lude Franke, Inês Bar-
893 roso, and Oliver Stegle. A linear mixed-model approach to study multivariate gene-environment
894 interactions. Nat Genet, 51(1):180–186, 2019. ISSN 1546-1718 (Electronic); 1061-4036 (Print);
895 1061-4036 (Linking). doi: 10.1038/s41588-018-0271-0.

896 66. Matthew Kerin and Jonathan Marchini. A non-linear regression method for estimation of gene–
897 environment heritability. Bioinformatics, 36(24):5632–5639, 2020. ISSN 1367-4803. doi: 10.1093/
898 bioinformatics/btaa1079. URL https://doi.org/10.1093/bioinformatics/btaa1079.

899 67. Carrie Zhu, Matthew J. Ming, Jared M. Cole, Mark Kirkpatrick, and Arbel Harpak. Amplifi-
900 cation is the primary mode of gene-by-sex interaction in complex human traits. bioRxiv, page
901 2022.05.06.490973, 2022. doi: 10.1101/2022.05.06.490973. URL http://biorxiv.org/content/
902 early/2022/05/08/2022.05.06.490973.abstract.

903 68. Daniel Runcie, Hao Cheng, and Lorin Crawford. Mega-scale linear mixed models for ge-
904 nomic predictions with thousands of traits. bioRxiv, page 2020.05.26.116814, 2020. doi:
905 10.1101/2020.05.26.116814. URL http://biorxiv.org/content/early/2020/05/29/2020.05.
906 26.116814.abstract.

907 69. Sahin Naqvi, Yoeri Sleyp, Hanne Hoskens, Karlijne Indencleef, Jeffrey P. Spence, Rose Bruffaerts,
908 Ahmed Radwan, Ryan J. Eller, Stephen Richmond, Mark D. Shriver, John R. Shaffer, Seth M.
909 Weinberg, Susan Walsh, James Thompson, Jonathan K. Pritchard, Stefan Sunaert, Hilde Peeters,
910 Joanna Wysocka, and Peter Claes. Shared heritability of human face and brain shape. Nature
911 Genetics, 53(6):830–839, 2021. doi: 10.1038/s41588-021-00827-w. URL https://doi.org/10.
912 1038/s41588-021-00827-w.

913 70. Yong Jiang and Jochen C. Reif. Modeling epistasis in genomic selection. Genetics, 201:759–768,
914 2015.

915 71. L. Isserlis. On a formula for the product-moment coefficient of any order of a normal frequency
916 distribution in any number of variables. Biometrika, 12(1-2):134–139, 1918. ISSN 0006-3444. doi:
917 10.1093/biomet/12.1-2.134. URL https://doi.org/10.1093/biomet/12.1-2.134.

918 72. Peter Carbonetto and Matthew Stephens. Scalable variational inference for bayesian variable

selection in regression, and its accuracy in genetic association studies. Bayesian Anal, 7(1):73–108, 2012.

73. Jennifer A. Hoeting, David Madigan, Adrian E. Raftery, and Chris T. Volinsky. Bayesian model averaging: a tutorial (with comments by m. clyde, david draper and e. i. george, and a rejoinder by the authors. Statist Sci, 14(4):382–417, 1999. doi: 10.1214/ss/1009212519. URL https://projecteuclid.org:443/euclid.ss/1009212519.

74. Shadi Zabad, Aaron P Ragsdale, Rosie Sun, Yue Li, and Simon Gravel. Assumptions about frequency-dependent architectures of complex traits bias measures of functional enrichment. Genetic epidemiology, 45(6):621–632, 2021.

75. Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. Gcta: a tool for genome-wide complex trait analysis. Am J Hum Genet, 88(1):76–82, 2011. ISSN 1537-6605 (Electronic); 0002-9297 (Print); 0002-9297 (Linking). doi: 10.1016/j.ajhg.2010.11.011.

76. Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Alan Young, Mark Effingham, Gil McVean, Stephen Leslie, Naomi Allen, Peter Donnelly, and Jonathan Marchini. The uk biobank resource with deep phenotyping and genomic data. Nature, 562(7726):203–209, 2018. doi: 10.1038/s41586-018-0579-z. URL https://doi.org/10.1038/s41586-018-0579-z.

77. Christopher C Chang, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. Second-generation plink: rising to the challenge of larger and richer datasets. Gigascience, 4(1):s13742–015, 2015.

78. Mashaal Sohail, Robert M Maier, Andrea Ganna, Alex Bloemendal, Alicia R Martin, Michael C Turchin, Charleston WK Chiang, Joel Hirschhorn, Mark J Daly, Nick Patterson, et al. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. Elife, 8:e39702, 2019.

79. Gad Abraham, Yixuan Qiu, and Michael Inouye. Flashpca2: principal component analysis of biobank-scale genotype datasets. Bioinformatics, 2017.

946 80. Masahiro Kanai, Masato Akiyama, Atsushi Takahashi, Nana Matoba, Yukihide Momozawa,
947 Masashi Ikeda, Nakao Iwata, Shiro Ikegawa, Makoto Hirata, Koichi Matsuda, et al. Genetic anal-
948 ysis of quantitative traits in the japanese population links cell types to complex human diseases.
949 Nature genetics, 50(3):390–400, 2018.

950 81. Masato Akiyama, Yukinori Okada, Masahiro Kanai, Atsushi Takahashi, Yukihide Momozawa,
951 Masashi Ikeda, Nakao Iwata, Shiro Ikegawa, Makoto Hirata, Koichi Matsuda, et al. Genome-wide
952 association study identifies 112 new loci for body mass index in the japanese population. Nature
953 genetics, 49(10):1458, 2017.

954 82. Masato Akiyama, Kazuyoshi Ishigaki, Saori Sakaue, Yukihide Momozawa, Momoko Horikoshi,
955 Makoto Hirata, Koichi Matsuda, Shiro Ikegawa, Atsushi Takahashi, Masahiro Kanai, et al. Char-
956 acterizing rare and low-frequency height-associated variants in the japanese population. Nature
957 communications, 10(1):1–11, 2019.