Epsilon-Genic Effects Bridge the Gap Between Polygenic and Omnigenic 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 are often 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 model to include small-yet-nonzero associations emitted by interactive variants, which we refer to as "epsilon-genic" effects. gene-ε efficiently identifies core genes under a variety of simulated genetic architectures, achieving up to $\sim$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 gene sets 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]. Although the traditional genome-wide association (GWA) framework, in which individual SNPs are tested independently for association with a trait of interest, has largely been regarded as a failure when applied to complex traits [2,3,6], 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, which we refer to as "causal"; or (ii) be non-causal (see Fig. 1a). This classification is 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 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 instead that each SNP in a GWA dataset in truth belongs to one of *three* categories: (i) causal; (ii) statistically associated with the trait but not causal, which we refer to as "interactive"; and (iii) non-causal (Fig. 1b) [18]. As Boyle et al. [4] noted in their recent "omnigenic" model of complex traits, causal SNPs lie in core genes that directly influence the trait of interest, while interactive SNPs may lie in core genes or covary with causal SNPs due to various degrees of linkage disequilibrium (LD), spurious

correlations, or through trans-interactions [19]. For complex traits under the omnigenic model (Fig. 1b), interactive SNPs will emit intermediate statistical noise (in some cases, even appearing indistinguishable from causal SNPs), thereby confounding traditional GWA tests. We refer to this noise as "epsilon-genic effects" (denoted hereafter as "ε-genic effects").

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. The key motivation 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 a modified null hypothesis of *approximately* no association for interactive and non-causal SNPs ($H_0 \colon \beta_j \approx 0$) and

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

where ρ denotes the correlation coefficient between the j-th and l-th SNPs, and σ_j^2 denotes the proportion of h^2 contributed by j-th SNP (with the constraint $\sum \sigma_j^2 = h^2$). Equivalently, we can restate this same hypothesis testing framework using the variance-component-based null hypothesis $H_0 \colon \sigma_j^2 \leq \sigma^2$, where σ^2 represents the maximum proportion-of-variance-explained (PVE) that is explained 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. $0 \leq \sigma_j^2 \leq \sigma^2$ for every j-th SNP in that region). Core genes, on the other hand, are genes that contain at least one causal SNP (i.e. $\sigma_j^2 > \sigma^2$ for at least one SNP j in that region). By modeling ε-genic effects (i.e., different values of σ^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 to omnigenic 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 (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^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, 12, 14]. 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 [20].

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; Equation 5 in Methods). This shrinkage step reduces the inflation of effect sizes for SNPs covarying with causal SNPs [21], 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) [22], the Elastic Net solution (gene-ε-EN) [23], and Ridge Regression (gene-ε-RR) [24]. 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. Each successive Gaussian mixture component has distinctively smaller variances ($\sigma_1^2 > \cdots > \sigma_K^2$) with the K-th component fixed at $\sigma_K^2 = 0$. In this study, causal SNPs are assumed to be 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, SNPs with ε-genic effects then have PVEs that fall at or below the variance of the second component (i.e., $\sigma_j^2 \leq \sigma_2^2$ for the j-th SNP). Figs. 1b and 2c illustrate 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 using a quadratic form (Fig. 2d). 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 [25] 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 Note.

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 19 from individuals of European ancestry in the UK Biobank (Methods). Following quality control procedures, our simulations included 11,263 SNPs distributed across 1,303 genes (Supplementary Note). 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); the percentage of core genes (set to 1% or 10%); and number of causal intergenic SNPs (drawn uniformly from $\{5, 15, 30\}$). 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 five 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 are based on 100 different simulated runs for each parameter combination.

We compared the performance of gene-ε against that of three competing gene-level association or enrichment methods: VEGAS [7], PEGASUS [12], and RSS [14]. 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. Both 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 [26]. 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 [27]. Conceptually, VEGAS and PEGASUS assume null models under the traditional GWA framework, while RSS and gene-ε allow for traits to also have omnigenic architectures.

For all methods, we assess the power and false discovery rates (FDR) for identifying core genes at a Bonferroni-corrected or median probability threshold ($P = 0.05/1303$ genes $= 3.83 \times 10^{-5}$ or posterior

enrichment probability > 0.5; Supplementary Tables 1-8). 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. 2-8). While we find RSS to have the overall best tradeoff between true and false positive rates, RSS does not scale well for genome-wide analyses (Supplementary Table 9). 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. Computationally, gene-ε gains speed by directly assessing evidence for rejecting the null hypothesis, whereas RSS must compute the posterior probability of being a core gene. 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 compare 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., VEGAS and PEGASUS) is limited to scenarios where the phenotypic variance is dominated by just a few core genes with large effects (Supplementary Figs. 2 and 3). RSS, gene-ε-EN, and gene-ε-LASSO robustly outperform these methods for the other trait architectures (Fig. 3 and Supplementary Figs. 4-8). 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).

Characterizing Genetic Architecture of Quantitative Traits in the UK Biobank

We applied gene-ε to 410,172 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 349,468 European-ancestry individuals in the UK Biobank (Supplementary Note) [20]. After quality control, there were a total of 17,652 genes analyzed. First, we regressed the top five 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; findings with the other approaches can be found in Supplementary Figures and Tables.

While estimating ε-genic effects, gene-ε provides insight into to the genetic architecture of a trait (Supplementary Table 10). For example, past studies have shown human height to have a higher narrow-sense heritability (estimates ranging from 45-80%; [6, 28–36]). Using Elastic Net regularized effect sizes, gene-ε estimated approximately 68% of SNPs in the UK Biobank to be associated with height. This meant approximately 68% SNPs had marginal PVEs $\sigma_j^2 > 0$ (Methods). This number is similar to the Boyle et al. [4] result, which estimated 62% SNPs in the 1000 Genomes Project data to be associated with height. Additionally, gene-ε identified approximately 4.55% 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 1000 Genomes Project.

Compared to height, narrow-sense heritability estimates for BMI have been considered both high and low (estimates ranging from 25-60%; [28, 30, 31, 33, 34, 36–40]). 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 [39]. Here, our results suggest BMI to have a narrow-sense heritability similar to height, but with a slightly different distribution of null and non-null SNP effects. Specifically, we found BMI to have 66.7% associated SNPs and 5.35% 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 (Figs. 4a,b and Supplementary Figs. 9a-12a). 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 analyses, $P = 0.05/17652$ autosomal genes $= 2.83 \times 10^{-6}$; Supplementary Tables 11-16). As a validation step, we used the gene set enrichment analysis tool Enrichr [41] to identify dbGaP categories with an overrepresentation of core genes reported by gene-ε (Figs. 4c,d and Supplementary Figs. 9b-12b). A comparison of gene-level associations and gene set enrichments between the different gene-ε approaches are also listed (Supplementary Table 17).

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). For example, gene-ε reports *HTF9C* as having a significant gene-level association with BMI ($P = 5.13 \times 10^{-7}$). Although the protein encoded by *HTF9C* has an unknown function, *HTF9C* is orthologous to *TRMT2A* in mice, which is known to effect mouse lean body mass and metabolism [42].

Additionally, 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. The top four categories with Q-values smaller than 0.05 for BMI are "Macular Degeneration", "Cholesterol", "Stroke", and "Glucose". While there has been much debate about the exact relationship between BMI and age-related macular degeneration [43–46], there have been many studies verifying the connection between BMI and cholesterol [47–49], blood glucose levels [50], and risk of stroke [51, 52].

Importantly, gene-ε can also identify 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 9.65×10^{-14}. Previous studies have shown a rare variant *rs147110934* within this gene to significantly affect adult height [35]. 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 [26, 36, 53–56]. 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 causal 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 [9,26,53,54], comparing the efficacy of tagging causal variants in different ancestries [57,58], integrating GWA data with functional -omics data [59–61] — 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 modifying 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^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 to omnigenic spectrum.

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. 1-8, and Supplementary Table 9). 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.

There are several potential extensions for the gene-ε framework described here. For example, in the current study, we only focused on applying gene-ε to quantitative traits (Fig. 4 and Supplementary Figs. 9-12). 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 [62–64]. Another possible 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 nonsignificant P-values in the original GWA analysis (after Bonferroni-correction; Supplementary Tables 11-16). While this could be attributed to the improved 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, 21, 27].

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/`; Enrichr software, `http://amp.pharm.mssm.edu/Enrichr/`; Combined Association Test (COMBAT) software `https://cran.r-project.org/web/packages/COMBAT`; 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 supported in part by IDeA grant P20GM109035 (Center for Computational Biology of Human Disease) from the National Institute of General Medical Sciences (NIGMS) of the NIH and an Alfred P. Sloan Research Fellowship 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: GWA marginal effect size estimates $\widehat{\boldsymbol{\beta}}$, and an empirical linkage disequilibrium (i.e., variance-covariance) matrix $\boldsymbol{\Sigma}$. We assumed the following generative linear model for complex phenotypes

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

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 marker; $\boldsymbol{\beta}$ is a J-dimensional vector containing the additive effect sizes for an additional copy of the reference allele 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 assume that the genotype matrix (column-wise) and trait of interest have been mean-centered and standardized. 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 Equation (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 tested is $H_0 : \beta_j = 0$ for all $j = 1, \ldots, J$. 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 linkage disequilibrium (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 GWA effect size estimates and the true coefficient values is

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

where ρ denotes the correlation coefficient between SNPs $\mathbf{x}_j$ and $\mathbf{x}_l$, and $\boldsymbol{\Sigma}$ is the corresponding LD matrix. 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 [65]. Thus, we derive the general regularized solution

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

where in addition to previous notation, the solution $\widetilde{\boldsymbol{\beta}}$ is used to denote the regularized effect sizes; 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 [22], $\alpha = 1$ equates to Ridge Regression [24], while $0 < \alpha < 1$ results in the combined Elastic Net [23]. 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 [66]. 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 for each of the gene-ε regularization implementations are given in the main text and Supplementary Note.

Estimating ε-Genic Effects

In contrast, to the traditional association tests described earlier, the gene-ε model assumes a (SNP-level) null hypothesis $H_0 \colon \beta_j \approx 0$ with

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

where σ_j^2 denotes the proportion of narrow-sense heritability h^2 contributed by the j-th SNP (with the constraint $\sum \sigma_j^2 = h^2$). To infer σ_j^2 for each SNP, gene-ε uses an empirical Bayes approach by fitting a K-mixture of normal distributions over the regularized effect sizes,

$$\widetilde{\beta}_j \sim \pi \sum_{k=1}^{K-1} \gamma_k \mathcal{N}(0, \sigma_k^2) + (1 - \pi)\delta_0, \tag{7}$$

where δ_0 is the Dirac delta function with fixed variance $\sigma_K^2 = 0$, and indicates the $(1 - \pi)$ fraction of SNPs that do not directly influence the trait of interest (i.e., non-causal). Equivalently, we say

$$\widetilde{\beta}_j \sim \sum_{k=1}^{K} \gamma_k \mathcal{N}(0, \sigma_k^2) \tag{8}$$

The above mixture allows for distinct clusters of nonzero effects through K different variance components $(\sigma_k^2, k = 1, \ldots, K)$ [18]. Here, we consider sequential fractions $(\gamma_1, \ldots, \gamma_{K-1})$ of SNPs to have distinctively 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 effects will belong to the latter groups. Throughout this study, we use an ε-genic threshold $\sigma^2 = \sigma_2^2$. This means that we consider SNP j to be causal if it is grouped in the largest fraction (i.e., $\sigma_j^2 > \sigma_2^2$), and interactive or non-causal otherwise. Given Equations (7) and (8), we fit the mixtures for all J SNPs by aiming to maximize the joint log-likelihood

$$\log p(\widetilde{\boldsymbol{\beta}} \,|\, \boldsymbol{\Theta}) = \sum_{j=1}^{J} \log p(\widetilde{\beta}_j \,|\, \boldsymbol{\theta}_j), \tag{9}$$

where $\boldsymbol{\theta}_j = (\gamma_1, \ldots, \gamma_K, \sigma_1^2, \ldots, \sigma_K^2)$ is the set of mixture parameters to be estimated for the j-th SNP, and $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J)$. In the gene-$\varepsilon$ framework, estimation is done using an expectation-maximization (EM) algorithm to maximize the joint distribution in Equation (9). The number of components K is chosen via a Bayesian Information Criterion (BIC). Details of this algorithm are given in the Supplementary Note.

The gene-ε Gene-level Association Test Statistic

We now derive the gene-ε gene-level association test statistic for complex traits using GWA summary statistics. Denote gene (or genomic region) g with a known set of SNPs $j \in \mathcal{J}_g$ — where, in practice, $\mathcal{J}_g$ may include SNPs within the boundaries of g and/or within its corresponding regulatory region. We conformably partition the regularized SNP-level effect sizes and define a gene-level test statistic

$$\widetilde{Q}_g = \widetilde{\boldsymbol{\beta}}_g^{\intercal} \mathbf{A} \widetilde{\boldsymbol{\beta}}_g, \tag{10}$$

where $\mathbf{A}$ is a predefined symmetric and positive semi-definite weight matrix. Each $\widetilde{Q}_g$ is used to test for significant enrichment of associated mutations at the gene level. This leads to an ε-genic inspired (gene-level) null hypothesis $H_0 : \widetilde{Q}_g \approx 0$ where, because of the normality assumption in Equation (6), $\widetilde{Q}_g$ is assumed to follow a mixture of chi-square distributions,

$$\widetilde{Q}_g \sim \sum_{j=1}^{|\mathcal{J}_g|} \lambda_j \chi^2_{1,j}, \tag{11}$$

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

$$\widetilde{\boldsymbol{\Sigma}}_{g,0}^{1/2} \mathbf{A} \widetilde{\boldsymbol{\Sigma}}_{g,0}^{1/2},$$

where $\widetilde{\boldsymbol{\Sigma}}_{g,0} = \sigma_2^2 \boldsymbol{\Sigma}_g$ is the covariance structure of gene g with only SNPs emitting ε-genic effects under the null hypothesis. Several approximate and exact methods have been suggested to obtain P-values under this distribution. In this study, we use Imhof's method [25].

We highlighted some of the additional features of the gene-ε association test statistics. 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. To see this more formally, consider the expansion of Equation (10) derived from the expectation of quadratic forms,

$$\mathbb{E}[\widetilde{Q}_g] = \mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}_g \mathbf{A}) = \sum_r \sum_s (\widetilde{\boldsymbol{\Sigma}}_g \circ \mathbf{A})_{rs} = h_g^2, \tag{12}$$

where $\mathrm{tr}(\bullet)$ denotes the matrix trace function, $(\widetilde{\boldsymbol{\Sigma}}_g \circ \mathbf{A})$ 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 hypothesis test 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 [9, 12, 26, 53, 67–69]. The variances of the gene-ε gene-level test statistics are given by

$$\mathbb{V}[Q_g] = 2\mathrm{tr}(\mathbf{A} \widetilde{\boldsymbol{\Sigma}}_g \mathbf{A} \widetilde{\boldsymbol{\Sigma}}_g) \tag{13}$$

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

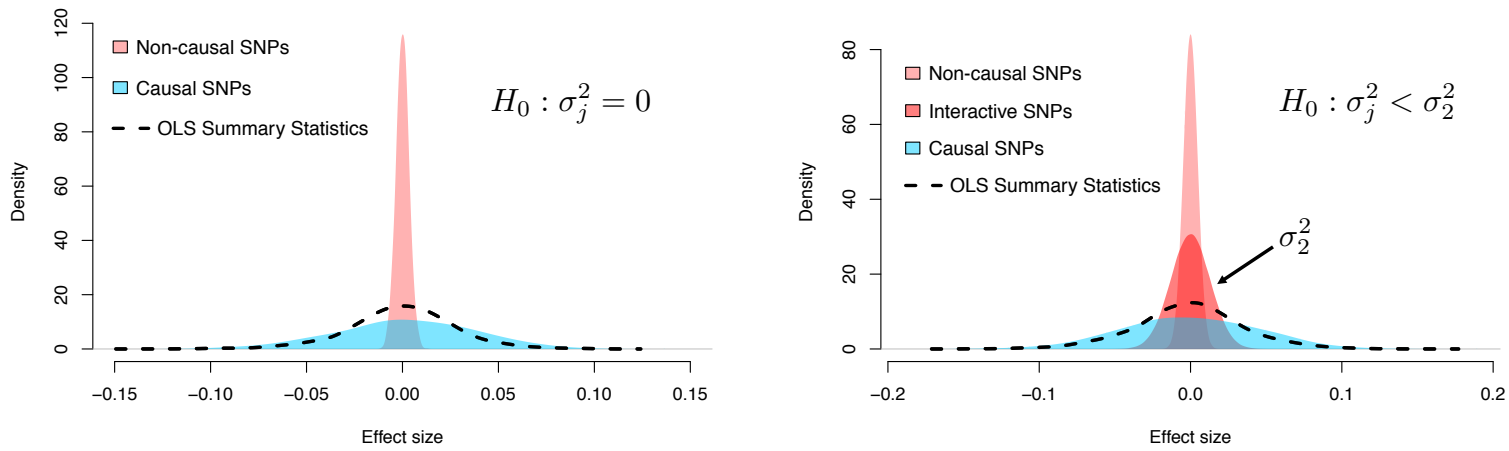
$$\mathbf{y} = \mathbf{W}\boldsymbol{\alpha} + \sum_{c \in \mathcal{C}} \mathbf{x}_c \beta_c + \mathbf{e}, \qquad \mathbf{e} \sim \mathcal{N}(\mathbf{0}, \tau^2 \mathbf{I}), \tag{14}$$

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 five principal components from the genotype matrix) with corresponding fixed effects $\boldsymbol{\alpha}$; 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 $\mathbb{V}[\mathbf{W}\boldsymbol{\alpha}] + \mathbb{V}[\mathbf{e}] = (1 - h^2)$. 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 $\boldsymbol{\Sigma}$ 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, and RSS). Given different model parameters, we simulate data mirroring a wide range of genetic architectures (Supplementary Note).

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 Model

(b) Omnigenic Model

Figure 1. 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. The black dashed line represents the distribution of estimated marginal effect sizes from a single-SNP univariate linear model (ordinary least squares, or "OLS", summary statistics). **(a)** The traditional model of complex traits simply assumes SNPs are causal or non-causal. Under this null model, 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 traits. **(b)** Under the omnigenic model [4], 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 different degrees of connectedness (e.g., through linkage disequilibrium) or interactions with core genes. 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.

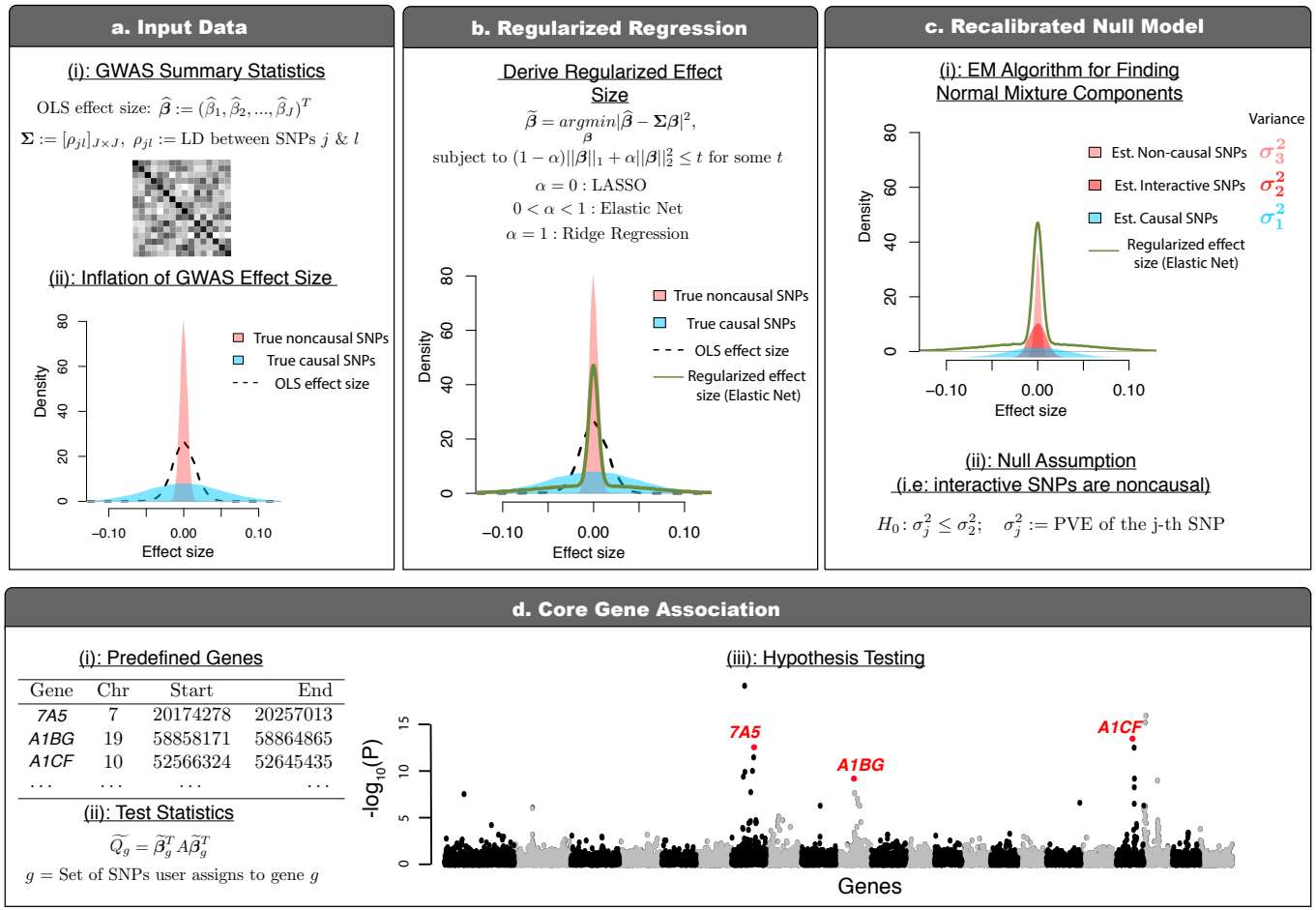
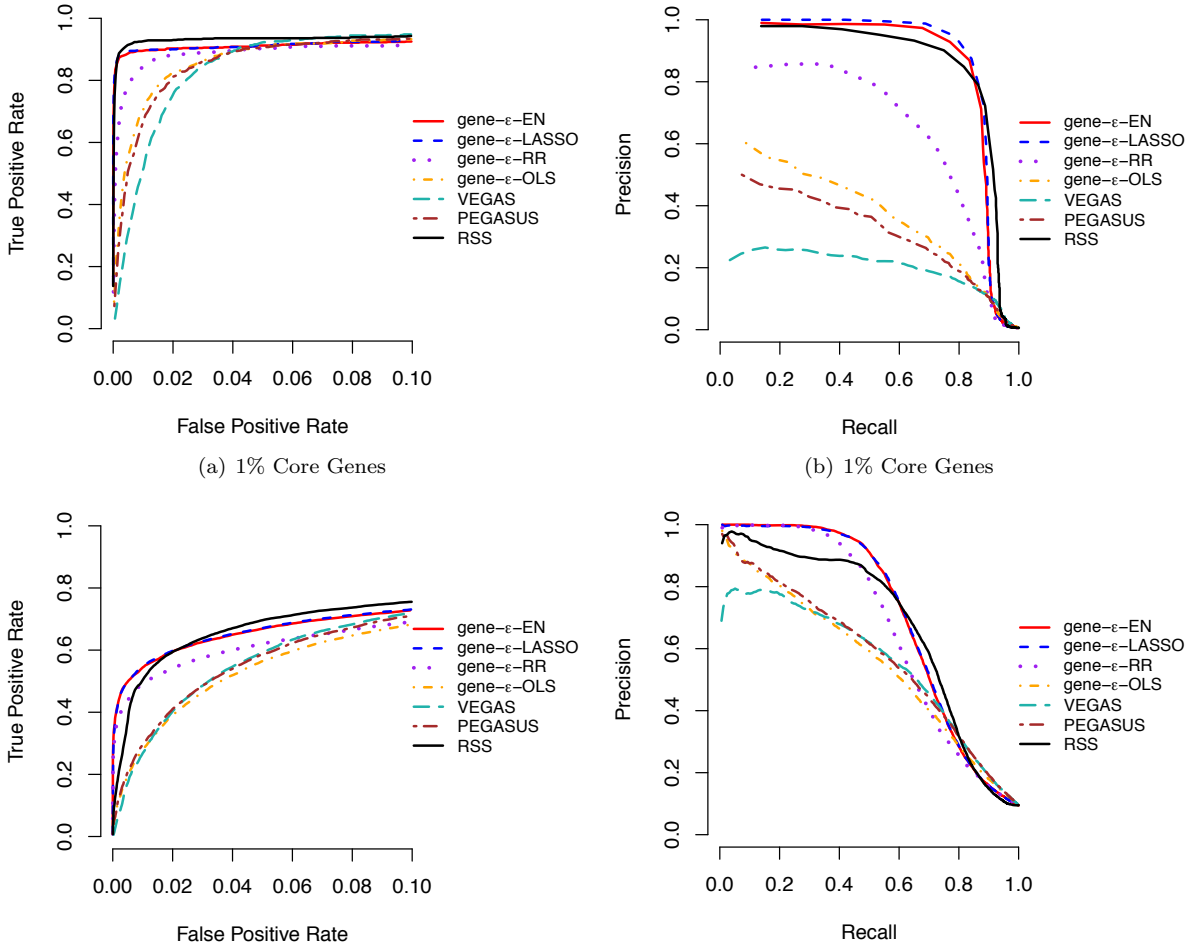


Figure 2. Schematic overview of gene-ε: our new gene-level association approach modeling ε-genic effects. (a) gene-ε takes SNP-level GWA marginal effect sizes (OLS estimates $\widehat{\boldsymbol{\beta}}$) and a linkage disequilibrium (LD) matrix ($\boldsymbol{\Sigma}$) as input. It is well-known that 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 Note 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 (SNP-level) null hypothesis $H_0 : \sigma_j^2 \leq \sigma^2$, where σ^2 is the ε-genic threshold and represents the maximum proportion-of-variance-explained (PVE) that is explained 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 $\sigma^2 = \sigma_2^2$. **(d)** Lastly, gene-ε computes gene-level association test statistics using quadratic forms, and estimates corresponding P-values using Imhof's method. For more details, see Methods.



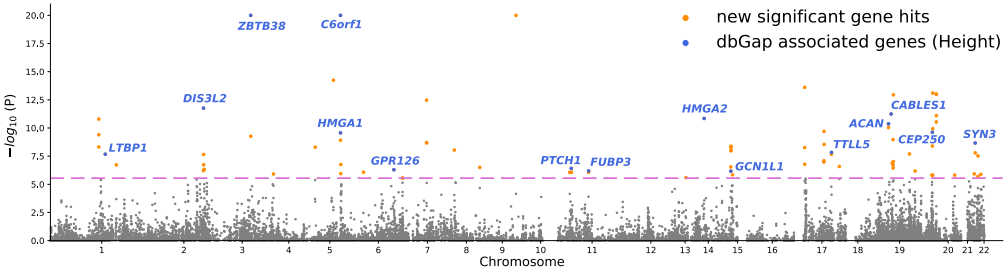
(a) 1% Core Genes

(b) 1% Core Genes

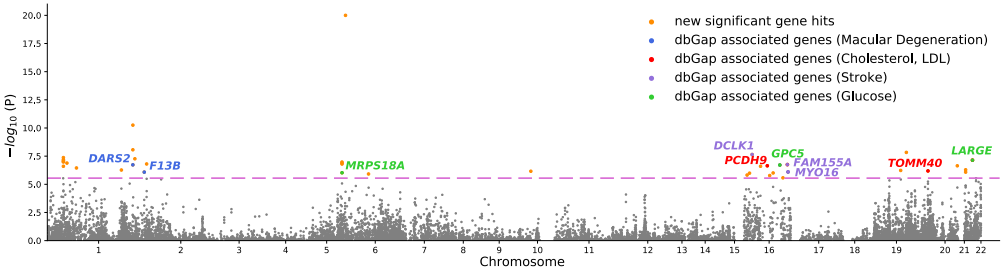
(c) 10% Core Genes

(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 Note). 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 three existing methods: PEGASUS (brown) [12], VEGAS (teal) [7], and the Bayesian approach RSS (black) [14]. **(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.



(a)



(b)

	p value	*q* value	*z* value	Combined score	# of genes in dbGap
Body Height	6.239e-11	2.059e-9	-2.36	55.44	15
Respiratory Function Tests	0.2577	0.3194	-2.23	8.16	3
Prion Diseases	0.2904	0.3194	-1.60	5.65	1
Alcoholism	0.05833	0.3770	-1.97	5.60	2
Smoking	0.04927	0.3770	-1.68	5.07	1
Stroke	0.1219	0.3770	-2.00	4.20	3
Leprosy	0.09993	0.3770	-1.70	3.92	1
Waist Circumference	0.1255	0.3770	-1.78	3.69	2
Hippocampus	0.1150	0.3770	-1.63	3.53	1
Iron	0.1368	0.3770	-1.68	3.34	2

(c)

	p value	*q* value	*z* value	Combined score	# of genes in dbGap
Macular Degeneration	0.001687	0.04567	-2.23	14.23	3
Cholesterol, LDL	0.003109	0.04567	-2.43	14.01	4
Stroke	0.002592	0.04567	-2.33	13.89	4
Glucose	0.003150	0.04567	-2.13	12.28	3
Cholesterol, HDL	0.005497	0.05258	-2.15	11.19	4
Coronary Artery Disease	0.007934	0.05258	-1.93	9.34	3
Alzheimer Disease	0.01117	0.05258	-1.54	6.94	2
Cholesterol	0.01632	0.05258	-1.56	6.41	3
C-Reactive Protein	0.01340	0.05258	-1.44	6.20	2
Lipoproteins, VLDL	0.01282	0.05258	-1.39	6.04	2

(d)

Figure 4. Gene-level association results from applying gene-ε to Height (panels a and c) and Body Mass Index (BMI; panels b and d), assayed in European-ancestry individuals in the UK Biobank. Manhattan plots of gene-$\boldsymbol{\varepsilon}$ gene-level association P-values using Elastic Net regularized effect sizes for **(a)** Body Height and **(b)** BMI. The purple dashed line indicates a log-transformed Bonferroni-corrected significance threshold ($P = 2.83 \times 10^{-6}$ correcting for 17652 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 [41, 70] 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 15 of the genes identified by gene-ε appearing in this category. For BMI, the four significant dbGAP categories are Macular Degeneration, Cholesterol, Stroke, and Glucose — all of which have been connected to BMI [47, 51, 71–74].

Trait	Gene	Chr	gene-ε P-Value	Rank	h_g^2	Biological Relevance to Trait	Ref(s)
Height	*CHD8*	14	2.45×10^{-14}	5	1.09×10^{-2}	Disruption in zebrafish recapitulates features of the human phenotype, including increased head size.	[75]
Height	*ZNF628*	19	9.65×10^{-14}	7	2.58×10^{-2}	Rare variants within the gene have been associated with adult height.	[35]
Height	*MATN3*	2	1.63×10^{-11}	15	9.70×10^{-3}	Mutations in *MATN3* can cause Multiple epiphyseal dysplasia (MED) and affect bone length.	[76]
BMI	*C1orf91*	1	4.24×10^{-8}	6	3.66×10^{-3}	Associated with nonalcoholic fatty liver disease (NAFLD) which has risk of affecting BMI.	[77]
BMI	*HTF9C*	22	5.13×10^{-7}	26	2.34×10^{-3}	Orthologous gene in mice affects lean body mass.	[42, 78–80]
BMI	*USPL1*	13	1.02×10^{-6}	36	3.03×10^{-3}	*USPL1* is a member of a short region in which deletions can cause obesity.	[81]
MCV	*LRCH4*	7	1×10^{-20}	1*	1.71×10^{-2}	Involved in ligand binding which can affect deformability of red blood cell surface necessary for merozoite invasion.	[82]
MCV	*FAM21B-2*	10	1×10^{-20}	1*	3.28×10^{-3}	Involved in the endocytosis pathway, which can affect risks of megaloblastic anaemia (i.e., unusually large red blood cells).	[83–86]
MCV	*LOC284194*	17	1×10^{-20}	1*	1.65×10^{-3}	—	—
MPV	*ZNF385A*	12	1×10^{-20}	1*	1.56×10^{-2}	Mutations in mouse orthologue can cause gastrointestinal and intracranial hemorrhages.	[78–80]
MPV	*EXDL2*	14	1×10^{-20}	1*	4.55×10^{-2}	Identified as an important enhancer of the gene *DCAF5*, which affects MPV.	[87]
MPV	*GIT1*	17	1×10^{-20}	1*	3.18×10^{-2}	Identified as an important enhancer of the gene *ENSG00000266111*, which affects MPV.	[87]
PLC	*A26B3*	21	1×10^{-20}	1*	2.22×10^{-4}	Indicated as being important for attachment of integral membrane proteins to the spectrin-actin based membrane skeleton, which affects platelet formation.	[88]
PLC	*CD302*	2	6.55×10^{-15}	13	1.91×10^{-2}	Identified as an important enhancer of the gene *MARCH7*, which affects platelet counts.	[87]
PLC	*FAM131A*	3	5.33×10^{-14}	17	2.33×10^{-2}	Identified as an important enhancer of the gene *POLR2H*, which affects platelet counts.	[87]
WHR	*LRFN4*	11	1.09×10^{-7}	6*	2.86×10^{-3}	Promotes neurite outgrowth and affects cognitive function, possibly impairing abdominal obesity.	[89]
WHR	*C1orf91*	1	2.12×10^{-7}	9	4.4×10^{-3}	Involved in formation of transmembrane proteins, such as insulin receptor, which affect central (i.e., intra-abdominal) obesity.	[83, 85, 86, 90]
WHR	*PTRHD1*	2	2.64×10^{-7}	11	6.08×10^{-3}	Previously associated with BMI, diabetes, and height.	[87]

Table 1. Top three novel candidate core genes reported by gene-ε for the six quantitative traits studied in the UK Biobank. 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; here, they are 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 17652 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.

19. Liu X, Li YI, Pritchard JK. Trans effects on gene expression can drive omnigenic inheritance. bioRxiv. 2018;p. 425108. Available from: http://biorxiv.org/content/early/2018/09/24/425108.abstract.

20. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. Genome-wide genetic data on ∼500,000 UK Biobank participants. bioRxiv. 2017;p. 166298. Available from: http://biorxiv.org/content/early/2017/07/20/166298.abstract.

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

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

23. 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.

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

25. 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.

26. 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.

27. 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.

28. 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.

29. 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.

30. 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.

31. 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.

32. 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.

33. 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.

34. 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.

35. 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.

36. 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.

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

38. 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.

39. 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.

40. 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.

41. 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.

42. 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.

43. Seddon JM, Cote J, Davis N, Rosner B. Progression of age-related macular degeneration: association with body mass index, waist circumference, and waist-hip ratio. JAMA Ophthalmol. 2003;121(6):785–792. Available from: https://doi.org/10.1001/archopht.121.6.785.

44. Peeters A, Magliano DJ, Stevens J, Duncan BB, Klein R, Wong TY. Changes in abdominal obesity and age-related macular degeneration: the atherosclerosis risk in communities study. JAMA Ophthalmol. 2008;126(11):1554–1560. Available from: https://doi.org/10.1001/archopht.126.11.1554.

45. Adams MKM, Simpson JA, Aung KZ, Makeyeva GA, Giles GG, English DR, et al. Abdominal obesity and age-related macular degeneration. Am J Epidemiol. 2011;173(11):1246–1255. Available from: https://doi.org/10.1093/aje/kwr005.

46. Neale BM, Fagerness J, Reynolds R, Sobrin L, Parker M, Raychaudhuri S, et al. Genome-wide association study of advanced age-related macular degeneration identifies a role of the hepatic lipase gene (*LIPC*). Proc Natl Acad Sci U S A. 2010;107(16):7395–7400. Available from: http://www.pnas.org/content/107/16/7395.abstract.

47. Faeh D, Braun J, Bopp M. Body mass index vs cholesterol in cardiovascular disease risk prediction models. JAMA Intern Med. 2012;172(22):1766–1768.

48. 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.

49. Zhu Z, Zheng Z, Zhang F, Wu Y, Trzaskowski M, Maier R, et al. Causal associations between risk factors and common diseases inferred from GWAS summary data. Nat Comm. 2018;9(1):224. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29335400.

50. Fesinmeyer MD, Meigs JB, North KE, Schumacher FR, Bůžková P, Franceschini N, et al. Genetic variants associated with fasting glucose and insulin concentrations in an ethnically diverse population: results from the Population Architecture using Genomics and Epidemiology (PAGE) study. BMC Med Genet. 2013;14:98–98. Available from: https://www.ncbi.nlm.nih.gov/pubmed/24063630.

51. Kurth T, Gaziano JM, Berger K, Kase CS, Rexrode KM, Cook NR, et al. Body mass index and the risk of stroke in men. JAMA Intern Med. 2002;162(22):2557–2562.

52. Larsson SC, Scott RA, Traylor M, Langenberg CC, Hindy G, Melander O, et al. Type 2 diabetes, glucose, insulin, BMI, and ischemic stroke subtypes: Mendelian randomization study. Neurology. 2017;89(5):454–460. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28667182.

53. 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.

54. 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.

55. 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.

56. 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.

57. 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.

58. 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.

59. GTEx Consortium. Genetic effects on gene expression across human tissues. Nature. 2017 10;550:204–213. Available from: https://doi.org/10.1038/nature24277.

60. 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.

61. 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.

62. 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/.

63. 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.

64. 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.

65. 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.

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

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

68. 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.

69. 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.

70. 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.

71. Song YM, Sung J, Smith GD, Ebrahim S. Body mass index and ischemic and hemorrhagic stroke: a prospective study in Korean men. Stroke. 2004;35(4):831–836.

72. Faheem M, Qureshi S, Ali J, Hameed H, Zahoor Z, Abbas F, et al. Does BMI affect cholesterol, sugar, and blood pressure in general population? J Ayub Med Coll Abbottabad. 2010;22(4):74–77.

73. Zhang QY, Tie LJ, Wu SS, Lv PL, Huang HW, Wang WQ, et al. Overweight, obesity, and risk of age-related macular degeneration. Investig Ophthalmol Vis Sci. 2016;57(3):1276–1283.

74. Agrawal N, Agrawal MK, Kumari T, Kumar S. Correlation between body mass index and blood glucose levels in Jharkhand population. IJCMR. 2017;4(8):1633–6.

75. Bernier R, Golzio C, Xiong B, Stessman HA, Coe BP, Penn O, et al. Disruptive CHD8 mutations define a subtype of autism early in development. Cell. 2014;158(2):263–276.

76. Seo SG, Song HR, Kim HW, Yoo WJ, Shim JS, Chung CY, et al. Comparison of orthopaedic manifestations of multiple epiphyseal dysplasia caused by MATN3 versus COMP mutations: a case control study. BMC Musculoskelet Disord. 2014;15(1):84.

77. Wu J, Zhang R, Shen F, Yang R, Zhou D, Cao H, et al. Altered DNA methylation sites in peripheral blood leukocytes from patients with simple steatosis and nonalcoholic steatohepatitis (NASH). Med Sci Monit. 2018;24:6946.

78. 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.

79. 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.

80. 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.

81. D'Angelo CS, Varela MC, de Castro CIE, Otto PA, Perez ABA, Lourenço CM, et al. Chromosomal microarray analysis in the genetic evaluation of 279 patients with syndromic obesity. Mol Cytogenet. 2018;11(1):14.

82. Sisquella X, Nebl T, Thompson JK, Whitehead L, Malpede BM, Salinas ND, et al. Plasmodium falciparum ligand binding to erythrocytes induce alterations in deformability essential for invasion. Elife. 2017;6:e21083.

83. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30.

84. Castellanos-Sinco H, Ramos-Peñafiel C, Santoyo-Sánchez A, Collazo-Jaloma J, Martínez-Murillo C, Montaño-Figueroa E, et al. Megaloblastic anaemia: Folic acid and vitamin B12 metabolism. Revista Médica Del Hospital General De México. 2015;78(3):135–143.

85. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2016;45(D1):D353–D361.

86. Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M. New approach for understanding genome variations in KEGG. Nucleic Acids Res. 2018;47(D1):D590–D595.

87. Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database. 2017;2017.

88. Thon JN, Italiano JE. Does size matter in platelet production? Blood. 2012;120(8):1552–1561.

89. Lackner N, Bengesser S, Birner A, Painold A, Fellendorf F, Platzer M, et al. Abdominal obesity is associated with impaired cognitive function in euthymic bipolar individuals. World J Biol Psychiatry. 2016;17(7):535–546.

90. Papaetis GS, Papakyriakou P, Panagiotou TN. Central obesity, type 2 diabetes and insulin: exploring a pathway full of thorns. Arch Med Sci. 2015;11(3):463.