Recalibrating the Genome-wide Association Null Hypothesis Enables the Detection of Core Genes Underlying Complex Traits

Wei Cheng[1,2], Sohini Ramachandran[1,2†], and Lorin Crawford[2-4†]

1 Department of Ecology and Evolutionary Biology, Brown University, Providence, RI, USA
2 Center for Computational Molecular Biology, Brown University, Providence, RI, USA
3 Department of Biostatistics, Brown University, Providence, RI, USA
4 Center for Statistical Sciences, Brown University, Providence, RI, USA

† E-mail: sramachandran@brown.edu; lorin_crawford@brown.edu

Abstract

Traditional univariate genome-wide association studies generate false positives and negatives due to difficulties distinguishing causal variants from "interactive" variants (i.e., variants correlated with causal variants without directly influencing the trait). Recent efforts have been directed at identifying gene or pathway associations, but these can be computationally costly and hampered by strict model assumptions. Here, we present gene-ε, a new approach for identifying statistical associations between sets of variants and quantitative traits. Our key innovation is a recalibration of the genome-wide null hypothesis to treat small-yet-nonzero associations emitted by interactive variants as non-causal. gene-ε efficiently identifies core genes under a variety of simulated genetic architectures, achieving greater than a 90% true positive rate at 1% false positive rate for polygenic traits. Lastly, we apply gene-ε to summary statistics derived from six quantitative traits using European-ancestry individuals in the UK Biobank, and identify genes that are enriched in biological relevant pathways.

Introduction

Over the last decade, there has been considerable debate surrounding whether genome-wide single-nucleotide polymorphism (SNP) genotype data can offer insight into the genetic architecture of complex traits [1–5]. In the traditional genome-wide association (GWA) framework, individual SNPs are tested independently for association with a trait of interest. While this approach has drawbacks [2, 3, 6], more recent approaches that combine SNPs within a region have gained power to detect biologically relevant genes and pathways enriched for correlations with complex traits [7–14]. Reconciling these two observations is crucial for biomedical genomics.

In the traditional GWA model, each SNP is assumed to either (*i*) directly influence (or perfectly tag a variant that directly influences) the trait of interest; or (*ii*) have no affect on the trait at all (see Fig. 1a). Throughout this manuscript, for simplicity, we refer to SNPs under the former as "causal" and those under latter as "non-causal". These classifications are based on ordinary least squares (OLS) effect size estimates $\widehat{\beta}_j$ for each j-th SNP in a regression framework, where the null hypothesis assumes no association for the true effects of non-causal SNPs ($H_0\colon \beta_j = 0$). The traditional GWA model is agnostic to trait architecture, and is underpowered with a high false-positive rate for "polygenic" traits or traits which are generated by many mutations of small effect [5, 15–17].

Suppose that in truth each SNP in a GWA dataset instead belongs to one of *three* categories: (*i*) causal or truly associated; (*ii*) statistically associated with the trait but not directly influencing it, which we refer to as "interactive"; and (*iii*) non-causal (Fig. 1b) [18]. Equivalently, these three categories can be

defined using the relationship between the true SNP effects β and the observed effect size estimates $\widehat{\beta}$: (i) for causal SNPs, the true effects $\beta \neq 0$ and the observed effect size estimates $\widehat{\beta} \neq 0$; (ii) for "interactive" SNPs, $\beta = 0$ or $\beta \approx 0$, and $\widehat{\beta} \neq 0$; and (iii) for non-causal SNPs, $\beta = 0$ and $\widehat{\beta} = 0$. More specifically, causal SNPs lie in core genes that directly influence the trait of interest. Interactive SNPs then follow one of two definitions. They are either non-causal SNPs that are statistically associated with the trait due to some varying degree of linkage disequilibrium (LD) with causal SNPs ($\beta = 0$); or they are SNPs that are not directly causal, but are located within a gene that covaries with a core gene ($\beta \approx 0$) [19]. In either setting, interactive SNPs will emit intermediate statistical noise (in some cases, even appearing indistinguishable from causal SNPs), thereby confounding traditional GWA tests (Fig. 1b). Hereafter, we refer to this noise as "epsilon-genic effects" (denoted in shorthand as "ε-genic effects"). There is a need for a computational framework that has the ability to identify mutations associated with a wide range of traits, regardless of whether heritability is sparsely or uniformly distributed across the genome.

Here, we develop a new and scalable quantitative approach for testing aggregated sets of SNP-level GWA summary statistics. In practice, our approach can be applied to any user-specified set of genomic regions, such as regulatory elements, intergenic regions, or gene sets. In this study, for simplicity, we refer to our method as a gene-level test (i.e., an annotated collection of SNPs within the boundary of a gene). The key contribution for our approaches is that gene-level association tests should treat interactive SNPs with ε-genic effects as non-causal. Conceptually, this requires assessing whether a given SNP explains more than an "epsilon" proportion of narrow-sense heritability, h^2. In this generalized model, we assume that true SNP-level effects for complex traits are jointly distributed as

$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}), \qquad \boldsymbol{\Omega}_{jj} = \sigma_j^2, \qquad \boldsymbol{\Omega}_{jl} = \sigma_j \rho(\mathbf{x}_j, \mathbf{x}_l)\sigma_l, \tag{1}$$

where $\boldsymbol{\Omega} = \mathbf{H}^{1/2}\boldsymbol{\Sigma}\mathbf{H}^{1/2}$ is a $J \times J$ variance-covariance matrix, and $\mathbf{H} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_J^2)$ is a $J \times J$ diagonal matrix with elements σ_j^2 denoting the proportion of narrow-sense heritability h^2 contributed by the j-th SNP (with the constraint $\sum \sigma_j^2 = h^2$). Under our framework, we recalibrate the GWA null hypothesis to assume *approximately* no association for interactive and non-causal SNPs, $H_0 : \beta_j \approx 0$. This can be formally restated using the variance-component-based null hypothesis, $H_0 : \sigma_j^2 \leq \sigma_\varepsilon^2$, where σ_ε^2 represents the maximum proportion-of-variance-explained (PVE) by an interactive or non-causal SNP (Fig. 1b). Non-core genes are then defined as genes that only contain SNPs with ε-genic effects (i.e. $0 \leq \sigma_j^2 \leq \sigma_\varepsilon^2$ for every j-th SNP within that region). Core genes, on the other hand, are genes that contain at least one causal SNP (i.e. $\sigma_j^2 > \sigma_\varepsilon^2$ for at least one SNP j within that region). By modeling ε-genic effects (i.e., different values of σ_ε^2 for interactive SNPs), our approach flexibly constructs an appropriate null hypothesis for a wide range of traits with genetic architectures that land anywhere on the polygenic spectrum (see Methods).

We refer to our gene-level association framework as "gene-ε" (pronounced "genie"). gene-ε lowers false positive rates and increases power for identifying gene-level associations from GWA studies via two key conceptual insights. First, gene-ε regularizes observed (and inflated) GWA summary statistics so that SNP-level effect size estimates are positively correlated with the assumed generative model of complex traits. Second, it examines the distribution of regularized effect sizes to determine the ε-genic threshold σ_ε^2 of interactive and non-causal SNPs. This makes for an improved hypothesis testing strategy for identifying core genes associated with complex traits. With detailed simulations, we assess the power of gene-ε to identify core genes under a variety of genetic architectures, and compare its performance against multiple competing approaches [7, 10, 12, 14, 20]. We also apply gene-ε to the SNP-level summary statistics of six quantitative traits assayed in individuals of European ancestry from the UK Biobank [21].

Results

Overview of gene-ε

The gene-ε framework requires two inputs: GWA SNP-level effect size estimates, and an empirical linkage disequilibrium (LD, or variance-covariance) matrix. The LD matrix can be estimated directly from genotype data, or from an ancestry-matched set of samples if genotype data are not available to the user. We use these inputs to both estimate gene-level contributions to narrow-sense heritability h^2, and perform gene-level association tests. After preparing the input data, there are three steps implemented in gene-ε, which are detailed below (Fig. 2).

First, we shrink the GWA effect size estimators via regularized regression (Figs. 2a,b; Eq. (5) in Methods). This shrinkage step reduces the inflation of effect sizes for interactive SNPs [22], and increases their correlation with the assumed generative model for the trait of interest (particularly for traits with high heritability; Supplementary Fig. 1). When assessing the performance of gene-ε in simulations, we considered different types of regularization for the effect size estimates: the Least Absolute Shrinkage And Selection Operator (gene-ε-LASSO) [23], the Elastic Net solution (gene-ε-EN) [24], and Ridge Regression (gene-ε-RR) [25]. We also assessed our framework using the observed ordinary least squares (OLS) estimates without any shrinkage (gene-ε-OLS) to serve as motivation for having regularization as a step in the framework.

Second, we fit a K-mixture Gaussian model to the regularized effect sizes with the goal of classifying SNPs as causal, intermediate, or non-causal (Fig. 2c). Each successive Gaussian mixture component has distinctly smaller variances ($\sigma_1^2 > \cdots > \sigma_K^2$) with the K-th component fixed at $\sigma_K^2 = 0$. In this study, we assume causal SNPs to appear in the first mixture component with the largest variance σ_1^2, while non-causal SNPs appear in the last component σ_K^2. By definition, interactive SNPs with ε-genic effects then have PVEs that fall at or below the variance of the second component (i.e., $\sigma_\varepsilon^2 = \sigma_2^2$ and $H_0 \colon \sigma_j^2 \leq \sigma_2^2$ for the j-th SNP). Indeed, while core genes may contain important variants that appear in intermediate mixture components, the biological definitions of these components may not be consistent across trait architectures. As a result, for results shown here, we take a conservative approach in our definition of causal SNPs within core genes. Fig. 1b illustrates this concept with a three-component mixture model (see also [18]). Intuitively, the intermediate mixture components may represent varying degrees of connectivity to core genes via LD or trans-interactions. Thus, gene-ε allows for flexibility in the number of Gaussians that specify the range of null and non-null SNP effects. For scalability, we estimate parameters of the K-mixture model using an expectation-maximization (EM) algorithm.

Third, we group the regularized GWA summary statistics according to gene boundaries (or user-specified regions) and compute a gene-level association statistic based on a commonly used quadratic form (Fig. 2d) [7,12,20]. In expectation, these test statistics can be naturally interpreted as the contribution of each gene to the narrow-sense heritability. We use Imhof's method [26] to derive a P-value for assessing evidence in support of an association between a given gene and the trait of interest. Details for each of these steps can be found in Methods and Supplementary Notes.

Performance Comparisons in Simulation Studies

To assess the performance of gene-ε, we simulated complex traits under multiple genetic architectures using real genotype data on chromosome 1 from individuals of European ancestry in the UK Biobank (Methods). Following quality control procedures, our simulations included 36,518 SNPs (Supplementary Notes). Next, we used the NCBI's Reference Sequence (RefSeq) database in the UCSC Genome Browser [27] to annotate SNPs with the appropriate genes. Simulations were conducted using two different SNP-to-gene assignments. In the first, we directly used the UCSC annotations which resulted in 1,408 genes to be used in the simulation study. In the second, we augmented the UCSC gene boundaries to include SNPs within ±50kb, which resulted in 1,916 genes in the simulation study. For both cases, we assumed a linear

additive model for quantitative traits, while varying the following parameters: sample size ($N = 5,000$ or 10,000); narrow-sense heritability ($h^2 = 0.2$ or 0.6); and the percentage of core genes (set to 1% or 10%). In each scenario, we considered traits being generated with and without additional population structure. In the latter setting, traits are simulated while also using the top ten principal components of the genotype matrix as covariates to create stratification. Regardless of the setting, GWA summary statistics were computed by fitting a single-SNP univariate linear model (via OLS) without any control for population structure. Comparisons were based on 100 different simulated runs for each parameter combination.

We compared the performance of gene-ε against that of five competing gene-level association or enrichment methods: SKAT [20], VEGAS [7], MAGMA [10], PEGASUS [12], and RSS [14] (Supplementary Notes). As previously noted, we also explored the performance of gene-ε while using various degrees of regularization on effect size estimates, with gene-ε-OLS being treated as a baseline. SKAT, VEGAS, and PEGASUS are frequentist approaches, in which SNP-level GWA P-values are drawn from a correlated chi-squared distribution with covariance estimated using an empirical LD matrix [28]. MAGMA is also a frequentist approach in which gene-level p-values are derived from distributions of SNP-level effect sizes using an F-test [10]. RSS is a Bayesian model-based enrichment method which places a likelihood on the observed SNP-level GWA effect sizes (using their standard errors and LD estimates), and assumes a spike-and-slab shrinkage prior on the true SNP effects [29]. Conceptually, SKAT, MAGMA, VEGAS, and PEGASUS assume null models under the traditional GWA framework, while RSS and gene-ε allow for traits to have architectures with interactive SNPs.

For all methods, we assess the power and false discovery rates (FDR) for identifying core genes at a Bonferroni-corrected threshold ($P = 0.05/1,408$ genes $= 3.55 \times 10^{-5}$ and $P = 0.05/1,916$ genes $= 2.61 \times 10^{-5}$, depending on if the $\pm$50kb buffer was used) or median probability model (posterior enrichment probability > 0.5; see [30]) (Supplementary Tables 1-16). We also compare their ability to rank true positives over false positives via receiver operating characteristic (ROC) and precision-recall curves (Fig. 3 and Supplementary Figs. 2-16). While we find gene-ε and RSS have the best tradeoff between true and false positive rates, RSS does not scale well for genome-wide analyses (Supplementary Table 17). In many settings, gene-ε has similar power to RSS (while maintaining a considerably lower FDR), and generally outperforms RSS in precision-versus-recall. gene-ε also stands out as the best approach in scenarios where the observed OLS summary statistics were produced without first controlling for confounding stratification effects in more heritable traits (i.e., $h^2 = 0.6$). Computationally, gene-ε gains speed by directly assessing evidence for rejecting the recalibrated GWA null hypothesis, whereas RSS must compute the posterior probability of being a core gene (which can suffer from convergence issues; Supplementary Notes). For context, an analysis of just 1,000 genes takes gene-ε an average of 140 seconds to run on a personal laptop, while RSS takes around 9,400 seconds to complete.

When using GWA summary statistics to identify genotype-phenotype associations, modeling the appropriate trait architecture is crucial. As expected, all methods we compared in this study have relatively more power for traits with high h^2. However, our simulation studies confirm the expectation that the max utility for methods assuming the traditional GWA framework (i.e., SKAT, MAGMA, VEGAS, and PEGASUS) is limited to scenarios where heritability is low, phenotypic variance is dominated by just a few core genes with large effects, and summary statistics are not confounded by population structure (Supplementary Figs. 2, 3, 9, and 10). RSS, gene-ε-EN, and gene-ε-LASSO robustly outperform these methods for the other trait architectures (Fig. 3 and Supplementary Figs. 4-8 and 11-16). One major reason for this result is that shrinkage and penalized regression methods appropriately correct for inflation in GWA summary statistics (Supplementary Fig. 1). For example, we find that the regularization used by gene-ε-EN and gene-ε-LASSO is able to recover effect size estimators that are almost perfectly correlated ($r^2 > 0.9$) with the true effect sizes used to simulate sparse architectures (e.g., simulations with 1% core genes). Regularization also allows gene-ε to preserve type 1 error when traits are generated under the null hypothesis. Importantly, our method is relatively conservative when GWA summary statistics are

less precise and derived from studies with smaller sample sizes (e.g., N = 5,000; Supplementary Table 18).

Characterizing Genetic Architecture of Quantitative Traits in the UK Biobank

We applied gene-ε to 1,070,306 genome-wide SNPs and six quantitative traits — height, body mass index (BMI), mean red blood cell volume (MCV), mean platelet volume (MPV), platelet count (PLC), waist-hip ratio (WHR) — assayed in 349,414 European-ancestry individuals in the UK Biobank (Supplementary Notes) [21]. After quality control, we regressed the top ten principal components of the genotype data onto each trait to control for population structure, and then we derived OLS SNP-level effect sizes using the traditional GWA framework. For completeness, we then analyzed these GWA effect size estimates with the four different implementations of gene-ε. In the main text, we highlight results under the Elastic Net solution; detailed findings with the other gene-ε approaches can be found in Supplementary Notes.

While estimating ε-genic effects, gene-ε provides insight into to the genetic architecture of a trait (Supplementary Table 19). For example, past studies have shown human height to have a higher narrow-sense heritability (estimates ranging from 45-80%; [6, 31–39]). Using Elastic Net regularized effect sizes, gene-ε estimated approximately 11% of SNPs in the UK Biobank to be associated with height. This meant approximately 110,000 SNPs had marginal PVEs $\sigma_j^2 > 0$ (Methods). This number is similar to the 93,000 and 100,000 height associated variants previously estimated by Goldstein [40] and Boyle et al. [4], respectively. Additionally, gene-ε identified approximately 2% of SNPs to be causal (meaning they had PVEs greater than the ε-genic threshold, $\sigma_j^2 > \sigma_2^2$); again similar to the Boyle et al. [4] estimate of 3.8% causal SNPs for height using data from the GIANT Consortium [32], and the Lello et al. [41] estimate of 3.1% causal SNPs for height using European-ancestry individuals in the UK Biobank.

Compared to body height, narrow-sense heritability estimates for BMI have been considered both high and low (estimates ranging from 25-60%; [31, 33, 34, 36, 37, 39, 42–45]). Such inconsistency is likely due to difference in study design (e.g., twin, family, population-based studies), many of which have been known to produce different levels of bias [44]. Here, our results suggest BMI to have a lower narrow-sense heritability than height, with a slightly different distribution of null and non-null SNP effects. Specifically, we found BMI to have 13% associated SNPs and 6% causal SNPs.

In general, we found our genetic architecture characterizations in the UK Biobank to reflect the same general themes we saw in the simulation study. Less aggressive shrinkage approaches (e.g., OLS and Ridge) are subject to misclassifications of causal, interactive, and non-causal SNPs. As result, these methods struggle to reproduce well-known narrow-sense heritability estimates from the literature, across all six traits. This once again highlights the need for computational frameworks that are able to appropriately correct for inflation in summary statistics.

gene-ε Identifies New Core Genes and Novel Genetic Associations

Next, we applied gene-ε to the summary statistics from the UK Biobank and generated genome-wide gene-level association P-values (panels a,b of Fig. 4 and Supplementary Figs. 17-21). As in the simulation study, we conducted two separate analyses using two different SNP-to-gene annotations: *(i)* we used the RefSeq database gene boundary definitions directly, or *(b)* we augmented the gene boundaries by adding SNPs within a ±50 kilobase (kb) buffer to account for possible regulatory elements. A total of 14,322 genes were analyzed when using the UCSC boundaries as defined, and a total of 17,680 genes were analyzed when including the 50kb buffer. The ultimate objective of gene-ε is to identify core genes, which we define as containing at least one causal SNP and achieving a gene-level association P-value below a Bonferroni-corrected significance threshold (in our two analyses, $P = 0.05/14,322$ genes $= 3.49 \times 10^{-6}$ and $P = 0.05/17,680$ genes 2.83×10^{-6}, respectively; Supplementary Tables 20-25). As a validation step, we compared gene-ε P-values to RSS posterior enrichment probabilities for each gene. We also used the gene set enrichment analysis tool Enrichr [46] to identify dbGaP categories with an overrepresentation

of core genes reported by gene-ε (panels c,d of Fig. 4 and Supplementary Figs. 17-21). A comparison of gene-level associations and gene set enrichments between the different gene-ε approaches are also listed (Supplementary Tables 26-28).

Many of the candidate core genes we identified by applying gene-ε were not previously annotated as having trait-specific associations in either dbGaP or the GWAS catalog (Fig. 4); however, many of these same candidate core genes have been identified by past publications as related to the phenotype of interest (Table 1). Additionally, 45% of the core genes selected by gene-ε were also selected by RSS. For example, gene-ε reports *C1orf150* as having a significant gene-level association with MPV ($P = 1 \times 10^{-20}$ and RSS posterior enrichment probability of 1), which is known to be associated with germinal center signaling and the differentiation of mature B cells that mutually activate platelets [47–49]. Importantly, nearly all of the core genes reported by gene-ε had evidence of overrepresentation in gene set categories that were at least related to the trait of interest. As expected, the top categories with Enrichr Q-values smaller than 0.05 for height and MPV were "Body Height" and "Platelet Count", respectively. Even for the less heritable MCV, the top significant gene sets included hematological categories such as "Transferrin", "Erythrocyte Indices", "Hematocrit", "Narcolepsy", and "Iron" — all of which have verified and clinically relevant connections to trait [50–57].

Lastly, gene-ε also identified genes with rare causal variants. For example, *ZNF628* (which is not mapped to height in the GWAS catalog) was detected by gene-ε with a significant P-value of 1×10^{-20} (and $P = 4.58 \times 10^{-8}$ when the gene annotation included a 50kb buffer). Previous studies have shown a rare variant *rs147110934* within this gene to significantly affect adult height [38]. Rare and low-frequency variants are generally harder to detect under the traditional GWA framework. However, rare variants have been shown to be important for explaining the variation of complex traits [28, 39, 58–61]. With regularization and testing for ε-genic effects, gene-ε is able to distinguish between rare variants that are causal and SNPs with larger effect sizes due various types of correlations. This only enhances the power of gene-ε to identify potential novel core genes.

Discussion

During the past decade, it has been repeatedly observed that the traditional GWA framework struggles to accurately differentiate between truly associated (which we refer to as "causal" for simplicity) and interactive SNPs (which we define as SNPs that covary with causal SNPs but do not directly influence the trait of interest). As a result, the traditional GWA approach is prone to generating false positives, and detects variant-level associations spread widely across the genome rather than aggregated sets in disease-relevant pathways [4]. While this observation has spurred to many interesting lines of inquiry — such as investigating the role of rare variants in generating complex traits [9, 28, 58, 59], comparing the efficacy of tagging causal variants in different ancestries [62, 63], and integrating GWA data with functional -omics data [64–66] — the focus of GWA studies and studies integrating GWA data with other -omics data is still largely based on the role of individual variants, acting independently.

Here, we challenge the view that signals in GWA datasets cannot reveal the architecture of complex traits by recalibrating the traditional GWA null hypothesis from $H_0 : \beta_j = 0$ (i.e., the j-th SNP has zero statistical association with the trait of interest) to $H_0 : \beta_j \approx 0$. We accomplish this by testing for ε-genic effects: small-yet-nonzero effect sizes emitted by interactive SNPs. We use an empirical Bayesian approach to learn the distributions of ε-genic effects, and then we aggregate regularized SNP-level association signals into a gene-level test statistic that represents the gene's contribution to the narrow-sense heritability of the trait of interest. Together, these two steps reduce false positives and increase power to identify the mutations, genes, and pathways that directly influence a trait's genetic architecture. By considering different levels of ε-genic effects (i.e., different values of σ_ε^2 for interactive SNPs; Figs. 1 and 2), gene-ε offers the flexibility to construct an appropriate null hypothesis for a wide range of traits with genetic architectures that land anywhere on the polygenic spectrum. It is important

to stress that while we repeatedly point to our improved ability distinguish "causal" variants in core genes, gene-ε is by no means a causal inference procedure. Instead, it is an association test which highlights genes in enriched pathways that are most likely to be associated with the trait of interest.

Through simulations, we showed the gene-ε framework outperforms other widely used gene-level association methods (particularly for highly heritable traits), while also maintaining scalability for genome-wide analyses (Fig. 3, Supplementary Figs. 2-16, and Supplementary Tables 1-18). Indeed, all the approaches we compared in this study showed improved performance when they used summary statistics derived from studies with larger sample sizes (i.e., simulations with $N = 10,000$). This is because the quality of summary statistics also improves in these settings (via the asymptotic properties of OLS estimators). Nonetheless, our results suggest that applying gene-ε to summary statistics from previously published studies will increase the return made on investments in GWA studies over the last decade.

Like any aggregated SNP-set association method, gene-ε has its limitations. Perhaps the most obvious limitation is that annotations can bias the interpretation of results and lead to erroneous scientific conclusions (i.e., cause us to highlight the "wrong" gene [14, 67, 68]). We observed some instances of this during the UK Biobank analyses. For example, when studying MPV, *CAPN10* only appeared to be a core gene after its UCSC annotated boundary was augmented by a $\pm$50kb buffer window ($P = 1.85 \times 10^{-1}$ and $P = 1.17 \times 10^{-7}$ before and after the buffer was added, respectively; Supplementary Table 23). After further investigation, this result occurred because the augmented definition of *CAPN10* included nearly all causal SNPs from the significant neighboring gene *RNPEPL1* ($P = 1 \times 10^{-20}$ and $P = 2.07 \times 10^{-9}$ before and after the buffer window was added, respectively). While this shows the need for careful biological interpretation of the results, it also highlights the power of gene-ε to prioritize true genetic signal effectively.

There are several potential extensions for the gene-ε framework described here. First, in the current study, we only focused on applying gene-ε to quantitative traits (Fig. 4 and Supplementary Figs. 17-21). Future studies extending this approach to binary traits (e.g. case-control studies) should explore controlling for additional confounders that can occur within these phenotypes, such as ascertainment [69–71]. Second, we only focus on data consisting of common variants; however, it would be interesting to extend gene-ε for (i) rare variant association testing and (ii) studies that consider the combined effect between rare and common variants. A significant challenge, in either case, would be to adaptively adjust the strength of the regularization penalty on the observed OLS summary statistics for causal rare variants, so as to not misclassify them as interactive SNPs. Previous approaches with specific re-weighting functions for rare variants may help here [9, 28, 58] (Methods). A final related extension of gene-ε is to include information about standard errors when estimating ε-genic effects. In our analyses using the UK Biobank, some of the newly identified core genes contained SNPs that had large effect sizes but insignificant P-values in the original GWA analysis (after Bonferroni-correction; Supplementary Tables 20-25). While this could be attributed to the improved and recalibrated hypothesis test in gene-ε, it also motivates a regularization model that accounts for the standard error of effect size estimates from GWA studies [14, 22, 29].

URLs

gene-ε software, `https://github.com/ramachandran-lab/genee`; UK Biobank, `https://www.ukbiobank.ac.uk`; Database of Genotypes and Phenotypes (dbGaP), `https://www.ncbi.nlm.nih.gov/gap`; NHGRI-EBI GWAS Catalog, `https://www.ebi.ac.uk/gwas/`; UCSC Genome Browser, `https://genome.ucsc.edu/index.html`; Enrichr software, `http://amp.pharm.mssm.edu/Enrichr/`; SNP-set (Sequence) Kernel Association Test (SKAT) software, `https://www.hsph.harvard.edu/skat`; Multi-marker Analysis of GenoMic Annotation (MAGMA) software, `https://ctg.cncr.nl/software/magma`; Precise, Efficient Gene Association Score Using SNPs (PEGASUS) software, `https://github.com/ramachandran-lab/PEGASUS`; Regression with Summary Statistics (RSS) enrichment software, `https://github.com/stephenslab/`

`rss`; Versatile Gene-based Association Study (VEGAS) version 2, `https://vegas2.qimrberghofer.edu.au`.

Acknowledgements

The authors thank Xiang Zhu (Stanford University) for help with the implementation of RSS. They also thank Sam Smith (Brown University) for help with the management of the UK Biobank data. This research was conducted using the UK Biobank Resource under Application Number 22419, and part of this research was conducted using computational resources and services at the Center for Computation and Visualization, Brown University. This research was supported in part by US National Institutes of Health (NIH) grant R01GM118652 and National Science Foundation (NSF) CAREER award DBI-1452622 to S. Ramachandran. This research was also partly supported by grants P20GM109035 (COBRE Center for Computational Biology of Human Disease; PI Rand) and P20GM103645 (COBRE Center for Central Nervous; PI Sanes) from the NIH NIGMS, 2U10CA180794-06 from the NIH NCI and the Dana Farber Cancer Institute (PIs Gray and Gatsonis), as well as by an Alfred P. Sloan Research Fellowship (No. FG-2019-11622) awarded to L. Crawford.

Author Contributions

W.C., S.R., and L.C. conceived the methods. W.C. carried out all analyses. W.C., S.R., and L.C. wrote and reviewed the manuscript.

Competing Interests

The authors declare no competing interests.

Methods

Traditional Association Tests using Summary Statistics

gene-ε requires two inputs: genome-wide association (GWA) marginal effect size estimates $\widehat{\boldsymbol{\beta}}$, and an empirical linkage disequilibrium (LD) matrix $\boldsymbol{\Sigma}$. We assumed the following generative linear model for complex traits

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \qquad \mathbf{e} \sim \mathcal{N}(\mathbf{0}, \tau^2\mathbf{I}), \tag{2}$$

where $\mathbf{y}$ denotes an N-dimensional vector of phenotypic states for a quantitative trait of interest measured in N individuals; $\mathbf{X}$ is an $N \times J$ matrix of genotypes, with J denoting the number of single nucleotide polymorphisms (SNPs) encoded as $\{0, 1, 2\}$ copies of a reference allele at each locus; $\boldsymbol{\beta}$ is a J-dimensional vector containing the additive effect sizes for an additional copy of the reference allele at each locus on $\mathbf{y}$; $\mathbf{e}$ is a normally distributed error term with mean zero and scaled variance τ^2; and $\mathbf{I}$ is an $N \times N$ identity matrix. For convenience, we assumed that the genotype matrix (column-wise) and trait of interest have been mean-centered and standardized. We also treat $\boldsymbol{\beta}$ as a fixed effect. A central step in GWA studies is to infer $\boldsymbol{\beta}$ for each SNP, given both genotypic and phenotypic measurements for each individual sample. For every SNP j, gene-ε takes in the ordinary least squares (OLS) estimates based on Eq. (2)

$$\widehat{\beta}_j = (\mathbf{x}_j^\intercal \mathbf{x}_j)^{-1}\mathbf{x}_j^\intercal \mathbf{y}, \tag{3}$$

where $\mathbf{x}_j$ is the j-th column of the genotype matrix $\mathbf{X}$, and $\widehat{\beta}_j$ is the j-th entry of the vector $\widehat{\boldsymbol{\beta}}$.

In traditional GWA studies, the null hypothesis for statistical tests assumes $H_0 : \beta_j = 0$ for all $j = 1, \ldots, J$ SNPs. When both the trait and SNP measurements are standardized, it can be shown that the correlation between any two regression coefficients $\widehat{\beta}_j$ and $\widehat{\beta}_l$ $(j \neq l)$ is proportional to the correlation between genotypic variants $\mathbf{x}_j$ and $\mathbf{x}_l$. Therefore, under the null hypothesis, $\boldsymbol{\beta}$ is assumed to follow a multivariate normal distribution with mean vector zero and a correlation structure defined by the pairwise SNP-by-SNP LD matrix. For the applications considered here, the LD matrix is 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).

Regularization of GWA Effect Size Estimators

gene-ε uses regularization on the observed GWA summary statistics to reduce inflation of SNP-level effect size estimates and increase their correlation with the assumed generative model of complex traits. For large sample size N, note that the asymptotic relationship between the observed GWA effect size estimates $\widehat{\boldsymbol{\beta}}$ and the true coefficient values $\boldsymbol{\beta}$ is

$$\mathbb{E}[\widehat{\beta}_l] = \sum_{j=1}^{J} \rho(\mathbf{x}_j, \mathbf{x}_l)\beta_j \qquad \Longleftrightarrow \qquad \mathbb{E}[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\Sigma}\boldsymbol{\beta}, \tag{4}$$

where $\boldsymbol{\Sigma}_{jl} = \rho(\mathbf{x}_j, \mathbf{x}_l)$ denotes the correlation coefficient between SNPs $\mathbf{x}_j$ and $\mathbf{x}_l$. The above mirrors a high-dimensional regression model with the misestimated OLS summary statistics as the response variables and the LD matrix as the design matrix. Theoretically, the resulting output coefficients from this model are the desired effect size estimators. Due to the multi-collinear structure of GWA data, we cannot reuse the ordinary least squares solution reliably [72]. Thus, we derive the general regularization

$$\widetilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\Sigma}\boldsymbol{\beta}\|^2, \qquad \text{subject to } (1-\alpha)\|\boldsymbol{\beta}\|_1 + \alpha\|\boldsymbol{\beta}\|_2^2 \leq t \text{ for some } t, \tag{5}$$

where, in addition to previous notation, the solution $\widetilde{\boldsymbol{\beta}}$ is used to denote the regularized solution of the observed GWA effect sizes $\widehat{\boldsymbol{\beta}}$; and $\| \bullet \|_1$ and $\| \bullet \|_2^2$ denote L_1 and L_2 penalties, respectively. The term α

distinguishes the type of regularization used, and can be chosen to induce various degrees of shrinkage on the effect size estimators. Specifically, $\alpha = 0$ corresponds to the "Least Absolute Shrinkage and Selection Operator" or LASSO solution [23], $\alpha = 1$ equates to Ridge Regression [25], while $0 < \alpha < 1$ results in the Elastic Net [24]. The LASSO solution forces some inflated coefficients to be zero; while the Ridge shrinks the magnitudes of all coefficients but does not set any of them to be exactly zero. Intuitively, the LASSO will create a regularized set of effect sizes where causal SNPs have larger effects, interactive SNPs have small-yet-nonzero effects, and non-causal SNPs are set zero. It has been suggested that the L_1-penalty can suffer from a lack of stability [73]. Therefore, in the main text, we also highlighted gene-ε using the Elastic Net (with $\alpha = 0.5$). The Elastic Net is a convex combination of the LASSO and Ridge penalties, but still results in distinguishable sets of causal, interactive, and non-causal SNPs. Results comparing each of the gene-ε regularization implementations are given in the main text (Fig. 3) and Supplementary Notes (Supplementary Figs. 2-16 and Supplementary Tables 1-19 and 26-28).

Recalibration of the GWA Null Hypothesis

The gene-ε framework assumes that true SNP-level effects for complex traits are jointly distributed as

$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}), \qquad \boldsymbol{\Omega}_{jj} = \sigma_j^2, \qquad \boldsymbol{\Omega}_{jl} = \sigma_j \rho(\mathbf{x}_j, \mathbf{x}_l)\sigma_l, \tag{6}$$

where $\boldsymbol{\Omega} = \mathbf{H}^{1/2}\boldsymbol{\Sigma}\mathbf{H}^{1/2}$ is a $J \times J$ variance-covariance matrix, and $\mathbf{H} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_J^2)$ is a $J \times J$ diagonal matrix with elements σ_j^2 denoting the proportion of narrow-sense heritability h^2 contributed by the j-th SNP (with the constraint $\sum \sigma_j^2 = h^2$). The main innovation of gene-ε is to treat interactive SNPs with ε-genic effects as non-causal. This leads to the recalibrated GWA null hypothesis that assumes every SNP makes a negligible contribution to the genotypic variation of the trait of interest. Formally, we write $H_0 \colon \beta_j \approx 0$ and use Eq. (6) to derive the (limiting) joint null distribution as

$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_0), \qquad \boldsymbol{\Omega}_{0,jj} = \sigma_\varepsilon^2, \qquad \boldsymbol{\Omega}_{0,jl} = \sigma_\varepsilon \rho(\mathbf{x}_j, \mathbf{x}_l)\sigma_\varepsilon, \tag{7}$$

where $\boldsymbol{\Omega}_0 = \mathbf{H}_0^{1/2}\boldsymbol{\Sigma}\mathbf{H}_0^{1/2} = \sigma_\varepsilon^2\boldsymbol{\Sigma}$ with $\mathbf{H}_0 = \mathrm{diag}(\sigma_\varepsilon^2, \ldots, \sigma_\varepsilon^2)$. We refer to σ_ε^2 as the "ε-genic threshold". Intuitively, σ_ε^2 represents the maximum proportion of variance explained (PVE) by an interactive or non-causal SNP. To this end, we may equivalently state the recalibrated GWA null hypothesis as $H_0 \colon \sigma_j^2 \leq \sigma_\varepsilon^2$. To infer the proportion of PVE contributed by each SNP, gene-ε takes an empirical Bayes approach by fitting a K-mixture of normal distributions over the regularized effect size estimates $\widetilde{\beta}_j$ [18], such that

$$\widetilde{\beta}_j \,|\, z_j = k \sim \mathcal{N}(0, \sigma_k^2), \tag{8}$$

where $z_j \in \{1, \ldots, K\}$ is a latent variable representing the mixture membership for the j-th SNP. When summing over all components, Eq. (8) corresponds to the following marginal distribution

$$\widetilde{\beta}_j \sim \sum_{k=1}^{K} \pi_k \mathcal{N}(0, \sigma_k^2), \tag{9}$$

where π_k is a mixture weight representing the marginal (unconditional) probability that a randomly selected SNP belongs to the k-th component, with $\sum_k \pi_k = 1$. The above mixture allows for distinct clusters of nonzero effects through K different variance components (σ_k^2, $k = 1, \ldots, K$) [18]. Here, we consider sequential fractions $(\pi_1, \ldots, \pi_K)$ of SNPs to correspond to distinctly smaller effects ($\sigma_1^2 > \cdots > \sigma_K^2 = 0$) [18]. Intuitively, causal SNPs will appear in the first set of fraction groups, while SNPs with ε-genic or non-causal effects will belong to the latter groups. Since our objective is to find the core genes underlying complex traits, results in the current study are based on considering an ε-genic threshold $\sigma_\varepsilon^2 = \sigma_2^2$. This means that we consider SNP j to be causal if it is grouped in the largest fraction (i.e., the

alternative $H_A : \sigma_j^2 > \sigma_2^2$), and interactive or non-causal otherwise (i.e., the null $H_0 : \sigma_j^2 \leq \sigma_2^2$). Given Eqs. (8) and (9), we fit the mixtures for all J SNPs by aiming to maximize the joint log-likelihood

$$\log p(\widetilde{\boldsymbol{\beta}} \,|\, \boldsymbol{\Theta}) = \sum_{j=1}^{J} \log p(\widetilde{\beta}_j \,|\, \boldsymbol{\Theta}) = \sum_{j=1}^{J} \log \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(0, \sigma_k^2) \right\}, \tag{10}$$

where $\boldsymbol{\Theta} = (\pi_1, \ldots, \pi_K, \sigma_1^2, \ldots, \sigma_K^2)$ is the complete set of parameters for the mixture model. In the gene-ε framework, estimation is done using an expectation-maximization (EM) algorithm to maximize the joint distribution in Eq. (10). Here, we compute the probability that each SNP is a member of the K different components. The number of components K is chosen via a Bayesian Information Criterion (BIC). Details of this algorithm are given in the first section of the Supplementary Notes.

Hypothesis Testing for the Detection of Core Genes

We now formalize the statistical test for identifying core genes using the recalibrated GWA null hypothesis under the gene-ε framework. This test statistic is based on a quadratic form using GWA summary statistics, and is a commonly used approach for generating gene-level association statistics for complex traits. Let gene (or genomic region) g represent a known set of SNPs $j \in \mathcal{J}_g$; for example, $\mathcal{J}_g$ may include SNPs within the boundaries of g and/or within its corresponding regulatory region. Here, we conformably partition the regularized GWA effect size estimates $\widetilde{\boldsymbol{\beta}}$ and define the gene-level test statistic

$$\widehat{Q}_g = \widetilde{\boldsymbol{\beta}}_g^{\intercal} \mathbf{A} \widetilde{\boldsymbol{\beta}}_g, \tag{11}$$

where $\mathbf{A}$ is an arbitrary symmetric and positive semi-definite weight matrix, which we set to be the identity $\mathbf{I}$ for all analyses in the current study. Indeed, similar quadratic forms have been implemented to assess the enrichment of mutations at the gene level [7, 12] and across general SNP-sets [9, 20, 28, 58]. A key feature of the gene-ε framework is to test the statistics in Eq. (11) against a recalibrated gene-level association null hypothesis $H_0 : \widehat{Q}_g \approx 0$. Due to the asymptotic normality assumption for each SNP effect (Eq. (6)), $\widehat{Q}_g$ is theoretically assumed to follow a mixture of chi-square distributions,

$$\widehat{Q}_g \sim \sum_{j=1}^{|\mathcal{J}_g|} \lambda_j \chi_{1,j}^2, \tag{12}$$

where $|\mathcal{J}_g|$ denotes the cardinality of the set of SNPs $\mathcal{J}_g$ and $\chi_{1,j}^2$ are standard chi-square random variables with one degree of freedom. Most importantly, $(\lambda_1, \ldots, \lambda_{|\mathcal{J}_g|})$ are the eigenvalues of the matrix

$$\boldsymbol{\Omega}_{0,g}^{1/2} \mathbf{A} \boldsymbol{\Omega}_{0,g}^{1/2}, \tag{13}$$

where $\boldsymbol{\Omega}_{0,g} = \mathbf{H}_{0,g}^{1/2} \boldsymbol{\Sigma}_g \mathbf{H}_{0,g}^{1/2} = \sigma_\varepsilon^2 \boldsymbol{\Sigma}_g$ is the variance-covariance structure of gene g with SNPs only emitting interactive and non-causal effects under the recalibrated GWA null hypothesis (Eq. (7)). Several approximate and exact methods have been suggested to obtain P-values under a mixture of chi-square distributions. In this study, we use Imhof's method [26].

In the main text, we highlight some of the additional features of the gene-ε gene-level association test statistic. First, the expected enrichment for trait-associated mutations in a given gene is equal to the heritability explained by the SNPs contained in said gene. More formally, consider the expansion of Eq. (11) derived from the expectation of quadratic forms,

$$\mathbb{E}[\widehat{Q}_g] = \mathrm{tr}(\boldsymbol{\Omega}_g \mathbf{A}) = \sum_r \sum_s (\boldsymbol{\Omega}_g \circ \mathbf{A})_{rs} = h_g^2, \tag{14}$$

where $\mathbf{\Omega}_g = \mathbf{H}_g^{1/2}\mathbf{\Sigma}_g\mathbf{H}_g^{1/2}$ is the variance-covariance matrix of the SNPs within gene g, $\text{tr}(\bullet)$ denotes the matrix trace function, $(\mathbf{\Omega}_g \circ \mathbf{A})$ represents the Hadamard (i.e. element-wise) product between the two matrices, r and s are row and column indices, and h_g^2 denotes the heritability contributed by gene g. When $\mathbf{A} = \mathbf{I}$ (as in the current study), the gene-ε recalibrated hypothesis test for identifying core genes is based on the LD map between SNPs within the gene of interest, scaled by the individual SNP contributions to the narrow-sense heritability. Alternatively, one could choose to re-weight these contributions by specifying $\mathbf{A}$ otherwise [12, 20, 74–76]. For example, if SNP j has a small effect size but is known to be functionally associated with the trait of interest, then increasing $\mathbf{A}_{jj}$ will reflect this knowledge. Specific weight functions have also been suggested for dealing with rarer variants [9, 28, 58]. The variances of the gene-ε gene-level statistics are given by

$$\mathbb{V}[\widehat{Q}_g] = 2\text{tr}\left(\mathbf{A}\mathbf{\Omega}_g\mathbf{A}\mathbf{\Omega}_g\right) \tag{15}$$

and can be used to derive measures of uncertainty regarding the strength of gene associations (e.g., building confidence intervals).

Simulation Studies

We used a simulation scheme to generate SNP-level summary statistics for GWA studies. First, we randomly select core genes and assume that complex traits (under various genetic architectures) are generated via a linear model

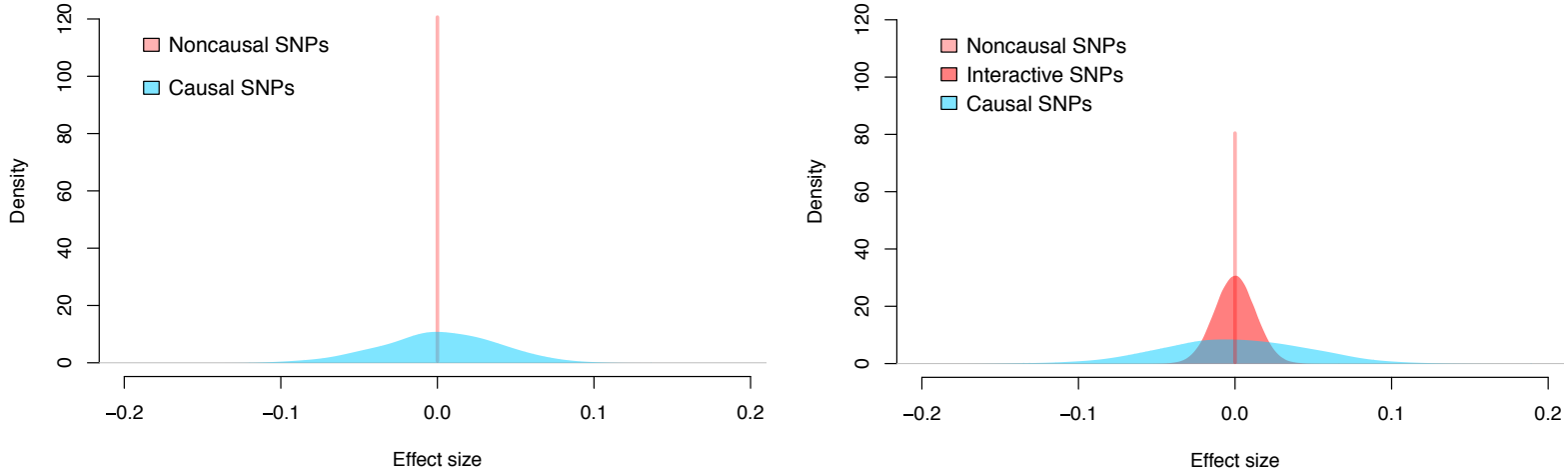
$$\mathbf{y} = \mathbf{W}\mathbf{b} + \sum_{c\in\mathcal{C}} \mathbf{x}_c\beta_c + \mathbf{e}, \qquad \mathbf{e} \sim \mathcal{N}(\mathbf{0}, \tau^2\mathbf{I}), \tag{16}$$

where $\mathbf{y}$ is an N-dimensional vector containing all the phenotypes; $\mathcal{C}$ represents the set of causal SNPs contained within the core genes; $\mathbf{x}_c$ is the genotype for the c-th causal SNP encoded as 0, 1, or 2 copies of a reference allele; β_c is the additive effect size for the c-th SNP; $\mathbf{W}$ is an $N \times M$ matrix of covariates representing additional population structure (e.g., the top ten principal components from the genotype matrix) with corresponding fixed effects $\mathbf{b}$; and $\mathbf{e}$ is an N-dimensional vector of environmental noise. The phenotypic variance is assumed $\mathbb{V}[\mathbf{y}] = 1$. The effect sizes of SNPs in core genes are randomly drawn from standard normal distributions and then rescaled so they explain a fixed proportion of the narrow-sense heritability $\mathbb{V}[\sum \mathbf{x}_c\beta_c] = h^2$. The covariate coefficients are also drawn from standard normal distributions and then rescaled such that $\mathbb{V}[\mathbf{W}\mathbf{b}] + \mathbb{V}[\mathbf{e}] = (1 - h^2)$. GWA summary statistics are then computed by fitting a single-SNP univariate linear model via ordinary least squares (OLS): $\widehat{\beta}_j = (\mathbf{x}_j^\intercal\mathbf{x}_j)^{-1}\mathbf{x}_j^\intercal\mathbf{y}$ for every SNP in the data $j = 1, \ldots J$. These effect size estimates, along with an LD matrix $\widehat{\mathbf{\Sigma}}$ computed directly from the full $N \times J$ genotype matrix $\mathbf{X}$, are given to gene-ε . We also retain standard errors and P-values for implementation of the competing methods (VEGAS, PEGASUS, RSS, SKAT, and MAGMA). Given different model parameters, we simulate data mirroring a wide range of genetic architectures (Supplementary Notes).

Data Availability

Source code implementing gene-ε and tutorials are available on line (`https://github.com/ramachandran-lab/genee`). Links to the other competing methods and UK Biobank data are also provided (See URLs).

Figures and Tables



(a) Traditional GWA Hypothesis

(b) Recalibrated GWA Hypothesis

Figure 1. Illustration of null hypothesis assumptions for the distribution of GWA SNP-level effect sizes according to different views on underlying genetic architectures. The effect sizes of "non-causal" (pink), "intermediate" (red), and "causal" (blue) SNPs were drawn from normal distributions with successively larger variances. **(a)** The traditional GWA model of complex traits simply assumes SNPs are causal or non-causal. Under the corresponding null hypothesis, causal SNPs are more likely to emit large effect sizes while non-causal SNPs will have effect sizes of zero. When there are many causal variants, we refer to the traits as polygenic. **(b)** Under our recalibrated GWA model, there are causal SNPs, non-causal SNPs, and interactive SNPs. We propose a multi-component framework (see also [18]), in which interactive SNPs can emit different levels of statistical signals based on (*i*) different degrees of connectedness (e.g., through linkage disequilibrium), or (*ii*) its regulated gene interacts with a true core gene. While causal SNPs are still more likely to emit large effect sizes than SNPs in the other categories, interactive SNPs can emit intermediate effect sizes. Here, our goal here is to classify interactive SNPs as non-causal.

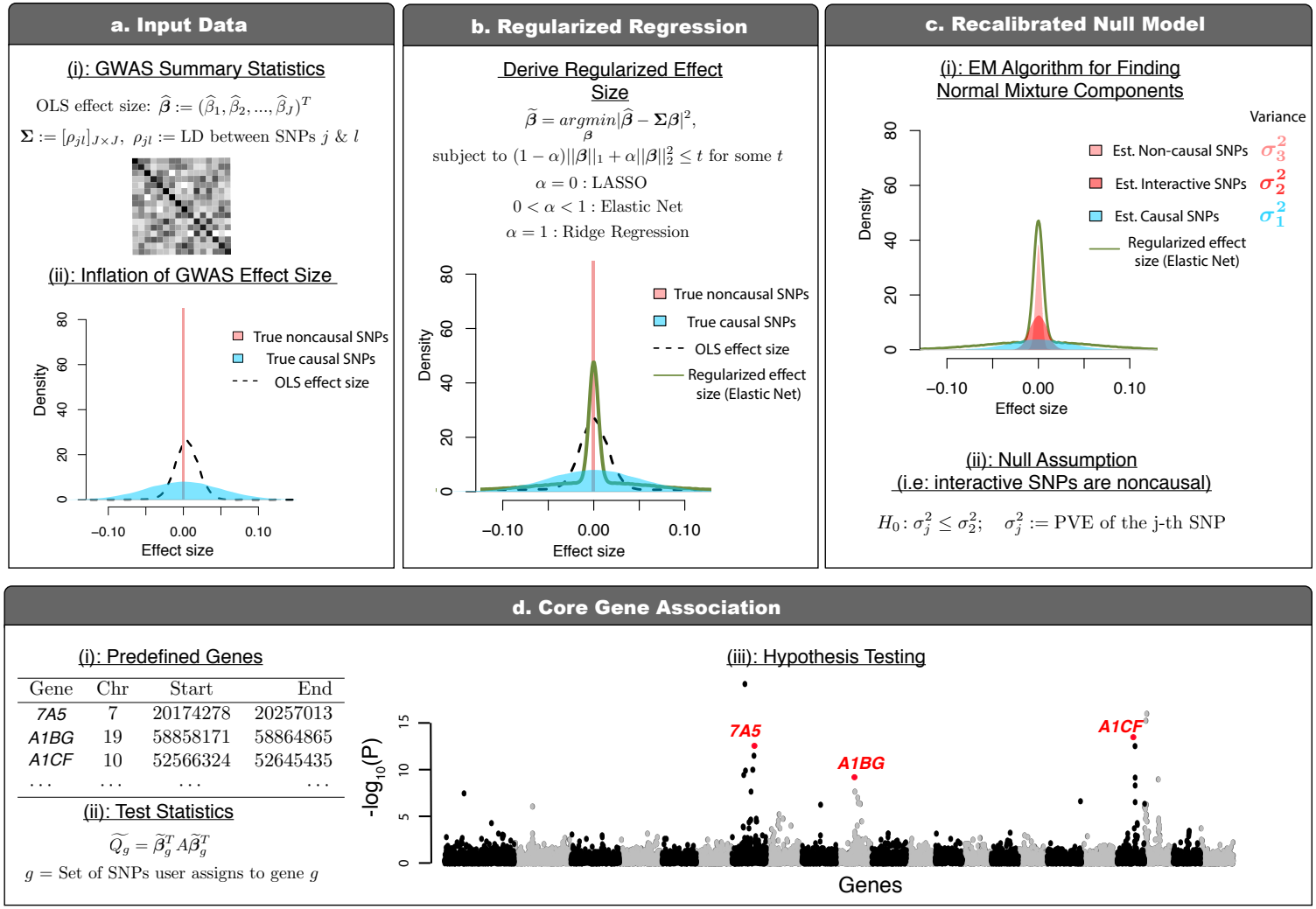
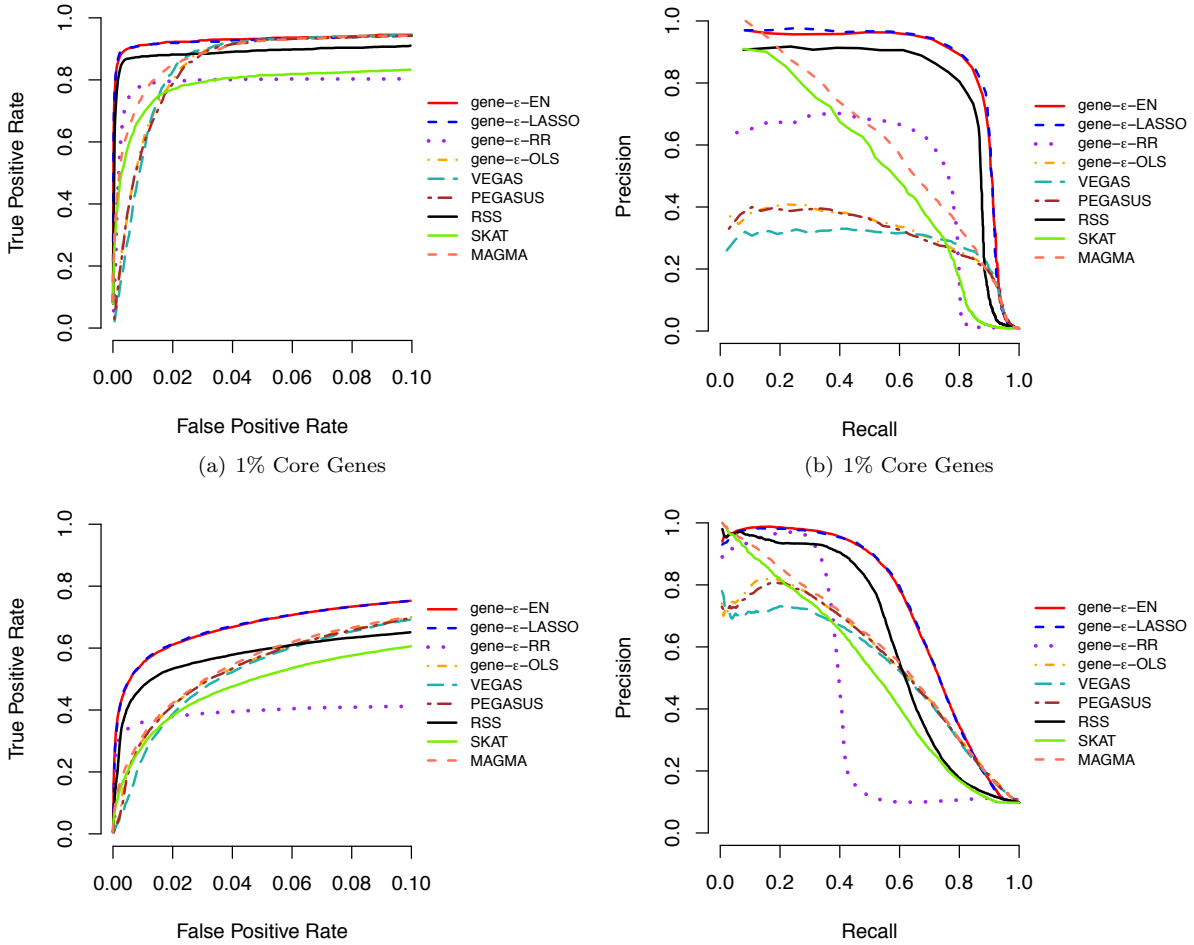


Figure 2. Schematic overview of gene-ε: our new gene-level association approach modeling ε-genic effects. **(a)** gene-ε takes SNP-level GWA marginal effect sizes (OLS estimates $\widehat{\boldsymbol{\beta}}$) and a linkage disequilibrium (LD) matrix ($\boldsymbol{\Sigma}$) as input. It is well-known that OLS effect size estimates are inflated due to LD (i.e., correlation structures) among genome-wide genotypes. **(b)** gene-ε first uses its inputs to derive regularized effect size estimators ($\widetilde{\boldsymbol{\beta}}$) through shrinkage methods (LASSO, Elastic Net and Ridge Regression; we explore performance of each solution under a variety of simulated trait architectures, Supplementary Notes and Figures). Marginally, $\widetilde{\beta}_j \sim \mathcal{N}(0, \sigma_j^2)$, where σ_j^2 denotes the proportion of narrow-sense heritability h^2 contributed by the j-th SNP. **(c)** A unique feature of gene-ε is that it treats interactive SNPs as non-causal. gene-ε assumes a recalibrated (SNP-level) null hypothesis $H_0: \sigma_j^2 \leq \sigma_\varepsilon^2$, where σ_ε^2 is the ε-genic threshold and represents the maximum proportion-of-variance-explained (PVE) by an interactive or non-causal SNP. To infer σ_j^2, gene-ε fits a K-mixture of normal distributions over the regularized effect sizes with successively smaller variances ($\sigma_1^2 > \cdots > \sigma_K^2$; with $\sigma_K^2 = 0$). In this study, we assume that causal SNPs will appear in the first set, while non-causal SNPs appear in the last set. By definition, the ε-genic threshold is then $\sigma_\varepsilon^2 = \sigma_2^2$. **(d)** Lastly, gene-$\varepsilon$ computes gene-level association test statistics using quadratic forms, and estimates corresponding P-values using Imhof's method. For more details, see Methods.



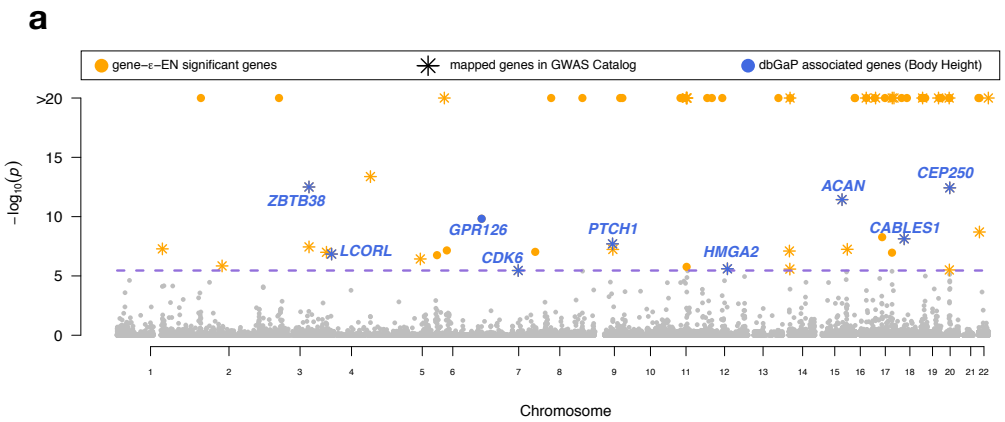
(a) 1% Core Genes

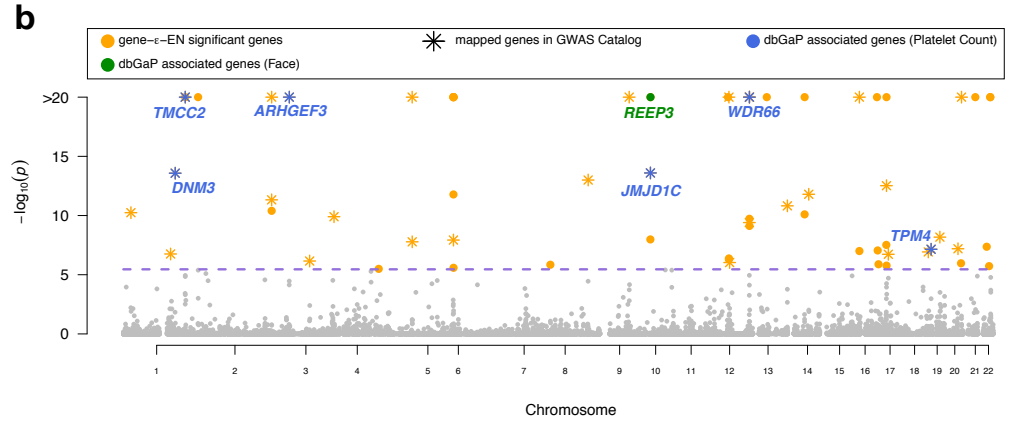
(b) 1% Core Genes

(c) 10% Core Genes

(d) 10% Core Genes

Figure 3. Receiver operating characteristic (ROC) and precision-recall curves comparing the performance of gene-ε and competing approaches in simulations ($N = 10,000$; $h^2 = 0.6$). We simulate complex traits under different genetic architectures and GWA study scenarios, varying the following parameters: narrow sense heritability, proportion of core genes, and sample size (Supplementary Notes). Here, the sample size $N = 10,000$ and the narrow-sense heritability $h^2 = 0.6$. We compute standard GWA SNP-level effect sizes (estimated using ordinary least squares). Results for gene-ε are shown with LASSO (blue), Elastic Net (EN; red), and Ridge Regression (RR; purple) regularizations. We also show the results of gene-ε without regularization to illustrate the importance of this step (labeled OLS; orange). We further compare gene-ε with five existing methods: PEGASUS (brown) [12], VEGAS (teal) [7], the Bayesian approach RSS (black) [14], SKAT (green) [20], and MAGMA (peach) [10]. **(a, c)** ROC curves show power versus false positive rate for each approach of sparse (1% core genes) and polygenic (10% core genes) architectures, respectively. Note that the upper limit of the x-axis has been truncated at 0.1. **(b, d)**. Precision-Recall curves for each method applied to the simulations. Note that, in the sparse case (1% core genes), the top ranked genes are always true positives, and therefore the minimal recall is not 0. All results are based on 100 replicates.





c

	p value	*q* value	Odds.ratio	Combined score	# of sig. genes in dbGaP
Body Height	9.65e-06	3.33e-03	6.49	74.99	9
Menopause	5.55e-03	9.58e-01	17.92	93.07	2
Prion Diseases	2.49e-02	1.00	39.68	146.49	1
Intercellular Adhesion Molecule-1	4.23e-02	1.00	23.15	73.18	1
Smoking	4.23e-02	1.00	23.15	73.18	1
Menarche	8.62e-02	1.00	11.11	27.23	1
Hippocampus	9.94e-02	1.00	9.58	22.12	1
Respiratory Function Tests	1.03e-01	1.00	3.68	8.37	2
Brain	1.15e-01	1.00	8.17	17.64	1
Sleep	1.19e-01	1.00	7.94	16.92	1

d

	p value	*q* value	Odds.ratio	Combined score	# of sig. genes in dbGaP
Platelet Count	1.93e-08	6.67e-06	34.72	616.70	6
Face	6.02e-05	1.04e-02	156.25	1518.33	2
Behcet Syndrome	8.94e-04	1.03e-01	15.89	111.54	3
Macular Degeneration	6.42e-03	5.54e-01	7.94	40.11	3
Hearing Loss	9.57e-03	6.60e-01	104.17	484.29	1
Erythrocytes	1.21e-02	6.93e-01	12.02	53.10	2
Sclerosis	1.90e-02	9.39e-01	52.08	206.29	1
Smoking	3.77e-02	1.00	26.04	85.34	1
Body Mass Index	5.13e-02	1.00	2.86	8.50	4
Natriuretic Peptide, Brain	5.61e-02	1.00	17.36	50.02	1

Figure 4. Gene-level association results from applying gene-ε to body height (panels a and c) and mean platelet volume (MPV; panels b and d), assayed in European-ancestry individuals in the UK Biobank. Body height has been estimated to have a narrow-sense heritability h^2 in the range of 0.45 to 0.80 [6, 31–39]; while, MPV has been estimated to have h^2 between 0.50 and 0.70 [33, 34, 77]. Manhattan plots of gene-ε gene-level association P-values using Elastic Net regularized effect sizes for **(a)** body height and **(b)** MPV. The purple dashed line indicates a log-transformed Bonferroni-corrected significance threshold ($P = 3.49 \times 10^{-6}$ correcting for 14,322 autosomal genes analyzed). We color code all significant genes identified by gene-ε in orange, and annotate genes overlapping with the database of Genotypes and Phenotypes (dbGaP). In **(c)** and **(d)**, we conduct gene set enrichment analysis using Enrichr [46, 78] to identify dbGaP categories enriched for significant gene-level associations reported by gene-ε. We highlight categories with Q-values (i.e., false discovery rates) less than 0.05 and annotate corresponding genes in the Manhattan plots in **(a)** and **(b)**, respectively. For height, the only significant dbGAP category is "Body Height", with 9 of the genes identified by gene-ε appearing in this category. For MPV, the two significant dbGAP categories are "Platelet Count" and "Face" — the first of which is directly connected to trait [57, 79, 80].

Trait	Gene	Chr	gene-ε P-Value	Rank	h_g^2	Post. Prob.	Biological Relevance to Trait	Ref(s)
Height	*EZH2*	7	9.34×10^{-8}	61	7.23×10^{-3}	1.000	Associated with diseases Adamantinoma of Long Bone and Weaver Syndrome (characterized by rapid growth).	[81]
Height	*C17orf42*	17	5.38×10^{-9}	52	4.54×10^{-3}	1.000	Known as the transcription elongation factor of mitochondria (TEFM) which regulates transcription and can affect body height.	[82]
Height	*KISS1R*	19	1×10^{-20}	1*	5.27×10^{-4}	0.970	Associated with disorders of puberty and final height.	[83]
BMI	*ZC3H4*	19	1.62×10^{-14}	20	7.84×10^{-3}	1.000	BMI-inducer known to be associated with adiposity and obesity.	[84–87]
BMI	*PTOV1*	19	1×10^{-20}	1*	2.26×10^{-3}	0.990	Found to be overexpressed in prostate adenocarcinomas which can be induced by obesity.	[88]
BMI	*FBXO45*	3	6.52×10^{-7}	23	1.82×10^{-3}	0.029	Reported to be involved in children syndromic obesity.	[89]
MCV	*SLC24A1*	15	1.74×10^{-7}	50	4.66×10^{-3}	0.140	Encoded protein is involved in glucose transportation pathway and MCV is reported to be associated with glucose level.	[88]
MCV	*PDX1*	13	1×10^{-20}	1*	2.31×10^{-4}	0.019	Associated with Glycated hemoglobin which is affected by MCV	[90]
MCV	*RHOD*	11	1×10^{-20}	1*	3.35×10^{-4}	0.002	Associated with Wiskott-Aldrich Syndrome which is characterized by abnormal immune system function (immune deficiency) and a reduced ability to form blood clots.	[88, 91]
MPV	*C1orf150*	1	1×10^{-20}	1*	3.44×10^{-2}	1.000	Known as *GCSAML* which is involved with germinal center signaling and differentiation of mature B cells that mutually activate platelets.	[47–49]
MPV	*KIAA0922*	4	3.20×10^{-6}	64	7.17×10^{-3}	1.000	Known as *TMEM131L* which is associated with canonical Wnt signaling and can effect platelet formation.	[92, 93]
MPV	*TPT1*	13	1×10^{-20}	1*	3.25×10^{-4}	0.051	mRNA expression is identified in platelets.	[88]
PLC	*C1orf150*	1	1×10^{-20}	1*	2.51×10^{-2}	1.000	Known as *GCSAML* which is involved with germinal center signaling and differentiation of mature B cells that mutually activate platelets.	[47–49]
PLC	*PSMD2*	3	1.42×10^{-9}	29	7.40×10^{-3}	1.000	Also known as the 26S proteasome which is found to be important for platelet production.	[88]
PLC	*APOB48R*	16	1×10^{-20}	1*	1.36×10^{-3}	0.003	Involved in Lipoprotein metabolism pathway which can affect platelet.	[88]
WHR	*TFAP2B*	6	3.92×10^{-7}	21	3.60×10^{-3}	1.000	Dietary protein associated with weight maintenance.	[86, 94]
WHR	*WDR68*	17	1.05×10^{-7}	20	1.10×10^{-3}	0.990	Also known as *DCAF7* which has been shown to bind Huntingtin-associated protein 1 (HAP1) and affect weight.	[95]
WHR	*MLL*	11	8.14×10^{-8}	19	2.43×10^{-3}	0.940	Orthologous gene in mice that affects skeleton, body size, and growth.	[86, 96–98]

Table 1. Top three novel candidate core genes reported by gene-ε for the six quantitative traits studied in the UK Biobank (using imputed genotypes with gene boundaries defined by the NCBI's RefSeq database in the UCSC Genome Browser [27]). We call these novel candidate core genes because they are not listed as being associated with the trait of interest in either the GWAS catalog or dbGaP, and they have top posterior enrichment probabilities with the trait using RSS analysis. Each gene is annotated with past functional studies that link them to the trait of interest. We also report each gene's overall trait-specific significance rank (out of 14,322 autosomal genes analyzed for each trait), as well as their heritability estimates from gene-ε using Elastic Net to regularize GWA SNP-level effect size estimates. The traits are: height; body mass index (BMI); mean corpuscular volume (MCV); mean platelet volume (MPV); platelet count (PLC); and waist-hip ratio (WHR). *: Multiple genes were tied for this ranking.

References

1. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era–concepts and misconceptions. Nat Rev Genet. 2008;9(4):255–266.

2. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747–753. Available from: https://www.ncbi.nlm.nih.gov/pubmed/19812666.

3. Visscher PM, Brown MA, McCarthy MI, Yang J. Five Years of GWAS Discovery. Am J Hum Genet. 2012;90(1):7–24. Available from: http://www.sciencedirect.com/science/article/pii/S0002929711005337.

4. Boyle EA, Li YI, Pritchard JK. An expanded view of complex traits: from polygenic to omnigenic. Cell. 2017;169(7):1177–1186.

5. Wray NR, Wijmenga C, Sullivan PF, Yang J, Visscher PM. Common disease is more complex than implied by the core gene omnigenic model. Cell. 2018;173(7):1573–1580. Available from: https://doi.org/10.1016/j.cell.2018.05.051.

6. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010;42(7):565–569.

7. Liu JZ, Mcrae AF, Nyholt DR, Medland SE, Wray NR, Brown KM, et al. A versatile gene-based test for genome-wide association studies. Am J Hum Genet. 2010;87(1):139–145.

8. Carbonetto P, Stephens M. Integrated enrichment analysis of variants and pathways in genome-wide association studies indicates central role for IL-2 signaling genes in type 1 diabetes, and cytokine signaling genes in Crohn's disease. PLoS Genet. 2013;9(10):e1003770–. Available from: https://doi.org/10.1371/journal.pgen.1003770.

9. Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. Sequence kernel association tests for the combined effect of rare and common variants. Am J Hum Genet. 2013;92(6):841–853. Available from: http://www.sciencedirect.com/science/article/pii/S0002929713001766.

10. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. PLOS Comput Biol. 2015;11(4):e1004219–. Available from: https://doi.org/10.1371/journal.pcbi.1004219.

11. Lamparter D, Marbach D, Rueedi R, Kutalik Z, Bergmann S. Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. PLOS Comput Biol. 2016;12(1):e1004714–. Available from: https://doi.org/10.1371/journal.pcbi.1004714.

12. Nakka P, Raphael BJ, Ramachandran S. Gene and network analysis of common variants reveals novel associations in multiple complex diseases. Genetics. 2016;204(2):783–798. Available from: http://www.genetics.org/content/204/2/783.abstract.

13. Wang M, Huang J, Liu Y, Ma L, Potash JB, Han S. COMBAT: a combined association test for genes using summary statistics. Genetics. 2017;207(3):883–891.

14. Zhu X, Stephens M. Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. Nat Comm. 2018;9(1):4361.

15. Zhou X, Carbonetto P, Stephens M. Polygenic modeling with Bayesian sparse linear mixed models. PLoS Genet. 2013;9(2):e1003264.

16. Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed-model association methods. Nat Genet. 2014;46(2):100–106.

17. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, of the Psychiatric Genomics Consortium SWG, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015;47:291–295. Available from: `http://dx.doi.org/10.1038/ng.3211`.

18. Zhang Y, Qi G, Park JH, Chatterjee N. Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. Nat Genet. 2018;50(9):1318–1326.

19. Liu X, Li YI, Pritchard JK. Trans effects on gene expression can drive omnigenic inheritance. Cell. 2019;177(4):1022–1034. Available from: `https://www.sciencedirect.com/science/article/pii/S0092867419304003`.

20. Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, et al. Powerful SNP-set analysis for case-control genome-wide association studies. Am J Hum Genet. 2010;86(6):929–942.

21. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562(7726):203–209. Available from: `https://doi.org/10.1038/s41586-018-0579-z`.

22. Stephens M. False discovery rates: a new deal. Biostatistics. 2017;18(2):275–294. Available from: `http://dx.doi.org/10.1093/biostatistics/kxw041`.

23. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol. 1996;58(1):267–288.

24. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Series B Stat Methodol. 2005;67(2):301–320.

25. Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics. 1970;12(1):55–67.

26. Imhof JP. Computing the distribution of quadratic forms in normal variables. Biometrika. 1961;48(3/4):419–426. Available from: `http://www.jstor.org/stable/2332763`.

27. Pruitt KD, Tatusova T, Maglott DR. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2005;33(Database issue):D501–4.

28. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet. 2012;91(2):224–237. Available from: `http://www.sciencedirect.com/science/article/pii/S0002929712003163`.

29. Zhu X, Stephens M. Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. Ann Appl Stat. 2017;11(3):1561–1592. Available from: `https://projecteuclid.org:443/euclid.aoas/1507168840`.

30. Barbieri MM, Berger JO. Optimal predictive model selection. Ann Statist. 2004;32(3):870–897. Available from: `http://projecteuclid.org/euclid.aos/1085408489`.

31. Zaitlen N, Kraft P, Patterson N, Pasaniuc B, Bhatia G, Pollack S, et al. Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. PLoS Genet. 2013;9(5):e1003520–. Available from: https://doi.org/10.1371/journal.pgen.1003520.

32. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet. 2014;46(11):1173–1186.

33. Heckerman D, Gurdasani D, Kadie C, Pomilla C, Carstensen T, Martin H, et al. Linear mixed model for heritability estimation that explicitly addresses environmental variation. Proc Natl Acad Sci U S A. 2016;113(27):7377–7382. Available from: http://www.pnas.org/content/113/27/7377.abstract.

34. Shi H, Kichaev G, Pasaniuc B. Contrasting the genetic architecture of 30 complex traits from summary association data. Am J Hum Genet. 2016;99(1):139–153. Available from: http://www.sciencedirect.com/science/article/pii/S0002929716301483.

35. Xia C, Amador C, Huffman J, Trochet H, Campbell A, Porteous D, et al. Pedigree- and SNP-associated genetics and recent environment are the major contributors to anthropometric and cardiometabolic trait variation. PLoS Genet. 2016;12(2):e1005804–. Available from: https://doi.org/10.1371/journal.pgen.1005804.

36. Ge T, Chen CY, Neale BM, Sabuncu MR, Smoller JW. Phenome-wide heritability analysis of the UK Biobank. PLoS Genet. 2017;13(4):e1006711–. Available from: https://doi.org/10.1371/journal.pgen.1006711.

37. Speed D, Cai N, The UCLEB Consortium, Johnson MR, Nejentsev S, Balding DJ. Reevaluation of SNP heritability in complex human traits. Nat Genet. 2017;49:986–992. Available from: https://doi.org/10.1038/ng.3865.

38. Marouli E, Graff M, Medina-Gomez C, Lo KS, Wood AR, Kjaer TR, et al. Rare and low-frequency coding variants alter human adult height. Nature. 2017;542(7640):186–190.

39. Wainschtein P, Jain DP, Yengo L, Zheng Z, TOPMed Anthropometry Working Group, Trans-Omics for Precision Medicine Consortium, et al. Recovery of trait heritability from whole genome sequence data. bioRxiv. 2019;p. 588020. Available from: http://biorxiv.org/content/early/2019/03/25/588020.abstract.

40. Goldstein DB. Common genetic variation and human traits. N Engl J Med. 2009;360(17):1696–1698.

41. Lello L, Avery SG, Tellier L, Vazquez AI, de los Campos G, Hsu SDH. Accurate Genomic Prediction of Human Height. Genetics. 2018;210(2):477–497. Available from: http://www.genetics.org/content/210/2/477.abstract.

42. Vattikuti S, Guo J, Chow CC. Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits. PLoS Genet. 2012;8(3):e1002637.

43. Yang J, Bakshi A, Zhu Z, Hemani G, Vinkhuyzen AA, Lee SH, et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat Genet. 2015;47(10):1114.

44. Robinson MR, English G, Moser G, Lloyd-Jones LR, Triplett MA, Zhu Z, et al. Genotype–covariate interaction effects and the heritability of adult body mass index. Nat Genet. 2017;49(8):1174.

45. Rothschild D, Weissbrod O, Barkan E, Kurilshikov A, Korem T, Zeevi D, et al. Environment dominates over host genetics in shaping human gut microbiota. Nature. 2018;555:210–215. Available from: https://doi.org/10.1038/nature25973.

46. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinform. 2013;14(1):128. Available from: https://doi.org/10.1186/1471-2105-14-128.

47. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518(7539):317–330. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25693563.

48. Eicher JD, Chami N, Kacprowski T, Nomura A, Chen MH, Yanek LR, et al. Platelet-Related Variants Identified by Exomechip Meta-analysis in 157,293 Individuals. Am J Hum Genet. 2016;99(1):40–55.

49. Iotchkova V, Huang J, Morris JA, Jain D, Barbieri C, Walter K, et al. Discovery and refinement of genetic loci associated with cardiometabolic risk using dense imputation maps. Nat Genet. 2016;48(11):1303–1312. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27668658.

50. Finberg KE, Heeney MM, Campagna DR, Aydinok Y, Pearson HA, Hartman KR, et al. Mutations in TMPRSS6 cause iron-refractory iron deficiency anemia (IRIDA). Nat Genet. 2008;40(5):569–571. Available from: https://www.ncbi.nlm.nih.gov/pubmed/18408718.

51. Andrews NC. Genes determining blood cell traits. Nat Genet. 2009;41:1161–1162. Available from: https://doi.org/10.1038/ng1109-1161.

52. Benyamin B, Ferreira MAR, Willemsen G, Gordon S, Middelberg RPS, McEvoy BP, et al. Common variants in TMPRSS6 are associated with iron status and erythrocyte volume. Nat Genet. 2009;41(11):1173–1175.

53. Chambers JC, Zhang W, Li Y, Sehmi J, Wass MN, Zabaneh D, et al. Genome-wide association study identifies variants in TMPRSS6 associated with hemoglobin levels. Nat Genet. 2009;41(11):1170–1172.

54. Soranzo N, Spector TD, Mangino M, Kühnel B, Rendon A, Teumer A, et al. A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. Nat Genet. 2009;41(11):1182–1190. Available from: https://www.ncbi.nlm.nih.gov/pubmed/19820697.

55. Ganesh SK, Zakai NA, van Rooij FJA, Soranzo N, Smith AV, Nalls MA, et al. Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. Nat Genet. 2009;41(11):1191–1198.

56. Li J, Glessner JT, Zhang H, Hou C, Wei Z, Bradfield JP, et al. GWAS of blood cell traits identifies novel associated loci and epistatic interactions in Caucasian and African-American children. Hum Mol Genet. 2013;22(7):1457–1464. Available from: https://www.ncbi.nlm.nih.gov/pubmed/23263863.

57. Astle WJ, Elding H, Jiang T, Allen D, Ruklisa D, Mann AL, et al. The allelic landscape of human blood cell trait variation and links to common complex disease. Cell. 2016;167(5):1415–1429. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27863252.

58. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011;89(1):82–93.

59. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet. 2014;95(1):5–23. Available from: http://www.sciencedirect.com/science/article/pii/S0002929714002717.

60. Zuk O, Schaffner SF, Samocha K, Do R, Hechter E, Kathiresan S, et al. Searching for missing heritability: designing rare variant association studies. Proc Natl Acad Sci U S A. 2014;111(4):E455–E464. Available from: http://www.pnas.org/content/111/4/E455.abstract.

61. Gazal S, Loh PR, Finucane HK, Ganna A, Schoech A, Sunyaev S, et al. Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations. Nat Genet. 2018;50(11):1600–1607. Available from: https://doi.org/10.1038/s41588-018-0231-8.

62. Wojcik G, Graff M, Nishimura KK, Tao R, Haessler J, Gignoux CR, et al. The PAGE Study: how genetic diversity improves our understanding of the architecture of complex traits. bioRxiv. 2018;p. 188094. Available from: http://biorxiv.org/content/early/2018/10/17/188094.abstract.

63. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet. 2019;51(4):584–591. Available from: https://doi.org/10.1038/s41588-019-0379-x.

64. GTEx Consortium. Genetic effects on gene expression across human tissues. Nature. 2017;550:204–213. Available from: https://doi.org/10.1038/nature24277.

65. Wu Y, Zeng J, Zhang F, Zhu Z, Qi T, Zheng Z, et al. Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. Nat Comm. 2018;9(1):918. Available from: https://doi.org/10.1038/s41467-018-03371-0.

66. Xue A, Wu Y, Zhu Z, Zhang F, Kemper KE, Zheng Z, et al. Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. Nat Comm. 2018;9(1):2941. Available from: https://doi.org/10.1038/s41467-018-04951-w.

67. Smemo S, Tena JJ, Kim KH, Gamazon ER, Sakabe NJ, Gomez-Marin C, et al. Obesity-associated variants within FTO form long-range functional connections with IRX3. Nature. 2014;507(7492):371–375.

68. Claussnitzer M, Dankel SN, Kim KH, Quon G, Meuleman W, Haugen C, et al. FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. N Engl J Med. 2015;373(10):895–907. Available from: https://doi.org/10.1056/NEJMoa1502214.

69. Lee SH, Wray NR, Goddard ME, Visscher PM. Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet. 2011;88(3):294–305. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3059431/.

70. Golan D, Lander ES, Rosset S. Measuring missing heritability: inferring the contribution of common variants. Proc Natl Acad Sci U S A. 2014;111(49):E5272–E5281. Available from: http://www.pnas.org/content/111/49/E5272.abstract.

71. Weissbrod O, Lippert C, Geiger D, Heckerman D. Accurate liability estimation improves power in ascertained case-control studies. Nat Meth. 2015;12:332–334. Available from: http://dx.doi.org/10.1038/nmeth.3285.

72. Wold S, Ruhe A, Wold H, Dunn W III. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. SIAM J Sci Comput. 1984;5(3):735–743.

73. Carvalho CM, Polson NG, Scott JG. The horseshoe estimator for sparse signals. Biometrika. 2010;97(2):465–480.

74. Chen Z, Lin T, Wang K. A powerful variant-set association test based on chi-square distribution. Genetics. 2017;207(3):903–910.

75. Zhongxue C, Yan L, Tong L, Qingzhong L, Kai W. Gene-based genetic association test with adaptive optimal weights. Genet Epidemiol. 2017;42(1):95–103. Available from: https://doi.org/10.1002/gepi.22098.

76. Zhou X. A unified framework for variance component estimation with summary statistics in genome-wide association studies. Ann Appl Stat. 2017;11(4):2027–2051. Available from: https://projecteuclid.org:443/euclid.aoas/1514430276.

77. Qayyum R, Snively BM, Ziv E, Nalls MA, Liu Y, Tang W, et al. A meta-analysis and genome-wide association study of platelet count and mean platelet volume in african americans. PLoS Genet. 2012;8(3):e1002491.

78. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016;44(W1):W90–W97. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27141961.

79. Lentaigne C, Freson K, Laffan MA, Turro E, Ouwehand WH, Consortium BB, et al. Inherited platelet disorders: toward DNA-based diagnosis. Blood. 2016;127(23):2814–2823. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27095789.

80. Mousas A, Ntritsos G, Chen MH, Song C, Huffman JE, Tzoulaki I, et al. Rare coding variants pinpoint genes that control human hematological traits. PLoS Genet. 2017;13(8):e1006925–. Available from: https://doi.org/10.1371/journal.pgen.1006925.

81. Gibson WT, Hood RL, Zhan SH, Bulman DE, Fejes AP, Moore R, et al. Mutations in EZH2 cause Weaver syndrome. Am J Hum Genet. 2012;90(1):110–118. Available from: https://www.cell.com/ajhg/fulltext/S0002-9297(11)00496-4.

82. Minczuk M, He J, Duch AM, Ettema TJ, Chlebowski A, Dzionek K, et al. TEFM (c17orf42) is necessary for transcription of human mtDNA. Nucleic Acids Res. 2011;39(10):4284–4299. Available from: https://www.ncbi.nlm.nih.gov/pubmed/21278163.

83. Carel JC, Lahlou N, Roger M, Chaussain JL. Precocious puberty and statural growth. Hum Reprod. 2004;10(2):135–147. Available from: https://academic.oup.com/humupd/article/10/2/135/617162.

84. Gong J, Schumacher F, Lim U, Hindorff LA, Haessler J, Buyske S, et al. Fine Mapping and Identification of BMI Loci in African Americans. Am J Hum Genet. 2013;93(4):661–671.

85. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015;518(7538):197–206. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25673413.

86. Dickinson ME, Flenniken AM, Ji X, Teboul L, Wong MD, White JK, et al. High-throughput discovery of novel developmental phenotypes. Nature. 2016;537:508–514. Available from: https://doi.org/10.1038/nature19356.

87. Baranski TJ, Kraja AT, Fink JL, Feitosa M, Lenzini PA, Borecki IB, et al. A high throughput, functional screen of human Body Mass Index GWAS loci using tissue-specific RNAi Drosophila melanogaster crosses. PLoS Genet. 2018;14(4):e1007222–. Available from: https://doi.org/10.1371/journal.pgen.1007222.

88. Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, Shmoish M, et al. GeneCards Version 3: the human gene integrator. Database. 2010;2010. Available from: https://academic.oup.com/database/article/doi/10.1093/database/baq020/407450.

89. Vuillaume ML, Naudion S, Banneau G, Diene G, Cartault A, Cailley D, et al. New candidate loci identified by array-CGH in a cohort of 100 children presenting with syndromic obesity. Am J Med Genet. 2014;164(8):1965–1975. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/ajmg.a.36587.

90. Wheeler E, Leong A, Liu CT, Hivert MF, Strawbridge RJ, Podmore C, et al. Impact of common genetic determinants of Hemoglobin A1c on type 2 diabetes risk and diagnosis in ancestrally diverse populations: A transethnic genome-wide meta-analysis. PLoS Med. 2017;14(9):e1002383. Available from: https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002383.

91. Linder S, Nelson D, Weiss M, Aepfelbacher M. Wiskott-Aldrich syndrome protein regulates podosomes in primary human macrophages. Proc Natl Acad Sci U S A. 1999;96(17):9648–9653. Available from: http://www.pnas.org/content/96/17/9648.abstract.

92. Steele BM, Harper MT, Macaulay IC, Morrell CN, Perez-Tamayo A, Foy M, et al. Canonical Wnt signaling negatively regulates platelet function. Proc Natl Acad Sci U S A. 2009;106(47):19836–19841.

93. Macaulay IC, Thon JN, Tijssen MR, Steele BM, MacDonald BT, Meade G, et al. Canonical Wnt signaling in megakaryocytes regulates proplatelet formation. Blood. 2013;121(1):188–196. Available from: http://www.bloodjournal.org/content/121/1/188.

94. Stocks T, Angquist L, Hager J, Charon C, Holst C, Martinez JA, et al. TFAP2B-dietary protein and glycemic index interactions and weight maintenance after weight loss in the DiOGenes trial. Hum Hered. 2013;75(2-4):213–219.

95. Xiang J, Yang S, Xin N, Gaertig MA, Reeves RH, Li S, et al. DYRK1A regulates Hap1–Dcaf7/WDR68 binding with implication for delayed growth in down syndrome. Proc Natl Acad Sci U S A. 2017;114(7):E1224–E1233. Available from: https://www.pnas.org/content/114/7/E1224.

96. Smith CM, Finger JH, Hayamizu TF, McCright IJ, Eppig JT, Kadin JA, et al. The mouse gene expression database (GXD): 2007 update. Nucleic Acids Res. 2006;35:D618–D623. Available from: https://academic.oup.com/nar/article/35/suppl_1/D618/1085755.

97. Bult CJ, Krupke DM, Begley DA, Richardson JE, Neuhauser SB, Sundberg JP, et al. Mouse Tumor Biology (MTB): a database of mouse models for human cancer. Nucleic Acids Res. 2014;43(D1):D818–D824. Available from: https://academic.oup.com/nar/article/43/D1/D818/2439858.

98. Smith CL, Blake JA, Kadin JA, Richardson JE, Bult CJ, Group MGD. Mouse Genome Database (MGD)-2018: knowledgebase for the laboratory mouse. Nucleic Acids Res. 2017;46(D1):D836–D842. Available from: https://academic.oup.com/nar/article/47/D1/D801/5165331.