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

Gregory Darnell\textsuperscript{1,2,*}, Samuel Pattillo Smith\textsuperscript{1,3,*}, Dana Udwin\textsuperscript{4}, Sohini Ramachandran\textsuperscript{1,3,5,§}, and Lorin Crawford\textsuperscript{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. Until recently, studies of genetic heritability in humans have been reliant on typically small sized family studies with known relatedness structure between individuals. 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. 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. 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. 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. The most widely used of these approaches is
linkage disequilibrium (LD) score regression and the corresponding LDSC software, 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 or because they have a trans-interaction effect with variants located within a gene enriched for associations with the trait of interest. 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. 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.

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. However, despite these controversies, many association mapping studies in humans have identified candidates of epistasis that notably contribute to trait variation, and some have recently shown that gene-by-gene interactions can drive heterogeneity of causal variant effect sizes across diverse human populations. Epistasis is a well-known contributor to trait architecture in several model organisms. 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. 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.”

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 $y$ 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 $\hat{\beta}$ and an empirical LD matrix $R$ (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

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

(1)

where $\beta = (\beta_1, \ldots, \beta_J)$ is a $J$-dimensional vector containing the true additive effect sizes for an additional copy of the reference allele at each locus on $y$; $W$ 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 $x_j \odot x_k$ for the $j$-th and $k$-th variants; $\theta = (\theta_1, \ldots, \theta_M)$ is an $M$-dimensional vector containing the interaction effect sizes; $\varepsilon$ 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 $H^2$. $I$ 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$^{47-51}$

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

(2)

where $\rho$ measures the proportion of total genetic effects that is contributed by the additive effects. Effectively, we say $\mathbb{V}[X\beta] = H^2 \rho = h^2$ is the narrow-sense heritability for a trait, while $\mathbb{V}[W\theta] = H^2 (1 - \rho)$ 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 $X$ and $W$ do not share the same column space (i.e., such that $X^T W = 0$). However, if we relax these assumptions, then the following relationship between the moment matrix $X^T y$, the observed marginal GWA summary statistics $\hat{\beta}$, and

\[
X^T y \sim \mathcal{N}(X^T \hat{\beta}, H^2 (1 - \rho) \mathbb{I}).
\]
the true coefficient values $\beta$ holds in expectation (see Materials and Methods)

$$X^\top y = (X^\top X)\beta + (X^\top W)\theta \iff \hat{\beta} = R\beta + V\theta$$ (3)

where $R$ is an empirical estimate of the LD matrix and $V$ represents an empirical estimate of the correlation between the individual-level genotypes $X$ and the span of genetic interactions between causal SNPs in $W$. Intuitively, the term $V\theta$ can be interpreted as “bias” in the additive effect estimate that stems 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., $E[\hat{\beta}] - \beta \neq 0$). Note that when either conditions (i) or (ii) are indeed met such that $V\theta = 0$, the equation above simplifies to a relationship between LD and summary statistics that is assumed in many common GWA studies\textsuperscript{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\textsuperscript{47}. Here, we note that, according to Eq. (3), $\hat{\beta} \sim \mathcal{N}(R\beta + V\theta, \lambda R)$ where $\lambda$ is a misestimation factor (i.e., inflation or deflation) due to uncontrolled confounding effects\textsuperscript{12,58}.

Next, we condition on $\Theta = (\beta, \theta)$ and take the expectation of chi-square statistics $\chi^2 = N\hat{\beta}\hat{\beta}^\top$ to yield

$$E[\hat{\beta}\hat{\beta}^\top] = E\left[E\left[\hat{\beta}\hat{\beta}^\top | \Theta \right]\right] = E\left[V E\left[\hat{\beta} | \Theta \right] + E\left[\hat{\beta} | \Theta \right] E\left[\hat{\beta} | \Theta \right]^\top \right]$$

$$= E\left[\lambda R + (R\beta + V\theta)(R\beta + V\theta)^\top \right]$$

$$= E\left[\lambda R + R\beta\beta^\top R + 2R\beta\theta^\top V + V\theta\theta^\top V \right]$$

$$= \lambda R + \left(\frac{H^2 \rho}{J}\right) R^2 + \left(\frac{H^2 (1-\rho)}{M}\right) V^2.$$ (4)

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

$$E[\chi^2] \propto 1 + \ell \tau + f \sigma$$ (5)
where $\chi^2 = (\chi_1^2, \ldots, \chi_J^2)$ is a $J$-dimensional vector of chi-square summary statistics, and $\ell = (\ell_1, \ldots, \ell_J)$ and $f = (f_1, \ldots, f_J)$ 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 $1$ 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 ($\ell$) and marginal epistatic ($f$) 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 $\tau$. Instead, the inclusion of $f$ 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 $\sigma$. In other words, we can interpret $\sigma$ 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; 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: ±10 and ±100 kb.

We also varied the correlation between SNP effect size and minor allele frequency (MAF) (as discussed in Schoech et al.). 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 \textit{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, \textit{LDSC} routinely overestimates the truth in our simulations that include nonzero epistatic effects. In contrast, \textit{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 \textit{MELD} and \textit{LDSC} are shown in Table S1 and S2, respectively. Generally, the error in narrow-sense heritability estimates is higher for \textit{LDSC} than it is for \textit{MELD} across each of the scenarios that we consider.

**Application of \textit{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 \textit{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 \textit{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.\textsuperscript{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\textsuperscript{12}. Generally, we find that the traditional \textit{LDSC} underestimates trait narrow-sense heritability when it does not consider this additional source of genetic signal as opposed to \textit{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 $\sigma$; 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. The inclusion of additional annotations has been shown to provide more refined narrow-sense heritability estimates from GWA summary statistics while accounting for linkage. 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. 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. 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.
URLs


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

\[ y = X\beta + W\theta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, (1 - H^2)I), \]  

where \( y \) denotes an \( N \)-dimensional vector of phenotypic states for a quantitative trait of interest measured in \( N \) individuals; \( X \) 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; \( \beta = (\beta_1, \ldots, \beta_J) \) is a \( J \)-dimensional vector containing the true additive effect sizes for an additional copy of the reference allele at each locus on \( y \); \( W \) 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 \( x_j \odot x_k \) for the \( j \)-th and \( k \)-th variants; \( \theta = (\theta_1, \ldots, \theta_M) \) is an \( M \)-dimensional vector containing the interaction effect sizes; \( \varepsilon \) 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 \( I \) 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 \( \beta \) and \( \theta \) 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 \( W \).
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., $E[\hat{\beta}] - \beta \neq 0$). We now formally derive this concept. Throughout this section, we will use $X^\top X/N$ to denote the linkage disequilibrium (LD) or pairwise correlation matrix between SNPs. We will then let $R$ 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

$$E[X^\top X] \approx NR, \quad E[x_j^\top x_j] \approx N, \quad E[x_j^\top x_k] \approx Nr_{jk} \quad (7)$$

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

A central goal in GWA studies is to jointly infer the true additive effects $\beta = (X^\top X)^{-1}X^\top y$ 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 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

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

where $\rho$ measures the proportion of total genetic effects that is contributed by the additive effects. Alternatively, we say that $\nabla[X\beta] = H^2\rho = h^2$ 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 $\nabla[W\theta] = H^2(1 - \rho)$ 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 \( \theta = 0 \))\(^{23,24,26} \); therefore, summary statistics of the true additive effects \( \beta \) in Eq. (6) are typically derived by computing a marginal least squares estimate with the observed data

\[
\hat{\beta}_j = (x_j^\top x_j)^{-1} x_j^\top y \iff \hat{\beta} = \text{diag}(X^\top X)^{-1} X^\top y. \tag{9}
\]

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 \( X^\top y \) and the additive effect size estimates \( \hat{\beta} \):

\[
X^\top y = \text{diag}(X^\top X) \hat{\beta} \approx N \hat{\beta}. \tag{10}
\]

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

\[
\beta = (X^\top X)^{-1} X^\top y \approx (NR)^{-1} N \hat{\beta} = R^{-1} \hat{\beta}. \tag{11}
\]

After some algebra, the above mirrors a high-dimensional regression model where \( \hat{\beta} = R\beta \) 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 \( \theta \neq 0 \)), 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 \( \beta \) and \( \theta \) from the generative model in Eq. (6)

\[
\begin{bmatrix}
\beta \\
\theta
\end{bmatrix} = \begin{bmatrix}
X^\top X & X^\top W \\
W^\top X & W^\top W
\end{bmatrix}^{-1} \begin{bmatrix}
X^\top \\
W^\top
\end{bmatrix} y, \tag{12}
\]
where the matrix $X^\top W$ 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 $X^\top y$, the observed marginal GWA summary statistics $\hat{\beta}$, and the true coefficient values $\beta$ where

$$X^\top y = (X^\top X)\beta + (X^\top W)\theta \quad \overset{\sim}{\Rightarrow} \quad \hat{\beta} = R\beta + V\theta.$$  \tag{13}

Here, we define $V$ to represent an empirical estimate of the correlation between the individual-level genotypes and the non-additive genetic interaction matrix such that $E[X^\top W] \approx NV$. Similar to the LD matrix $R$, the correlation matrix $V$ is also assumed to be computed from reference panel data. Intuitively, when $V\theta \neq 0$ there is additional bias in the effect size estimates, and when $V\theta = 0$ 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)\textsuperscript{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\textsuperscript{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 $\chi_j^2 = N\hat{\beta}_j^2$ for every $j = 1, \ldots, J$ SNP in the data. Taking the expectation of these chi-square statistics yields

$$E[\chi_j^2] = N E[\hat{\beta}_j^2] = N \left[ V[\hat{\beta}_j] + \left(E[\hat{\beta}_j]\right)^2 \right].$$  \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) $E[\hat{\beta}_j] = 0$ for all variants. Second, by conditioning on the generative model from
Eq. (6), we can use the law of total variance to simplify \( \mathbb{V}[^{\hat{\beta}}_{j}] \) where

\[
\mathbb{V}[^{\hat{\beta}}_{j}] = \mathbb{E}[\mathbb{V}[^{\hat{\beta}}_{j} | \mathbf{X}]] + \mathbb{V}[\mathbb{E}[^{\hat{\beta}}_{j} | \mathbf{X}]] \approx \mathbb{E}[\mathbb{V}[^{\hat{\beta}}_{j} \mathbf{y} / N | \mathbf{X}]] + 0
\]

\[
= \mathbb{E} \left[ \frac{1}{N^2} \mathbf{x}_{j}^\top \{ \mathbb{V}[\mathbf{y} | \mathbf{X}] \} \mathbf{x}_{j} \right]
\]

\[
= \mathbb{E} \left[ \frac{1}{N^2} \mathbf{x}_{j}^\top \left\{ \frac{H^2 \rho}{J} \mathbf{X}\mathbf{x}_{j}^\top + \frac{H^2 (1 - \rho)}{M} \mathbf{W}\mathbf{W}^\top +(1-H^2) \right\} \mathbf{x}_{j} \right]
\]

\[
= \mathbb{E} \left[ \frac{1}{N^2} \left\{ \frac{H^2 \rho}{J} \mathbf{x}_{j}^\top \mathbf{X}\mathbf{x}_{j}^\top + \frac{H^2 (1 - \rho)}{M} \mathbf{x}_{j}^\top \mathbf{W}\mathbf{W}^\top \mathbf{x}_{j} + N(1-H^2) \right\} \right].
\]

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

\[
\frac{1}{N^2} \mathbf{x}_{j}^\top \mathbf{X}\mathbf{x}_{j}^\top = \sum_{k=1}^{J} \tilde{r}_{jk}^2,
\]

\[
\frac{1}{N^2} \mathbf{x}_{j}^\top \mathbf{W}\mathbf{W}^\top \mathbf{x}_{j} = \sum_{m=1}^{M} \tilde{v}_{jm}^2
\]

(15)

where \( \tilde{r}_{jk} \) is used to denote the sample correlation between additively-coded genotypes at the \( j \)-th and \( k \)-th variants, and \( \tilde{v}_{jm} \) 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)

\[
\mathbb{E}[\tilde{r}_{jk}^2] \approx r_{jk}^2 + (1 - r_{jk}^2)/N,
\]

\[
\mathbb{E}[\tilde{v}_{jm}^2] \approx v_{jm}^2 + (1 - v_{jm}^2)/N.
\]

(16)

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

\[
\mathbb{E} \left[ \sum_{k=1}^{J} \tilde{r}_{jk}^2 \right] \approx \ell_{j} + (J - \ell_{j})/N,
\]

\[
\mathbb{E} \left[ \sum_{m=1}^{M} \tilde{v}_{jm}^2 \right] \approx f_{j} + (M - f_{j})/N
\]

(17)

where \( \ell_{j} \) 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

\[
\mathbb{E}[\chi_{j}^2] \approx N \left[ \left( \frac{H^2 \rho}{J} \right) \ell_{j} + \left( \frac{H^2 (1 - \rho)}{M} \right) f_{j} + \frac{1}{N}(1-H^2) \right] = \ell_{j} \tau + f_{j} \sigma + 1
\]

(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 $1$ 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

$$E[\chi^2] \approx \ell \tau + f \sigma + 1$$  \hspace{1cm} (19)$$

where $\chi = (\chi_1, \ldots, \chi_J)$ is a $J$-dimensional vector of chi-square summary statistics, and $\ell = (\ell_1, \ldots, \ell_J)$ and $f = (f_1, \ldots, f_J)$ are $J$-dimensional vectors of additive and marginal epistatic LD scores, respectively. It is important to note that, while $\chi$ must be recomputed for each trait of interest, both vectors $\ell$ and $f$ 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

$$\hat{\sigma} = (f^T \Psi f)^{-1} f^T \Psi \chi^2, \quad \psi_{jj} = [\ell_j \hat{\tau} + f_j \hat{\sigma} + 1]^{-2}$$  \hspace{1cm} (20)$$

where $\Psi$ 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. Standard errors for each coefficient estimate are derived via a delete-one jackknife over blocks of SNPs in the data, 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 $\hat{\tau} - \hat{\sigma}$ (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 \( W \) 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., \( x_j^T(x_l \circ x_l) \) for \( l \neq j \)) is bound to be much larger than the correlation between the genotype of the \( j \)-th SNP \( x_j \) and interactions involving some other SNP (e.g., \( x_j^T(x_k \circ x_l) \) for \( k \neq j \) and \( l \neq j \)). To that end, we can compute the MELD score for each SNP by replacing the full \( W \) matrix with \( W_j \) 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, \( W_j \), does not ever need to be explicitly generated. Referencing Eq. (15), the MELD scores are defined as \( x_j^T W_j x_j / N^2 \). This can be re-written as \( x_j^T (D_j X^{(j)}) (D_j X^{(j)})^T x_j \), where \( D_j = \text{diag}(x_j) \) is a diagonal matrix with the \( j \)-th genotype as its nonzero elements\(^{51} \) and \( X^{(j)} \) 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

\[
\left. \rightrfloor f_j \approx \frac{1}{N^2} \langle x_j^T \rangle^2 X^{(j)} X^{(j)^T} \langle x_j \rangle^2. \tag{21} \]

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 \( X^{(j)} \) (or equivalently \( W_j \) 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\textsuperscript{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 $f^{(l)}$ generated under the $L$-different window sizes, we compute normalized importance weights using their maximized likelihoods via the following

$$
\pi^{(l)} = \frac{L \left( \ell, f^{(l)}; \hat{\beta} \right)}{\sum_{l'} L \left( \ell, f^{(l')}; \hat{\beta} \right)}, \quad \sum_{l=1}^{L} \pi^{(l)} = 1.
$$

(22)

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

$$
\hat{\tau} = \sum_{l=1}^{L} \pi^{(l)} \hat{\tau}^{(l)}, \quad \hat{\sigma} = \sum_{l=1}^{L} \pi^{(l)} \hat{\sigma}^{(l)}.
$$

(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\textsuperscript{73}. In the current study, we average over estimated marginal epistatic LD scores generated using different windows of $\pm 5$, $\pm 10$, $\pm 25$, and $\pm 50$ 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 $\alpha$ to better represent the nonlinear relationship between MAF and variant effect size. The main idea is that this $\alpha$ 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\(^7^4\). Namely, we use

\[
\ell_j(c) := \sum_k L_{jk}(\alpha)a_c(k), \quad L_{jk}(\alpha) = r_{jk}^2 V[x_k]^{1-\alpha}
\] (24)

where \(a_c(k)\) is the annotation value for the \(c\)-th categorical bin. The \(\alpha\) parameter is unknown in practice and needs to be estimated for any given trait. While standard ranges for \(\alpha\) can be used for heritability estimates, we use a restricted maximum likelihood (REML) based method which was recently developed\(^5^9\). In the MELD software, we use this \(\alpha\) construction to handle the relationship between MAF and variant effect size for two specific reasons. First, by constructing the LD scores using \(\alpha\), we more accurately capture the variation in chi-square test statistics due to additive effects\(^7^4\). 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 \(\alpha\) 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 \(\alpha\) values for each trait that had not been analyzed by previous studies\(^5^9\). The \(\alpha\) estimates for each of the 25 traits analyzed in this study are shown in Table S4. Intuitively, \(\alpha\) 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 \(\alpha\) indicate that lower frequency variants contribute more to the observed variation in a trait of interest, whereas values of \(\alpha\) closer to zero indicate that common variants contribute a greater amount of variation to observed trait values.

We took \(\alpha\) values for 11 traits (again see Table S4) that had previously been calculated from Schoech et al.\(^5^9\). 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 \texttt{GRM-MAF-LD} software, https://github.com/arminschoech/GRM-MAF-LD. We then used the \texttt{GCTA} software\(^7\) to obtain heritability and likelihood estimates using REML for each \( \alpha \)-trait pairing. We then fit a trait-specific profile likelihood across the range of \( \alpha \) values and estimate the maximum likelihood value of \( \alpha \) 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

\[
y = X\beta + W\theta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \kappa^2 I), \tag{25}
\]

where \( y \) is an \( N \)-dimensional vector containing all the phenotypes; \( X \) is an \( N \times J \) matrix of genotypes encoded as 0, 1, or 2 copies of a reference allele; \( \beta \) is a \( J \)-dimensional vector of additive effect sizes for each SNP; \( W \) is an \( N \times M \) matrix which holds all pairwise interactions between the randomly selected subset of the interacting SNPs with corresponding effects \( \theta \); and \( \epsilon \) is an \( N \)-dimensional vector of environmental noise. The phenotypic variance is assumed to be \( \mathbb{V}[y] = 1 \). 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 \( \mathbb{V}[X\beta] + \mathbb{V}[W\theta] = H^2 \). Note that we do not assume any specific correlation structure between the effect sizes \( \beta \) and \( \theta \). We then rescale the random error term such that \( \mathbb{V}[\epsilon] = (1 - H^2) \). In the main text, we compare the traditional \texttt{LDSC} to its direct extension in \texttt{MELD}. For each method, GWA summary statistics are computed by fitting a single-SNP univariate linear model via least squares where \( \hat{\beta}_j = (x_j^T x_j)^{-1} x_j^T y \) for every \( j = 1, \ldots, J \) SNP in the data. These effect size estimates are used to derive the chi-square test statistics \( \chi_j^2 = N \hat{\beta}_j^2 \). We implement both \texttt{LDSC} and \texttt{MELD} with the LD matrix \( R = X^T X / N \) and the additive-epistatic correlation matrix \( V = X^T W / N \) being computed using a reference panel of 489 individuals from the European superpopulation (EUR) of the 1000 Genomes Project. The resulting matrices \( R \) and \( V \) 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 $\V{X\beta} = \rho H^2$, while the proportion detailed by genetic interactions is given as $\V{W\theta} = (1-\rho)H^2$. 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.\textsuperscript{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: ±10 and ±100 kilobase (kb);
- allele frequency parameter: $\alpha = -1$, -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 MELD framework to 25 continuous traits the UK Biobank\textsuperscript{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 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. Age, sex, and the first twenty principal components were included as covariates for all traits analyzed. Principal component analysis was performed using FlashPCA 2.0 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. 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 Resource76 (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.
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 α = 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 $\pm 5$ (green), $\pm 10$ (orange), $\pm 25$ (purple), and $\pm 50$ (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 ($\sigma$) in the MELD regression model and checking whether $P < 0.05$. (B) Estimates of the epistatic coefficient ($\sigma$). 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 ($\sigma$) in MELD. Results are based on 100 simulations per parameter combination and the horizontal bars represent standard errors. Similar plots for a range of $\alpha$ 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 ($\sigma$) for 25 quantitative 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 $\sigma$) between traits in the UK Biobank and BioBank Japan. (D) MELD estimates of the additive coefficient $\tau$. Note that the narrow-sense heritability estimates displayed in panels (A) and (B) are also given in Table S6.
### Tables

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

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 ($P = 0.530$).
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.
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 $\pm 5$ (green), $\pm 10$ (orange), $\pm 25$ (purple), and $\pm 50$ (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 ($\sigma$) in the MELD regression model and checking whether $P < 0.05$. (B) Estimates of the epistatic coefficient ($\sigma$). 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 ($\sigma$) 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 (\( \sigma \)) in the MELD regression model and checking whether \( P < 0.05 \). (B) Estimates of the epistatic coefficient (\( \sigma \)). 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 (\( \sigma \)) 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.
Biobank Japan

UK Biobank

\[ r^2 = 0.959 \]

\[ r^2 = 0.986 \]

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.
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 $h^2 = \rho \cdot h^2$. 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.
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.

<table>
<thead>
<tr>
<th>$h^2$</th>
<th>LDSC</th>
<th>MELD ($\pm 5$ SNPs)</th>
<th>MELD ($\pm 10$ SNPs)</th>
<th>MELD ($\pm 25$ SNPs)</th>
<th>MELD ($\pm 50$ SNPs)</th>
<th>MELD (Averaged)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>True $h^2$</td>
<td>Estimated $h^2$</td>
<td>MAE</td>
<td>Estimated $h^2$</td>
<td>MAE</td>
<td>Estimated $h^2$</td>
</tr>
<tr>
<td>0.12</td>
<td>0.315 (0.002)</td>
<td>0.194 (0.002)</td>
<td>-0.072 (0.004)</td>
<td>0.193 (0.004)</td>
<td>0.064 (0.003)</td>
<td>0.058 (0.003)</td>
</tr>
<tr>
<td>0.24</td>
<td>0.382 (0.002)</td>
<td>0.142 (0.002)</td>
<td>0.100 (0.004)</td>
<td>0.140 (0.004)</td>
<td>0.200 (0.004)</td>
<td>0.045 (0.003)</td>
</tr>
<tr>
<td>0.36</td>
<td>0.450 (0.003)</td>
<td>0.090 (0.003)</td>
<td>0.277 (0.004)</td>
<td>0.083 (0.004)</td>
<td>0.340 (0.004)</td>
<td>0.045 (0.003)</td>
</tr>
<tr>
<td>0.48</td>
<td>0.523 (0.003)</td>
<td>0.044 (0.002)</td>
<td>0.440 (0.004)</td>
<td>0.047 (0.004)</td>
<td>0.471 (0.004)</td>
<td>0.031 (0.002)</td>
</tr>
<tr>
<td></td>
<td>MELD ($\pm 50$ SNPs)</td>
<td>MELD (Averaged)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.12</td>
<td>0.229 (0.002)</td>
<td>0.109 (0.002)</td>
<td>0.100 (0.006)</td>
<td>0.105 (0.003)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.24</td>
<td>0.318 (0.003)</td>
<td>0.078 (0.003)</td>
<td>0.225 (0.005)</td>
<td>0.077 (0.003)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.36</td>
<td>0.414 (0.003)</td>
<td>0.055 (0.003)</td>
<td>0.360 (0.003)</td>
<td>0.053 (0.002)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.48</td>
<td>0.509 (0.003)</td>
<td>0.034 (0.002)</td>
<td>0.479 (0.002)</td>
<td>0.036 (0.001)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trait Name</td>
<td>Code</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------------------</td>
<td>--------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Body mass index</td>
<td>BMI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High density lipoprotein</td>
<td>HDL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low density lipoprotein</td>
<td>LDL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hemoglobin A1c</td>
<td>HBA1C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estimated glomerular filtration rate</td>
<td>EGFR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C-reactive protein</td>
<td>CRP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Systolic blood pressure</td>
<td>SBP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diastolic blood pressure</td>
<td>DBP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Platelet count</td>
<td>PLC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean corpuscular hemoglobin</td>
<td>MCHC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean corpuscular hemoglobin</td>
<td>MCH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean corpuscular volume</td>
<td>MCV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Red blood cell count</td>
<td>RBC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White blood cell count</td>
<td>WBC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**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.
Table S4. Trait-specific $\alpha$ parameters for each of the 25 traits analyzed. Here, $\alpha$ values are used to weight each variant based on its minor allele frequency to account for frequency dependent architectures in each trait. The * indicates $\alpha$ parameters that were taken directly from Schoech et al.\textsuperscript{59}. The $\alpha$ parameters for other traits were calculated using the protocol used in that paper. Expansion of trait abbreviations are given in Table S3.
<table>
<thead>
<tr>
<th>Trait Name or Code</th>
<th>Sample Size</th>
<th>Total SNPs</th>
<th>Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basophil count</td>
<td>62,076</td>
<td>5,653,566</td>
<td>80</td>
</tr>
<tr>
<td>BMI</td>
<td>158,284</td>
<td>5,653,566</td>
<td>81</td>
</tr>
<tr>
<td>CRP</td>
<td>75,391</td>
<td>5,608,701</td>
<td>80</td>
</tr>
<tr>
<td>DBP</td>
<td>136,615</td>
<td>5,653,566</td>
<td>80</td>
</tr>
<tr>
<td>eGFR</td>
<td>143,658</td>
<td>5,608,701</td>
<td>80</td>
</tr>
<tr>
<td>Eosinophil count</td>
<td>62,076</td>
<td>5,653,566</td>
<td>80</td>
</tr>
<tr>
<td>HDL</td>
<td>70,657</td>
<td>5,608,701</td>
<td>80</td>
</tr>
<tr>
<td>Height</td>
<td>159,095</td>
<td>6,296,332</td>
<td>82</td>
</tr>
<tr>
<td>Hematocrit</td>
<td>108,757</td>
<td>5,653,566</td>
<td>80</td>
</tr>
<tr>
<td>Hemoglobin</td>
<td>108,769</td>
<td>5,653,566</td>
<td>80</td>
</tr>
<tr>
<td>HbA1c</td>
<td>75,391</td>
<td>5,608,701</td>
<td>80</td>
</tr>
<tr>
<td>LDL</td>
<td>72,866</td>
<td>5,608,701</td>
<td>80</td>
</tr>
<tr>
<td>Lymphocyte count</td>
<td>62,076</td>
<td>5,653,566</td>
<td>80</td>
</tr>
<tr>
<td>MCH</td>
<td>108,054</td>
<td>5,653,566</td>
<td>80</td>
</tr>
<tr>
<td>MCHC</td>
<td>108,738</td>
<td>5,653,566</td>
<td>80</td>
</tr>
<tr>
<td>MCV</td>
<td>108,526</td>
<td>5,653,566</td>
<td>80</td>
</tr>
<tr>
<td>Monocyte count</td>
<td>62,076</td>
<td>5,653,566</td>
<td>80</td>
</tr>
<tr>
<td>Neutrophil count</td>
<td>62,076</td>
<td>5,653,566</td>
<td>80</td>
</tr>
<tr>
<td>PLC</td>
<td>108,208</td>
<td>5,653,566</td>
<td>80</td>
</tr>
<tr>
<td>RBC</td>
<td>108,794</td>
<td>5,653,566</td>
<td>80</td>
</tr>
<tr>
<td>SBP</td>
<td>136,597</td>
<td>5,653,566</td>
<td>80</td>
</tr>
<tr>
<td>Cholesterol</td>
<td>128,305</td>
<td>5,608,701</td>
<td>80</td>
</tr>
<tr>
<td>Triglyceride</td>
<td>105,597</td>
<td>5,608,701</td>
<td>80</td>
</tr>
<tr>
<td>Urate</td>
<td>109,029</td>
<td>5,608,701</td>
<td>80</td>
</tr>
<tr>
<td>WBC</td>
<td>107,694</td>
<td>5,653,566</td>
<td>80</td>
</tr>
</tbody>
</table>

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

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


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


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


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


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


