Detection of interactions between genetic marker sets and environment in a genome-wide study of hypertension

Linchuan Shen[1], Amei Amei[1], Bowen Liu[1], Yunqing Liu[2], Gang Xu[1,2], Edwin C. Oh[3,4], Zuoheng Wang[2]

[1]Department of Mathematical Sciences, University of Nevada, Las Vegas

[2]Department of Biostatistics, Yale School of Public Health

[3]Department of Internal Medicine, University of Nevada School of Medicine, Las Vegas

[4]Nevada Institute of Personalized Medicine, University of Nevada, Las Vegas

Correspondence

Amei Amei, PhD
Department of Mathematical Sciences
University of Nevada, Las Vegas
4505 S. Maryland Parkway
Las Vegas, NV 89154
Phone: (702) 895-5159
Fax: (702) 895-4343
Email: amei.amei@unlv.edu

Zuoheng Wang, PhD
Department of Biostatistics
Yale School of Public Health
60 College St.
New Haven, CT 06510
Phone: (203) 737-2672
Fax: (203) 785-6912
Email: zuoheng.wang@yale.edu

Summary

As human complex diseases are influenced by the interplay of genes and environment, detecting gene-environment interactions ($G \times E$) can shed light on biological mechanisms of diseases and play an important role in disease risk prediction. Development of powerful quantitative tools to incorporate $G \times E$ in complex diseases has potential to facilitate the accurate curation and analysis of large genetic epidemiological studies. However, most of existing methods that interrogate $G \times E$ focus on the interaction effects of an environmental factor and genetic variants, exclusively for common or rare variants. In this study, we proposed two tests, MAGEIT_RAN and MAGEIT_FIX, to detect interaction effects of an environmental factor and a set of genetic markers containing both rare and common variants, based on the MinQue for Summary statistics. The genetic main effects in MAGEIT_RAN and MAGEIT_FIX are modeled as random or fixed, respectively. Through simulation studies, we illustrated that both tests had type I error under control and MAGEIT_RAN was overall the most powerful test. We applied MAGEIT to a genome-wide analysis of gene-alcohol interactions on hypertension in the Multi-Ethnic Study of Atherosclerosis. We detected two genes, *CCNDBP1* and *EPB42*, that interact with alcohol usage to influence blood pressure. Pathway analysis identified sixteen significant pathways related to signal transduction and development that were associated with hypertension, and several of them were reported to have an interactive effect with alcohol intake. Our results demonstrated that MAGEIT can detect biologically relevant genes that interact with environmental factors to influence complex traits.

KEYWORDS

Gene-environment interaction; genome-wide study; method of moments; mixed effects model

1. Introduction

Causes of human complex diseases are multifactorial including the interplay of genes and environment. The effect of environment exposures on disease outcomes can vary across genotypic groups. It has been reported that individuals with certain genetic profiles have elevated disease risk only when they are exposed to an environment in many complex diseases (Lin *and others.*, 2013). For example, many environmental factors such as aging, sex, smoking, diet, stress, air quality and among others influence disease risk, progression and severity (Bhatnagar, 2017, Cosselman, Navas-Acien and Kaufman, 2015). As a result, incorporating gene-environment interactions ($G \times E$) has become crucial in the study of complex traits. Genome-wide association studies (GWAS) have successfully identified many genetic variants associated with human diseases. However, the estimated effects of these variants are small and only explain small portion of the heritability of complex diseases (Eichler *and others.*, 2010). Several studies have suggested that $G \times E$ may contribute partly to the missing heritability and the detection of $G \times E$ could lead to meaningful implication in fields of public health and personalized medicine (Eichler *and others.*, 2010, Thomas, 2010).

Traditional $G \times E$ analyses focus on evaluating the interactions with genetic variants one at a time (Aschard *and others.*, 2010, Kraft *and others.*, 2007, Manning *and others.*, 2011). Possible limitations in such approaches include the burden of multiple hypothesis testing and lacking consideration of joint effects shared by multiple variants with similar biological functions, resulting in power loss in the analysis (Lin *and others.*, 2013). In recent years, genome-wide search for $G \times E$ has been emerging (Khoury and Wacholder, 2009, Thomas, 2010) and several studies have investigated $G \times E$ from multiple variants in a genetic marker set (Chen, Meigs and Dupuis, 2014, Chi *and others.*, 2021, Jiao *and others.*, 2013, Lin *and others.*, 2019, Lin *and others.*, 2013, Lin *and others.*, 2016, Su, Di and Hsu, 2017, Tzeng *and others.*, 2011, Wang *and others.*, 2020). For a set of common genetic variants, gene-environment set association test (GESAT) was developed using a generalized linear model and ridge regression (Lin *and others.*, 2013). For rare variants, Chen *et al*. proposed INT-FIX and INT-RAN for testing $G \times E$ effect, as well as a joint test, JOINT, that detects the effects of a set of genetic variants as well as their interactions with an environmental factor simultaneously (Chen, Meigs and Dupuis, 2014). They used a beta density function for genetic effect to reflect larger contributions from rare genetic variants. Genetic main effects in their $G \times E$ tests were treated as fixed in INT-FIX or random in INT-RAN, respectively. The three tests were implemented as an R package called rareGE. To assess rare variants by environment interaction, Lin et al. developed the interaction sequence kernel association test (iSKAT) that modeled the main effects of rare variants using weighted ridge regression and allowed the interactions with environment across genetic variants to be correlated (Lin *and others.*, 2016). GESAT, the three tests in the rareGE package and iSKAT are all variance component-based tests that are robust to the signs and magnitudes of the $G \times E$ effects when many variants in a genetic region are non-causal and/or there are mixed beneficial and detrimental variants (Lee, Wu and Lin, 2012, Santorico and Hendricks, 2016, Wu *and others.*, 2011). A unified hierarchical modeling of $G \times E$ effects from a set of rare variants, called mixed effects score test for interaction (MiSTi), which models $G \times E$ effects by a fixed component as well as a random component was developed (Su, Di and Hsu, 2017). They constructed two independent score statistics and combined them using data-adaptive approaches. Simulation studies showed that MiSTi has greater than or comparable power to iSKAT. MiSTi provided a unified regression framework for testing interaction effects between a set of rare variants and an environmental factor where many existing methods can be derived from by constraining certain parameters to be zero. In addition to the above mentioned $G \times E$ tests that were developed under the regression framework, Lin *et al.* proposed a polygenic test of $G \times E$ effect using Bayes factors (Lin *and others.*, 2019). In their adaptive combination of Bayes factors method (ADABF), $G \times E$ effects are assumed to follow a normal distribution. Variants in a genetic region were sorted by Bayes factors and p-values were calculated using a resampling procedure. When there are a few genetic variants interacting with the environmental factor, ADABF had higher power than other methods for detecting $G \times E$ effects.

Complex diseases are influenced by many genetic variants including common and/or rare. Current methods in detecting $G \times E$ mainly focus on the interaction effects of an environmental factor and genetic variants, exclusively for common or rare. Although ADABF considers both common and rare variants in a genetic region, it does not distinguish the effects of the two types of variants in model fitting and hence may overlook the relatively larger contribution from rare variants. Recently, MQS (MinQue for Summary statistics) was developed for estimating variance components in linear mixed models (Zhou, 2017). MQS is based on the method of moments and the minimal norm quadratic unbiased estimation criterion. Compared to the restricted maximum likelihood estimation method (REML), MQS provided unbiased and statistically efficient estimates. It was extended to model the epistatic interactions between genetic variants (Crawford *and others.*, 2017). In this study, we propose two tests to detect interactions between an environmental factor and a set of genetic markers containing both rare and common variants based on the MQS method. We name it as MArginal Gene-Environment Interaction Test with RANdom or FIXed genetic effects (MAGEIT_RAN or MAGEIT_FIX). We assessed the performance of the two tests in detecting $G \times E$ for a set of genetic variants and compared it with existing set-based $G \times E$ methods via simulation studies. Our results demonstrated that both MAGEIT_RAN and MAGEIT_FIX had well controlled type I error. MAGEIT_RAN was most powerful in majority of the simulation scenarios. We applied MAGEIT_RAN and MAGEIT_FIX to a genome-wide analysis of gene-alcohol interaction on hypertension in the Multi-Ethnic Study of Atherosclerosis (MESA) and identified hypertension-related $G \times E$ and pathways.

2. Methods

Suppose a phenotype of interest, an environmental variable and genome-wide genetic variants are measured on n subjects. Let $y_k, E_k, \boldsymbol{G}_k = \left(G_{k1}, G_{k2}, \dots, G_{kp}\right)^T$ and $\boldsymbol{X}_k = (X_{k1}, X_{k2}, \dots, X_{km})^T$ denote the phenotype, environmental variable, genotypes of p variants in a genomic region, and m non-genetic covariates for the kth subject, respectively, for $k = 1, 2, \dots, n$, where $G_{kj} = 0$, 1 or 2 depending on whether subject k has 0, 1 or 2 copies of minor allele at the jth variant. We use $\boldsymbol{S}_k = \left(E_k G_{k1}, E_k G_{k2}, \dots, E_k G_{kp}\right)^T$ to denote the genetic variants by environment interaction for the kth subject. Our goal is to test whether there are interactions between the variant set and environment that influence the phenotype of interest.

2.1 Model for continuous phenotype

Let $\boldsymbol{y} = (y_1, y_2, \dots, y_n)^T$, $\boldsymbol{E} = (E_1, E_2, \dots, E_n)^T$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)^T$ denote vectors of the phenotype, environmental variable, and error term of length n. We further define an $n \times m$ covariate matrix $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n]^T$, an $n \times p$ genotype matrix $\boldsymbol{G} = [\boldsymbol{G}_1, \boldsymbol{G}_2, \cdots, \boldsymbol{G}_n]^T$, and an $n \times p$ matrix $\boldsymbol{S} = [\boldsymbol{S}_1, \boldsymbol{S}_2, \cdots, \boldsymbol{S}_n]^T$ of the $G \times E$. Then, the following model specifies the relationship between a continuous phenotype $\boldsymbol{Y}$ and $\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{G}$ and $\boldsymbol{S}$

$$\boldsymbol{y} = \alpha_0 \mathbf{1} + \boldsymbol{X}\boldsymbol{\alpha}_1 + \alpha_2 \boldsymbol{E} + \boldsymbol{G}\boldsymbol{\beta} + \boldsymbol{S}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{1}$ is an $n \times 1$ vector of 1, α_0 is an intercept term, $\boldsymbol{\alpha}_1 = (\alpha_{11}, \alpha_{12}, \dots, \alpha_{1m})^T$, α_2, $\boldsymbol{\beta} = \left(\beta_1, \beta_2, \dots, \beta_p\right)^T$ and $\boldsymbol{\gamma} = \left(\gamma_1, \gamma_2, \dots, \gamma_p\right)^T$ are regression coefficients for the covariates, environmental factor, genetic variants, and $G \times E$ terms. We further assume that $\boldsymbol{\gamma}$ and $\boldsymbol{\varepsilon}$ follow multivariate normal distributions with $\boldsymbol{\gamma} \sim \mathrm{MVN}(\mathbf{0}, \frac{\sigma^2}{p}\boldsymbol{W}_2^2)$ and $\boldsymbol{\varepsilon} \sim \mathrm{MVN}(\mathbf{0}, \tau^2\boldsymbol{I}_n)$, where $\boldsymbol{W}_2 = \mathrm{diag}(w_{21}, w_{22}, \cdots, w_{2p})$ contains weights of the p $G \times E$ terms and $\boldsymbol{I}_n$ is an identity matrix of dimension n.

2.2 Marginal gene-environment interaction test

We are interested in testing genetic variants by environment interactions in a genomic region, i.e., testing the null hypothesis $H_0: \boldsymbol{\gamma} = \mathbf{0}$, which is equivalent to testing $H_0: \sigma^2 = 0$. We develop two $G \times E$ tests, in which the genetic main effects $\boldsymbol{\beta}$ are modeled as random and fixed, respectively.

When we treat the genetic main effects $\boldsymbol{\beta}$ as random, we assume that $\boldsymbol{\beta} \sim \mathrm{MVN}(\mathbf{0}, \frac{\omega^2}{p}\boldsymbol{W}_1^2)$, where $\boldsymbol{W}_1 = \mathrm{diag}(w_{11}, w_{12}, \cdots, w_{1p})$ are weights of the p variants. We use the MQS method (Zhou, 2017) to estimate the three variance components ω^2, σ^2 and τ^2. In order to eliminate the fix effects $\alpha_0, \boldsymbol{\alpha}_1$ and α_2 in Model (1), we multiply both sides of the model, from left, by a projection matrix $\boldsymbol{M}$, where $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{b}(\boldsymbol{b}^T\boldsymbol{b})^{-1}\boldsymbol{b}^T$ with $\boldsymbol{b} = [\mathbf{1}, \boldsymbol{X}, \boldsymbol{E}]$. Then Model (1) becomes

$$\boldsymbol{y}^* = \boldsymbol{g}^* + \boldsymbol{s}^* + \boldsymbol{\varepsilon}^*,$$

where $\boldsymbol{y}^* = \boldsymbol{My}$, $\boldsymbol{g}^* = \boldsymbol{MG\beta}$, $\boldsymbol{s}^* = \boldsymbol{MS\gamma}$, and $\boldsymbol{\varepsilon}^* = \boldsymbol{M\varepsilon}$. It follows that $\boldsymbol{g}^* \sim \mathrm{MVN}(\mathbf{0}, \omega^2\boldsymbol{G}^*)$ with $\boldsymbol{G}^* = \frac{(\boldsymbol{MGW}_1)(\boldsymbol{MGW}_1)^T}{p}$, $\boldsymbol{s}^* \sim \mathrm{MVN}(\mathbf{0}, \sigma^2\boldsymbol{S}^*)$ with $\boldsymbol{S}^* = \frac{(\boldsymbol{MSW}_2)(\boldsymbol{MSW}_2)^T}{p}$, and $\boldsymbol{\varepsilon}^* \sim \mathrm{MVN}(\mathbf{0}, \tau^2\boldsymbol{M})$. Consequently, we have $\boldsymbol{y}^* \sim \mathrm{MVN}(\mathbf{0}, \omega^2\boldsymbol{G}^* + \sigma^2\boldsymbol{S}^* + \tau^2\boldsymbol{M})$.

We estimate the variance components using the method of moments based on the following set of second moment matching equations,

$$E(\boldsymbol{y}^{*T}\boldsymbol{A}\boldsymbol{y}^*) = \mathrm{tr}\big(\boldsymbol{A}(\omega^2\boldsymbol{G}^* + \sigma^2\boldsymbol{S}^* + \tau^2\boldsymbol{M})\big) = \omega^2\mathrm{tr}(\boldsymbol{AG}^*) + \sigma^2\mathrm{tr}(\boldsymbol{AS}^*) + \tau^2\mathrm{tr}(\boldsymbol{AM}), \quad (2)$$

where $\boldsymbol{A}$ is an arbitrary symmetric non-negative definite matrix (Zhou, 2017). Since there are three unknown parameters $(\omega^2, \sigma^2, \tau^2)$, three different $\boldsymbol{A}$'s are required to obtain parameter estimates.

In the method of moments, the expectation of Eq. (2) is usually replaced with the realized value $\boldsymbol{y}^{*T}\boldsymbol{A}\boldsymbol{y}^{*}$. Let $\boldsymbol{A}_1 = \boldsymbol{G}^*$, $\boldsymbol{A}_2 = \boldsymbol{S}^*$ and $\boldsymbol{A}_3 = \boldsymbol{M}$ (Zhou, 2017), then, the resulting estimates of the variance components are given in a matrix form as

$$\begin{bmatrix} \widehat{\omega}^2 \\ \hat{\sigma}^2 \\ \hat{\tau}^2 \end{bmatrix} = \boldsymbol{\Lambda}^{-1} \begin{bmatrix} \boldsymbol{y}^{*T}\boldsymbol{G}^*\boldsymbol{y}^* \\ \boldsymbol{y}^{*T}\boldsymbol{S}^*\boldsymbol{y}^* \\ \boldsymbol{y}^{*T}\boldsymbol{y}^* \end{bmatrix} = \begin{bmatrix} \mathrm{tr}(\boldsymbol{G}^*\boldsymbol{G}^*) & \mathrm{tr}(\boldsymbol{G}^*\boldsymbol{S}^*) & tr(\boldsymbol{G}^*) \\ \mathrm{tr}(\boldsymbol{S}^*\boldsymbol{G}^*) & \mathrm{tr}(\boldsymbol{S}^*\boldsymbol{S}^*) & tr(\boldsymbol{S}^*) \\ \mathrm{tr}(\boldsymbol{G}^*) & \mathrm{tr}(\boldsymbol{S}^*) & n-(m+2) \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{y}^{*T}\boldsymbol{G}^*\boldsymbol{y}^* \\ \boldsymbol{y}^{*T}\boldsymbol{S}^*\boldsymbol{y}^* \\ \boldsymbol{y}^{*T}\boldsymbol{y}^* \end{bmatrix},$$

where we used $\mathrm{tr}(\boldsymbol{G}^*\boldsymbol{M}) = \mathrm{tr}(\boldsymbol{M}\boldsymbol{G}^*) = \mathrm{tr}(\boldsymbol{G}^*)$, $\mathrm{tr}(\boldsymbol{S}^*\boldsymbol{M}) = \mathrm{tr}(\boldsymbol{M}\boldsymbol{S}^*) = \mathrm{tr}(\boldsymbol{S}^*)$, $\mathrm{tr}(\boldsymbol{M}\boldsymbol{M}) = \mathrm{tr}(\boldsymbol{M}) = n-(m+2)$, and $\boldsymbol{y}^{*T}\boldsymbol{M}\boldsymbol{y}^* = \boldsymbol{y}^{*T}\boldsymbol{y}^*$. The variance component estimator $\hat{\sigma}^2$ is considered as the test statistic, which we named as MArginal Gene-Environment Interaction Test with RANdom genetic main effects (MAGEIT_RAN). Specifically, the MAGEIT_RAN test statistic is

$$\hat{\sigma}^2 = \boldsymbol{y}^{*T}\{(\boldsymbol{\Lambda}^{-1})_{21}\boldsymbol{G}^* + (\boldsymbol{\Lambda}^{-1})_{22}\boldsymbol{S}^* + (\boldsymbol{\Lambda}^{-1})_{23}\boldsymbol{I}\}\boldsymbol{y}^* = \boldsymbol{y}^{*T}\boldsymbol{H}\boldsymbol{y}^*, \tag{3}$$

where $\boldsymbol{H} = (\boldsymbol{\Lambda}^{-1})_{21}\boldsymbol{G}^* + (\boldsymbol{\Lambda}^{-1})_{22}\boldsymbol{S}^* + (\boldsymbol{\Lambda}^{-1})_{23}\boldsymbol{I}$.

Under H_0: $\sigma^2 = 0$, $\boldsymbol{y}^* \sim \mathrm{MVN}(\boldsymbol{0}, \omega^2\boldsymbol{G}^* + \tau^2\boldsymbol{M})$, suggesting that $\boldsymbol{y}^*$ has the same distribution as $(\omega^2\boldsymbol{G}^* + \tau^2\boldsymbol{M})^{\frac{1}{2}}\boldsymbol{Z}$ with $\boldsymbol{Z} \sim \mathrm{MVN}(\boldsymbol{0}, \boldsymbol{I}_n)$. Therefore, the method of moments estimator $\hat{\sigma}^2$ follows the same distribution as $\boldsymbol{Z}^T((\widehat{\omega}_0^2\boldsymbol{G}^* + \hat{\tau}_0^2\boldsymbol{M})^{\frac{1}{2}})^T\boldsymbol{H}(\widehat{\omega}_0^2\boldsymbol{G}^* + \hat{\tau}_0^2\boldsymbol{M})^{\frac{1}{2}}\boldsymbol{Z}$, which has a mixture of χ^2 distribution $\hat{\sigma}^2 \sim \sum_{i=1}^{n}\lambda_i\chi_{1,i}^2$. Here, $(\widehat{\omega}_0^2, \hat{\tau}_0^2)$ are estimates of (ω^2, τ^2) under the null hypothesis, $(\lambda_1, \cdots, \lambda_n)$ are eigenvalues of the matrix $((\widehat{\omega}_0^2\boldsymbol{G}^* + \hat{\tau}_0^2\boldsymbol{M})^{\frac{1}{2}})^T\boldsymbol{H}(\widehat{\omega}_0^2\boldsymbol{G}^* + \hat{\tau}_0^2\boldsymbol{M})^{\frac{1}{2}}$, and $\chi_{1,i}^2$ are independent χ_1^2 variables (Zhou, 2017).The p-value of $\hat{\sigma}^2$ can be evaluated by the Davies method (Davies, 1980, Wu *and others.*, 2011) and Liu-Tang-Zhang approximation (Liu, Tang and Zhang, 2009).

If we treat the genetic main effects $\boldsymbol{\beta}$ as fixed, we use the MQS method (Zhou, 2017) to estimate the two variance components σ^2 and τ^2. To eliminate the fix effect terms $\alpha_0, \boldsymbol{\alpha}_1$, α_2 and

$\boldsymbol{\beta}$ in Model (1), we left multiply the model by a projection matrix $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{b}(\boldsymbol{b}^T\boldsymbol{b})^{-1}\boldsymbol{b}^T$ with $\boldsymbol{b} = [\boldsymbol{1}, \boldsymbol{X}, \boldsymbol{E}, \boldsymbol{G}]$. Then the model becomes $\boldsymbol{y}^* = \boldsymbol{s}^* + \boldsymbol{\varepsilon}^*$ and it contains two variance components σ^2 and τ^2. Using the method of moments, we obtain the following estimates of the variance components,

$$\begin{bmatrix}\hat{\sigma}^2\\ \hat{\tau}^2\end{bmatrix} = \begin{bmatrix} tr(\boldsymbol{S}^*\boldsymbol{S}^*) & tr(\boldsymbol{S}^*) \\ tr(\boldsymbol{S}^*) & n-(m+p+2)\end{bmatrix}^{-1}\begin{bmatrix}\boldsymbol{y}^{*T}\boldsymbol{S}^*\boldsymbol{y}^* \\ \boldsymbol{y}^{*T}\boldsymbol{y}^*\end{bmatrix}.$$

The variance component estimator $\hat{\sigma}^2$ is considered as the test statistic, which we named as MArginal Gene-Environment Interaction Test with FIXed genetic main effects (MAGEIT_FIX). Specifically, the MAGEIT_FIX test statistic is

$$\hat{\sigma}^2 = \frac{\boldsymbol{y}^{*T}\{(n-(m+p+2))\boldsymbol{S}^* - tr(\boldsymbol{S}^*)\boldsymbol{I}\}\boldsymbol{y}^*}{(n-(m+p+2))tr(\boldsymbol{S}^*\boldsymbol{S}^*) - tr(\boldsymbol{S}^*)^2}. \tag{4}$$

Under $H_0: \sigma^2 = 0$, $\hat{\sigma}^2$ follows a mixture of χ^2 distribution $\hat{\sigma}^2 \sim \sum_{i=1}^{n}\lambda_i\chi_{1,i}^2$ with $(\lambda_1, \cdots, \lambda_n)$ being the eigenvalues of the matrix $((\hat{\tau}_0^2\boldsymbol{M})^{\frac{1}{2}})^T\boldsymbol{H}(\hat{\tau}_0^2\boldsymbol{M})^{\frac{1}{2}}$.

2.3 Model for binary phenotype

We consider a liability threshold model and assume the binary outcome y_k of the kth subject is determined by an unobserved continuous liability variable z_k, i.e.,

$$y_k = \begin{cases} 1, & z_k \geq 0 \\ 0, & z_k < 0 \end{cases} \quad \text{for } k = 1, \dots, n, \tag{5}$$

where the underlying liability vector $\boldsymbol{z} = (z_1, z_2, \cdots, z_n)^T$ is specified using Model (1). The full likelihood of the liability threshold mixed effects model is intractable due to an n-dimensional integration over the liability variable $\boldsymbol{z}$. Following the previous studies (Crawford and Zhou, 2018, Engel, Buist and Visscher, 1995, Kuss, Rasmussen and Herbrich, 2005, Tempelman and Gianola, 1993, Williams and Barber, 1998), the liability threshold mixed effects model can be approximated by a linear mixed effects model on $\hat{\boldsymbol{z}} = E(\boldsymbol{z}|\boldsymbol{y})$, an estimated posterior mean of the liabilities,

$$\hat{\boldsymbol{z}} = \alpha_0 \mathbf{1} + \boldsymbol{X}\boldsymbol{\alpha}_1 + \alpha_2 \boldsymbol{E} + \boldsymbol{G}\boldsymbol{\beta} + \boldsymbol{S}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}. \tag{6}$$

The posterior mean $\hat{\boldsymbol{z}}$ can be obtained by approximation under certain assumptions based on the properties of GWAS data (Crawford and Zhou, 2018). Specifically, we assume that (i) subjects are unrelated, and (ii) both the genetic main effects and interaction effects are small such that the terms $\boldsymbol{G}\boldsymbol{\beta}$ and $\boldsymbol{S}\boldsymbol{\gamma}$ can be ignored. Under these assumptions, the distribution of the liability variable can be approximated by $\boldsymbol{z} \sim \text{MVN}(\alpha_0 \mathbf{1} + \boldsymbol{X}\boldsymbol{\alpha}_1 + \alpha_2 \boldsymbol{E}, \boldsymbol{I}_n)$ and $\hat{\boldsymbol{z}}$ is computed as the mean of the following truncated normal distribution (Crawford and Zhou, 2018)

$$z_k | y_k \sim \begin{cases} N(\alpha_0 + \boldsymbol{X}_k^T \boldsymbol{\alpha}_1 + \alpha_2 E_k, 1) & \text{with } z_k \geq 0 \quad \text{if } y_k = 1 \\ N(\alpha_0 + \boldsymbol{X}_k^T \boldsymbol{\alpha}_1 + \alpha_2 E_k, 1) & \text{with } z_k < 0 \quad \text{if } y_k = 0 \end{cases} \quad \text{for } k = 1, 2, \ldots, n.$$

The parameters $\alpha_0, \boldsymbol{\alpha}_1$ and α_2 are estimated using a probit model on the phenotype $\boldsymbol{y}$.

To test the interaction effects between a set of genetic variants and an environmental variable on the binary phenotype $\boldsymbol{y}$, we implement MAGEIT_RAN and MAGEIT_FIX on the estimate of the liability variable $\hat{\boldsymbol{z}}$. To construct MAGEIT_RAN, the liability threshold mixed effects model specified in Eqs (5) and (6) contains three variance components $(\omega^2, \sigma^2, \tau^2)$, where σ^2 represents a measure of interactions between the p genetic variants and the environmental variable. In order for the model to be identifiable, we put a constrain on the variance of $\boldsymbol{z}$, e.g., $\omega^2 + \sigma^2 + \tau^2 = 1$ (Lee *and others*., 2011). Similarly, we set $\sigma^2 + \tau^2 = 1$ for MAGEIT_FIX.

3. Simulation Studies

We conducted simulation studies to evaluate the performance of MAGEIT_RAN and MAGEIT_FIX to detect set-based $G \times E$ effects for both continuous and binary phenotypes, where the variant set contains both common and rare variants. We assessed type I error and empirical power of MAGEIT_RAN and MAGEIT_FIX, and compared them with three set-based $G \times E$ tests, GESAT-W (Lin *and others*., 2013), aMiSTi (Su, Di and Hsu, 2017), and ADABF (Lin *and* *others.*, 2019). These three existing methods are popular for $G \times E$ analysis and have well-developed R packages. For fair comparisons, the same weights for rare and common variants were used in all methods except ADABF which does not distinguish common and rare variants and hence no weights were used in the implementation.

3.1 Simulation settings

To generate genotypes, we first simulated 100,000 chromosomes over a 5 Kb region using a coalescent model that mimics the linkage disequilibrium (LD) structure and recombination rates of the European population (Schaffner *and others.*, 2005, Shlyakhter, Sabeti and Schaffner, 2014). Then we randomly selected 10 common variants with minor allele frequency (MAF) > 0.05 and 40 rare variants with 0.005 < MAF < 0.05 to compose a set of 50 genetic variants.

We simulated a continuous phenotype using the following trait model,

$$y_k = 0.05X_{k1} + 0.057X_{k2} + 0.64E_k + \sum_{j=1}^{10} w_{1j}\beta_j G_{kj} + \sum_{l=1}^{10} w_{2l}\gamma_l E_k G_{kl} + \varepsilon_k,$$

where $X_{k1} \sim \mathrm{N}(62.4, 11.5^2)$ mimicking age and $X_{k2} \sim \mathrm{Bernoulli}(0.52)$ mimicking sex (Lin *and others.*, 2013). The 10 genetic variants with main effects and the 10 variants with interaction effects were randomly selected from the set of the 50 variants, independent of E. The environmental variable E is a Bernoulli random variable taking values of 0 or 1 with a probability of 0.5. The weight of a rare variant in w_{1j} or w_{2l} is set to $\mathrm{Beta}(\mathrm{MAF}; 1, 25)$, the beta density function with parameters 1 and 25 evaluated at the variant's MAF, and the weight of a common variant in w_{1j} or w_{2l} is set to $c\mathrm{Beta}(\mathrm{MAF}; 0.5, 0.5)$ with $c = \frac{\mathrm{Beta}(0.05;\, 1{,}25)}{\mathrm{Beta}(0.05;\, 0.5{,}0.5)}$ (Ionita-Laza *and others.*, 2013, Madsen and Browning, 2009). The error term $\varepsilon_k \sim \mathrm{N}(0, 1.5^2)$ indicates independent noise.

For a binary trait, we use the following logistic regression model,

$$\text{logit}\big(P(y_k = 1)\big) = -6.2 + 0.05X_{k1} + 0.057X_{k2} + 0.64E_k + \sum_{j=1}^{10} w_{1j}\beta_j G_{kj} + \sum_{l=1}^{10} w_{2l}\gamma_l E_k G_{kl},$$

where all parameters are the same as those used in the continuous phenotype model. In all simulation settings, each simulated dataset contains 5,000 subjects (2,500 cases and 2,500 controls for binary phenotype).

In the type I error assessment, we set all γ_l to be 0, i.e., no $G \times E$ effects, and generated 10^6 datasets containing 50 genetic variants (10 common and 40 rare variants). We considered three scenarios: (1) no genetic main effect, i.e., $\beta_j = 0$ for $j = 1, 2, \dots, 10$; (2) for continuous/binary phenotype, assigning $\beta_j \sim \text{U}(0.07,\ 0.11) / \text{U}(0.08,\ 0.12)$ to two randomly selected common variants and $\beta_j \sim \text{U}(0.15,\ 0.19)/\text{U}(0.18,\ 0.22)$ to eight randomly selected rare variants; (3) similar to scenario (2) except that half of the common/rare variants have negative effects.

In the power comparison, we designed eight simulation scenarios that differ in three key factors that represent different considerations in the simulation design. The first factor pertains to the presence or absence of genetic main effects; the second factor focuses on the allocation of contributions from common and rare variants; and the third factor considers the direction of genetic main effects and $G \times E$ effects, either all positive effects or half positive and half negative effects. We considered ten variants with $G \times E$ effects, either two common and eight rare variants, or four common and six rare variants. The $G \times E$ effect γ_l was generated from $\text{U}(0.17,\ 0.21)$ and $\text{U}(0.57,\ 0.61)$ for common and rare variants, respectively, for continuous phenotype; and from $\text{U}(0.28,\ 0.32)$ and $\text{U}(0.86,\ 0.90)$ for common and rare variants, respectively, for binary phenotype. The first four simulation scenarios have no genetic main effect and they are as follow: (1) two common and eight rare variants with positive $G \times E$ effects; (2) two common and eight rare variants with $G \times E$ effects, 50% of $\gamma_j > 0$ and 50% of $\gamma_j < 0$; (3) four common and six rare variants with positive $G \times E$ effects; and (4) four common and six rare variants with $G \times E$ effects, 50% of $\gamma_j > 0$ and 50% of $\gamma_j < 0$. The remaining four simulation scenarios have two common and eight rare variants with genetic main effects: (5) β_j was specified the same as in scenario (2) in the type I error assessment, two common and eight rare variants with positive $G \times E$ effects; (6) β_j was specified the same as in scenario (3) in the type I error assessment, two common and eight rare variants with $G \times E$ effects, 50% of $\gamma_j > 0$ and 50% of $\gamma_j < 0$; (7) β_j was specified the same as in scenario (2) in the type I error assessment, four common and six rare variants with positive $G \times E$ effects; and (8) β_j was specified the same as in scenario (3) in the type I error assessment, four common and six rare variants with $G \times E$ effects, 50% of $\gamma_j > 0$ and 50% of $\gamma_j < 0$. Power was evaluated using 1,000 simulated datasets in each scenario.

3.2 Simulation results

Empirical type I error rate was calculated at the nominal level α, for $\alpha = 0.01$, 0.001 and 0.0001, based on 10^6 replicates, under three simulation scenarios, for both continuous and binary phenotypes (Table 1). In most simulations, the type I error of MAGEIT_FIX was within the 95% confidence interval of the nominal level, while the type I error of MAGEIT_RAN was lower than the nominal level in all simulation settings, especially for binary phenotype, suggesting that the MQS-based testing procedure tends to produce conservative p-values due to the approximation we used to handle binary phenotype (Crawford and Zhou, 2018, Schweiger *and others.*, 2017).

Empirical power was calculated at the significant level of 10^{-4}, based on 1,000 simulation replicates. Figures 1 and 2 demonstrate the power results of the five methods, MAGEIT_RAN, MAGEIT_FIX, GESAT-W, aMiSTi and ADABF, under eight simulation scenarios, for continuous and binary phenotypes, respectively. MAGEIT_RAN had comparable to higher power than the other methods across all simulation scenarios. We observed similar patterns for continuous and binary phenotypes. MAGEIT_RAN was much more powerful than other tests when there was no genetic main effect (Scenarios 1-4). For continuous traits, MAGEIT_FIX had comparable power to GESAT-W and higher power than aMiSTi in all simulation scenarios. For binary phenotypes, GESAT-W was comparable or more powerful than MAGEIT_FIX and ADABF. When the $G \times E$ effects had mixed positive and negative directions (Scenarios 2,4,6,8), aMiSTi had the lowest power for both continuous and binary phenotypes. Since aMiSTi is a combination of burden and variance component test, it loses power when there are both protective and detrimental variants in the genomic region being tested (Basu and Pan, 2011).

4. Application to MESA data

To demonstrate the utility of our proposed methods, we performed a genome-wide analysis of gene-alcohol interaction on hypertension in MESA (Bild *and others.*, 2002). MESA is a large longitudinal study of subclinical cardiovascular diseases including more than 6,800 participants. We analyzed the hypertension outcome measured at the first physical examination of 6,403 participants, consisting of 2,851 subjects with hypertension and 3,552 subjects without hypertension. The participants cover a diverse group of subjects including white (39.3%), African American (26.1%), Hispanic (22.5%), and Asian (12.1%). Alcohol usage (consumption of alcoholic beverages currently or formerly) was treated as an environmental variable, with 6,379 responses including 5,058 YESs and 1,321 NOs.

Samples were genotyped using the Affymetrix Human SNP Array 6.0. After data cleaning, IMPUTE2 (Howie, Donnelly and Marchini, 2009) was used for imputation with the 1000 Genome Phase 3 data as a reference panel. We excluded subjects whose proportion of successfully imputed variants < 5% or empirical inbreeding coefficients > 0.05, resulting in 6,424 subjects for further analysis. The following quality-control criteria were applied: (1) call rate > 95%, (2) MAF > 0.5%, and (3) Hardy-Weinberg χ^2 statistic p-value $> 10^{-6}$, resulting in a final set of 8,540,864 variants. In the gene-based $G \times E$ analysis, we restricted analysis on protein-coding regions based on the reference genome GRCh37 (Frankish *and others.*, 2019). In total, there were 18,977 genes on the 22 chromosomes and the number of variants in each gene region ranges from 2 to 5000, with a medium number of 383. Upon integrating the hypertension, alcohol usage and genotype data, a final set of 6,375 individuals are retained for downstream analyses.

4.1 Analysis of $G \times E$ effect

We performed genome-wide tests of gene-alcohol interaction effects on hypertension using all five methods, MAGEIT_RAN, MAGEIT_FIX, GESAT-W, aMiSTi, and ADABF. Age at the first exam, sex, and the top ten principal components (PCs) of the genetic relationship matrix were included in the analysis. The top ten PCs were calculated using the LD pruned variants with MAF > 0.05 to control for population structure.

MAGEIT_RAN and aMiSTi showed no evidence of inflation, with the genomic control inflation factors of 0.966 and 0.997, respectively. The $G \times E$ test assuming fixed genetic main effects, MAGEIT_FIX, and the Bayes factor-based test, ADABF, were conservative, with the genomic control inflation factors of 0.822 and 0.826, respectively. The genomic control inflation factor was 1.403 for GESAT-W. Therefore, we further adjusted the results of GESAT-W using genomic control.

No genes reached genome-wide significance at the p-value threshold of $\frac{0.05}{18{,}977} = 2.63 \times 10^{-6}$, commonly-used in gene-based analyses (Epstein *and others.*, 2015). Table 2 lists the top genes for which at least one of the five tests gives a p-value $< 10^{-4}$. The gene *CCNDBP1* had the smallest p-value, detected by MAGEIT_RAN (p-value = 2.80×10^{-5}) at a significance level of $\frac{1}{18{,}977} =$

5.27×10^{-5}, a suggestive significance threshold in genome-wide scan (Lander and Kruglyak, 1995). The p-value of *EPB42* (p-value = 5.98×10^{-5}) is close to the suggestive significance threshold, generated by MAGEIT_RAN. Both *CCNDBP1* and *EPB42* are located at 15q15.2. The cytogenetic region 15q15 has previously been reported to be associated with blood pressure (Kraja *and others.*, 2005). Moreover, *EPB42* was shown to be significantly down-regulated in heavy drinkers after exposed to psychological stress (Beech *and others.*, 2014, Chen *and others.*, 2021, Ma *and others.*, 2022).

4.2 Pathway analysis

Functional pathway analysis was conducted on genes that had $G \times E$ to identify enriched pathways related to hypertension, using MetaCore$^{\text{TM}}$. The top genes for which at least one of the five tests had a p-value $< 5 \times 10^{-3}$ were selected. Fisher's exact test was used to determine whether the gene list was enriched for a functional pathway. At the false discovery rate (FDR) < 0.05, there are 16 significant pathways that were reported to be related to hypertension (Table 3). Of particular interest are eight signaling pathways related to development and signal transduction that are relevant to hypertension and alcohol drinking. The first three pathways include a signal transduction pathway related to ERK1/2 signaling (p-value = 9.66×10^{-5}, FDR = 1.08×10^{-2}) and two development pathways related to activation of ERK by alpha-1 adrenergic receptors (p-value = 4.82×10^{-3}, FDR = 4.92×10^{-2}) and EPO-induced MAPK (p-value = 4.82×10^{-3}, FDR = 4.92×10^{-2}). Mitogen-activated protein kinases (MAPKs) are a group of serine/threonine kinases that include extracellular signal-regulated kinase 1/2 (ERK1/2), c-Jun N-terminal kinases (JNK1/2/3), and p38 (El-Mas and Abdel-Rahman, 2019). Highly conserved in eukaryotes, the MAPK signaling pathway has been implicated in cardiac remodeling and myocardial damage (Liu and Molkentin, 2016). For example, cardiac-specific ERK1/2 knockout mice developed cardiac dilation and eccentric growth of the heart (Kehat *and others.*, 2011), suggesting that ERK1/2 can regulate critical signal transduction pathways. In addition, p38 family members can participate in both protective and deleterious actions in the stress myocardium, demonstrating a key role for MAPKs proteins in cardiac physiology (Romero-Becerra *and others.*, 2020). The fourth pathway is a developmental module related to vascular endothelial growth factor (VEGA) signaling and activation (p-value = 3.97×10^{-3}, FDR = 4.50×10^{-2}). The VEGF signaling pathway plays a vital role in the vasculogenesis and angiogenesis in both embryo and adult (Zachary and Gliki, 2001). During neovascularization, VEGF is involved in gene expression, vascular permeability, and the migration, proliferation, and survival of cells (Shibuya, 2011). Studies indicated that VEGF blockade leads to endothelial dysfunction and the inhibition of VEGF-dependent vasodilatory pathways (Robinson *and others.*, 2010). These mechanisms together with the loss of microvascular capillary density through capillary rarefaction cause systemic vasoconstriction and hence resulting in hypertension (Robinson *and others.*, 2010). An animal study also suggested that physiologically relevant levels of alcohol consumption may associated with the stimulation of VEGF expression and angiogenesis (Tan *and others.*, 2007). The fifth one is a development pathway related to positive regulation of WNT/Beta-catenin signaling in the nucleus (p-value = 1.87×10^{-3}, FDR = 3.92×10^{-2}). WNT/Beta-catenin signaling pathway, also called canonical WNT signaling pathway in this context, is active in adult cardiac tissue after many cardiac injuries (Ozhan and Weidinger, 2015). WNT/Beta-catenin governs several elements of the renin-angiotensin system (RAS) containing angiotensinogen, renin, angiotensin-converting enzyme, and AT1 receptor (L Ruby *and others.*, 2010, Xiao *and others.*, 2019). Animal studies have shown that WNT2-deficient (another canonical Wnt signaling component) mice exhibit vascular abnormalities, abnormal vascular patterns, and increased vascular fragility (Cattelino *and others.*, 2003, Iso *and others.*, 2006). Interestingly, chronic ethanol consumption affects the activation of WNT/Beta-catenin signaling pathway and WNT/Beta-catenin directly controls alcohol-induced big potassium (BK) internalization (Mercer, Hennings and Ronis, 2015, Velázquez-Marrero *and others.*, 2016). The sixth one is a development pathway related to WNT and Notch signaling in early cardiac myogenesis (p-value = 2.35×10^{-3}, FDR = 4.00×10^{-2}). The Notch-mediated signaling, together with other signaling pathway such as VEGF and WNT, play a crucial role in vascular development and angiogenesis (Caliceti *and others.*, 2014). Alternations of Notch signaling is responsible for abnormal blood vessel and heart malformations (Zhou and Liu, 2014). Studies shown that in human coronary artery endothelial cells, ethanol activates notch pathway (Morrow *and others.*, 2010, Morrow *and others.*, 2014). The seventh pathway is a signal transduction pathway related to angiotensin II signaling via beta-arrestin (p-value = 9.17×10^{-4}, FDR = 3.00×10^{-2}). Angiotensin II (Ang II), a potent vasoconstrictor and a major effector molecule of renin-angiotensin system (RAS), participates in atherosclerosis and cardiovascular remodeling and raises blood pressure by exploiting various signaling cascades like WNT/beta-catenin (Benigni, Cassis and Remuzzi, 2010, Forrester *and others.*, 2018, Fyhrquist, Metsärinne and Tikkanen, 1995, Kawai *and others.*, 2017, Zhou and Liu, 2016). Animal studies reveal that angiotensinogen genes affect alcohol drinking behavior through Ang II (Fitts, 1993, Maul *and others.*, 2001). The last pathway is a signal transduction pathway related to adenosine A1 receptor signaling (p-value= 1.06×10^{-4}, FDR= 1.08×10^{-2}). Adenosine modulates cardiovascular function and produces bradycardia and hypotension when mediated systematically (Barraco *and others.*, 1987, Evoniuk, von Borstel and Wurtman, 1987). Activation of adenosine A1 receptor causes contraction of vascular smooth muscle and the adenosine A1 receptor agonists produce decreases in blood pressure and heart rate (Schindler *and others.*, 2005). It has been observed that raised adenosine levels mediate the ataxic and sedative/hypnotic effects of ethanol through activation of A1 receptors in the cerebellum, striatum, and cerebral cortex (L Ruby *and others.*, 2010). A1 agonists have been shown to decrease anxiety-like behavior, tremor, and seizures during acute ethanol withdrawal in mice (Kaplan *and others.*, 1999).

5. Conclusion

Human complex diseases are influenced by both genetic variation and interactions between gene and environmental factors. Many disease-associated genes have already been identified. Consequently, detecting and understanding gene-environment interactions becomes an important task for disease risk prediction (Hunter, 2005). In this study, we developed two methods MAGEIT_RAN and MAGEIT_FIX to detect the interaction between an environmental factor and gene sets where the genetic main effects were modeled as random or fixed, respectively. Both tests can be applied to continuous and binary phenotypes. Our methods is based on the MQS estimation (Zhou, 2017), which has been applied in MAPIT (Marginal ePIstasis Test) (Crawford *and others.*, 2017) and LT-MAPIT (liability threshold marginal epistasis test) (Crawford and Zhou, 2018) to detect gene-gene interactions. Our methods not only apply the MQS estimation to detect gene-environment interaction but also extends their methods by modeling genetic main effects as random in MAGEIT_RAN. Since variants in a genomic region can be either protective or deleterious and their effect sizes may vary, modeling genetic effects as random, as in MAGEIT_RAN, can capture different directions and magnitude of the genetic effects.

We compared the performance of MAGEIT_RAN and MAGEIT_FIX and three set-based $G \times E$ tests through simulations and real data analysis. In the simulation study, we demonstrated that MAGEIT_FIX had well-controlled type I error rate while MAGEIT_RAN was slightly conservative, especially for binary phenotypes, due to approximations used when specifying MAGEIT on binary phenotypes. MAGEIT_RAN was overall the most powerful among the five methods across all simulation settings. Application of MAGEIT_RAN and MAGEIT_FIX to the MESA hypertension data identified two genes, *CCNDBP1* and *EPB42*, located at the cytogenetic region 15q15.2 which has been reported to be associated with blood pressure. The *EPB42* gene was reported to be significantly down-regulated in heavy drinkers after exposed to psychological stress. Moreover, we identified 16 significant pathways that were related to hypertension, among them eight signaling transduction and development pathways are related to hypertension and alcohol usage. Given the established role of the genes and pathways we identified, MAGEIT has been shown to be able to detect biologically relevant genes that interact with environmental factors to influence complex traits.

There are several limitations in our methods. First, the type I error of MAGIT_RAN is conservative, particularly when there are genetic main effects, thus resulted in slight power loss in simulation scenarios 5-8. This is because the variation of the estimated variance component $\hat{\sigma}^2$ is larger when there are genetic main effects compared with no genetic main effect. Additionally, in MAGEIT_RAN, the regression coefficients β_j of the genetic main effects are assumed to be independent and γ_j of the $G \times E$ interactions as well. In reality, it is possible there exist correlations among these effects in a genomic region. This assumption may contribute to power loss, particularly in cases where most variants interact with the environmental factor and the effects of interactions are in the same direction. Nevertheless, considering the inherent complexities of linkage disequilibrium and haplotype effects, it is more appropriate to consider potential correlations among these coefficients. Given this, our model can be expanded to accommodate correlations among variants in a genomic region.

6. Software

Code to reproduce the results of the article is available at https://github.com/ZWang-Lab/MAGEIT.

FUNDING

This study was supported by National Science Foundation grant DMS1916246 and National Institutes of Health grants K01AA023321 and R01LM014087, and COBRE pilot grant GR13574.

Reference

ASCHARD, H., HANCOCK, D. B., LONDON, S. J. AND KRAFT, P. (2010). Genome-wide meta-analysis of joint tests for genetic and gene-environment interaction effects. *Human heredity* **70**, 292-300.

BARRACO, R. A., MARCANTONIO, D. R., PHILLIS, J. W. AND CAMPBELL, W. R. (1987). The effects of parenteral injections of adenosine and its analogs on blood pressure and heart rate in the rat. *General Pharmacology* **18**, 405-416.

BASU, S. AND PAN, W. (2011). Comparison of statistical tests for disease association with rare variants. *Genetic Epidemiology* **35**, 606-619.

BEECH, R. D., LEFFERT, J. J., LIN, A., HONG, K. A., HANSEN, J., UMLAUF, S., MANE, S., ZHAO, H. AND SINHA, R. (2014). Stress-Related Alcohol Consumption in Heavy Drinkers Correlates with Expression of mi R-10a, mi R-21, and Components of the TAR-RNA-Binding Protein-Associated Complex. *Alcoholism: Clinical and Experimental Research* **38**, 2743-2753.

BENIGNI, A., CASSIS, P. AND REMUZZI, G. (2010). Angiotensin II revisited: new roles in inflammation, immunology and aging. *EMBO molecular medicine* **2**, 247-257.

BHATNAGAR, A. (2017). Environmental determinants of cardiovascular disease. *Circulation research* **121**, 162-180.

BILD, D. E., BLUEMKE, D. A., BURKE, G. L., DETRANO, R., DIEZ ROUX, A. V., FOLSOM, A. R., GREENLAND, P., JACOBSJR, D. R., KRONMAL, R. AND LIU, K. (2002). Multi-ethnic study of atherosclerosis: objectives and design. *American journal of epidemiology* **156**, 871-881.

CALICETI, C., NIGRO, P., RIZZO, P. AND FERRARI, R. (2014). ROS, Notch, and Wnt signaling pathways: crosstalk between three major regulators of cardiovascular biology. *BioMed research international* **2014**.

CATTELINO, A., LIEBNER, S., GALLINI, R., ZANETTI, A., BALCONI, G., CORSI, A., BIANCO, P., WOLBURG, H., MOORE, R. AND OREDA, B. (2003). The conditional inactivation of the β-catenin gene in endothelial cells causes a defective vascular pattern and increased vascular fragility. *The Journal of cell biology* **162**, 1111-1122.

CHEN, H., MEIGS, J. B. AND DUPUIS, J. (2014). Incorporating gene-environment interaction in testing for association with rare genetic variants. *Human heredity* **78**, 81-90.

CHEN, J., YANG, C., YANG, Y., LIANG, Q., XIE, K., LIU, J. AND TANG, Y. (2021). Targeting DKK1 prevents development of alcohol-induced osteonecrosis of the femoral head in rats. *American Journal of Translational Research* **13**, 2320.

CHI, J. T., IPSEN, I. C., HSIAO, T.-H., LIN, C.-H., WANG, L.-S., LEE, W.-P., LU, T.-P. AND TZENG, J.-Y. (2021). SEAGLE: A Scalable Exact Algorithm for Large-Scale Set-Based Gene-Environment Interaction Tests in Biobank Data. *Frontiers in genetics* **12**.

COSSELMAN, K. E., NAVAS-ACIEN, A. AND KAUFMAN, J. D. (2015). Environmental factors in cardiovascular disease. *Nature Reviews Cardiology* **12**, 627-642.

CRAWFORD, L., ZENG, P., MUKHERJEE, S. AND ZHOU, X. (2017). Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. *PLoS genetics* **13**, e1006869.

CRAWFORD, L. AND ZHOU, X. (2018). Genome-wide marginal epistatic association mapping in case-control studies. *bioRxiv*, 374983.

DAVIES, R. B. (1980). Algorithm AS 155: The distribution of a linear combination of χ 2 random variables. *Applied Statistics*, 323-333.

EICHLER, E. E., FLINT, J., GIBSON, G., KONG, A., LEAL, S. M., MOORE, J. H. AND NADEAU, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature reviews genetics* **11**, 446-450.

EL-MAS, M. M. AND ABDEL-RAHMAN, A. A. (2019). Role of alcohol oxidative metabolism in its cardiovascular and autonomic effects. *Aldehyde dehydrogenases*, 1-33.

ENGEL, B., BUIST, W. AND VISSCHER, A. (1995). Inference for threshold models with variance components from the generalized linear mixed model perspective. *Genetics Selection Evolution* **27**, 15-32.

EPSTEIN, M. P., DUNCAN, R., WARE, E. B., JHUN, M. A., BIELAK, L. F., ZHAO, W., SMITH, J. A., PEYSER, P. A., KARDIA, S. L. AND SATTEN, G. A. (2015). A statistical approach for rare-variant association testing in affected sibships. *The American Journal of Human Genetics* **96**, 543-554.

EVONIUK, G., VON BORSTEL, R. W. AND WURTMAN, R. J. (1987). Antagonism of the cardiovascular effects of adenosine by caffeine or 8-(p-sulfophenyl) theophylline. *Journal of Pharmacology and Experimental Therapeutics* **240**, 428-432.

FITTS, D. A. (1993). Angiotensin and captopril increase alcohol intake. *Pharmacology Biochemistry and Behavior* **45**, 35-43.

FORRESTER, S. J., BOOZ, G. W., SIGMUND, C. D., COFFMAN, T. M., KAWAI, T., RIZZO, V., SCALIA, R. AND EGUCHI, S. (2018). Angiotensin II signal transduction: an update on mechanisms of physiology and pathophysiology. *Physiological reviews* **98**, 1627-1738.

FRANKISH, A., DIEKHANS, M., FERREIRA, A.-M., JOHNSON, R., JUNGREIS, I., LOVELAND, J., MUDGE, J. M., SISU, C., WRIGHT, J. AND ARMSTRONG, J. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic acids research* **47**, D766-D773.

FYHRQUIST, F., METSÄRINNE, K. AND TIKKANEN, I. (1995). Role of angiotensin II in blood pressure regulation and in the pathophysiology of cardiovascular disorders. *Journal of human hypertension* **9**, S19-24.

HOWIE, B. N., DONNELLY, P. AND MARCHINI, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics* **5**, e1000529.

HUNTER, D. J. (2005). Gene–environment interactions in human diseases. *Nature reviews genetics* **6**, 287-298.

IONITA-LAZA, I., LEE, S., MAKAROV, V., BUXBAUM, J. D. AND LIN, X. (2013). Sequence kernel association tests for the combined effect of rare and common variants. *The American Journal of Human Genetics* **92**, 841-853.

ISO, T., MAENO, T., OIKE, Y., YAMAZAKI, M., DOI, H., ARAI, M. AND KURABAYASHI, M. (2006). Dll4-selective Notch signaling induces ephrinB2 gene expression in endothelial cells. *Biochemical and biophysical research communications* **341**, 708-714.

JIAO, S., HSU, L., BÉZIEAU, S., BRENNER, H., CHAN, A. T., CHANG-CLAUDE, J., LE MARCHAND, L., LEMIRE, M., NEWCOMB, P. A. AND SLATTERY, M. L. (2013). SBERIA: set-based gene-environment interaction test for rare and common variants in complex diseases. *Genetic epidemiology* **37**, 452-464.

KAPLAN, G. B., BHARMAL, N. H., LEITE-MORRIS, K. A. AND ADAMS, W. R. (1999). Role of adenosine A1 and A2A receptors in the alcohol withdrawal syndrome. *Alcohol* **19**, 157-162.

KAWAI, T., FORRESTER, S. J., O'BRIEN, S., BAGGETT, A., RIZZO, V. AND EGUCHI, S. (2017). AT1 receptor signaling pathways in the cardiovascular system. *Pharmacological research* **125**, 4-13.

KEHAT, I., DAVIS, J., TIBURCY, M., ACCORNERO, F., SABA-EL-LEIL, M. K., MAILLET, M., YORK, A. J., LORENZ, J. N., ZIMMERMANN, W. H. AND MELOCHE, S. (2011). Extracellular signal-regulated kinases 1 and 2 regulate the balance between eccentric and concentric cardiac growth. *Circulation research* **108**, 176-183.

KHOURY, M. J. AND WACHOLDER, S. (2009). Invited commentary: from genome-wide association studies to gene-environment-wide interaction studies—challenges and opportunities. *American journal of epidemiology* **169**, 227-230.

KRAFT, P., YEN, Y.-C., STRAM, D. O., MORRISON, J. AND GAUDERMAN, W. J. (2007). Exploiting gene-environment interaction to detect genetic associations. *Human heredity* **63**, 111-119.

KRAJA, A. T., HUNT, S. C., PANKOW, J. S., MYERS, R. H., HEISS, G., LEWIS, C. E., RAO, D. AND PROVINCE, M. A. (2005). Quantitative trait loci for metabolic syndrome in the Hypertension Genetic Epidemiology Network study. *Obesity research* **13**, 1885-1890.

KUSS, M., RASMUSSEN, C. E. AND HERBRICH, R. (2005). Assessing Approximate Inference for Binary Gaussian Process Classification. *Journal of machine learning research* **6**.

L Ruby, C., A Adams, C., J Knight, E., Wook Nam, H. and Choi, D.-S. (2010). An essential role for adenosine signaling in alcohol abuse. *Current drug abuse reviews* **3**, 163-174.

Lander, E. and Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature genetics* **11**, 241-247.

Lee, S., Wu, M. C. and Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13**, 762-775.

Lee, S. H., Wray, N. R., Goddard, M. E. and Visscher, P. M. (2011). Estimating missing heritability for disease from genome-wide association studies. *The American Journal of Human Genetics* **88**, 294-305.

Lin, W.-Y., Huang, C.-C., Liu, Y.-L., Tsai, S.-J. and Kuo, P.-H. (2019). Genome-wide gene-environment interaction analysis using set-based association tests. *Frontiers in genetics* **9**, 715.

Lin, X., Lee, S., Christiani, D. C. and Lin, X. (2013). Test for interactions between a genetic marker set and environment in generalized linear models. *Biostatistics* **14**, 667-681.

Lin, X., Lee, S., Wu, M. C., Wang, C., Chen, H., Li, Z. and Lin, X. (2016). Test for rare variants by environment interactions in sequencing association studies. *Biometrics* **72**, 156-164.

Liu, H., Tang, Y. and Zhang, H. H. (2009). A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Computational Statistics & Data Analysis* **53**, 853-856.

Liu, R. and Molkentin, J. D. (2016). Regulation of cardiac hypertrophy and remodeling through the dual-specificity MAPK phosphatases (DUSPs). *Journal of molecular and cellular cardiology* **101**, 44-49.

MA, X., LIAO, Z., LI, R., XIA, W., GUO, H., LUO, J., SHENG, H., TIAN, M. AND CAO, Z. (2022). Myocardial Injury Caused by Chronic Alcohol Exposure—A Pilot Study Based on Proteomics. *Molecules* **27**, 4284.

MADSEN, B. E. AND BROWNING, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* **5**, e1000384.

MANNING, A. K., LAVALLEY, M., LIU, C. T., RICE, K., AN, P., LIU, Y., MILJKOVIC, I., RASMUSSEN-TORVIK, L., HARRIS, T. B. AND PROVINCE, M. A. (2011). Meta-analysis of gene-environment interaction: joint estimation of SNP and SNP× environment regression coefficients. *Genetic epidemiology* **35**, 11-18.

MAUL, B., SIEMS, W.-E., HOEHE, M. R., GRECKSCH, G., BADER, M. AND WALTHER, T. (2001). Alcohol consumption is controlled by angiotensin II. *The FASEB Journal* **15**, 1640-1642.

MERCER, K., HENNINGS, L. AND RONIS, M. (2015). Alcohol consumption, Wnt/β-catenin signaling, and hepatocarcinogenesis. *Biological Basis of Alcohol-Induced Cancer*, 185-195.

MORROW, D., CULLEN, J. P., LIU, W., CAHILL, P. A. AND REDMOND, E. M. (2010). Alcohol inhibits smooth muscle cell proliferation via regulation of the Notch signaling pathway. *Arteriosclerosis, thrombosis, and vascular biology* **30**, 2597-2603.

MORROW, D., HATCH, E., HAMM, K., CAHILL, P. A. AND REDMOND, E. M. (2014). Flk-1/KDR mediates ethanol-stimulated endothelial cell Notch signaling and angiogenic activity. *Journal of vascular research* **51**, 315-324.

OZHAN, G. AND WEIDINGER, G. (2015). Wnt/β-catenin signaling in heart regeneration. *Cell regeneration* **4**, 4: 3.

ROBINSON, E. S., KHANKIN, E. V., KARUMANCHI, S. A. AND HUMPHREYS, B. D. (Year). Hypertension induced by vascular endothelial growth factor signaling pathway inhibition:

mechanisms and potential use as a biomarker. Proceedings of the Seminars in nephrology, 591-601.

ROMERO-BECERRA, R., SANTAMANS, A. M., FOLGUEIRA, C. AND SABIO, G. (2020). p38 MAPK pathway in the heart: new insights in health and disease. *International Journal of Molecular Sciences* **21**, 7412.

SANTORICO, S. A. AND HENDRICKS, A. E. (Year). Progress in methods for rare variant association. Proceedings of the BMC genetics, 57-66.

SCHAFFNER, S. F., FOO, C., GABRIEL, S., REICH, D., DALY, M. J. AND ALTSHULER, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. *Genome research* **15**, 1576-1583.

SCHINDLER, C. W., KARCZ-KUBICHA, M., THORNDIKE, E. B., MÜLLER, C. E., TELLA, S. R., FERRÉ, S. AND GOLDBERG, S. R. (2005). Role of central and peripheral adenosine receptors in the cardiovascular responses to intraperitoneal injections of adenosine A1 and A2A subtype receptor agonists. *British journal of pharmacology* **144**, 642-650.

SCHWEIGER, R., WEISSBROD, O., RAHMANI, E., MÜLLER-NURASYID, M., KUNZE, S., GIEGER, C., WALDENBERGER, M., ROSSET, S. AND HALPERIN, E. (2017). RL-SKAT: an exact and efficient score test for heritability and set tests. *Genetics* **207**, 1275-1283.

SHIBUYA, M. (2011). Vascular endothelial growth factor (VEGF) and its receptor (VEGFR) signaling in angiogenesis: a crucial target for anti-and pro-angiogenic therapies. *Genes & cancer* **2**, 1097-1105.

SHLYAKHTER, I., SABETI, P. C. AND SCHAFFNER, S. F. (2014). Cosi2: an efficient simulator of exact and approximate coalescent with selection. *Bioinformatics* **30**, 3427-3429.

SU, Y.-R., DI, C.-Z. AND HSU, L. (2017). A unified powerful set-based test for sequencing data analysis of GxE interactions. *Biostatistics* **18**, 119-131.

TAN, W., BAILEY, A. P., SHPARAGO, M., BUSBY, B., COVINGTON, J., JOHNSON, J. W., YOUNG, E. AND GU, J.-W. (2007). Chronic alcohol consumption stimulates VEGF expression, tumor angiogenesis and progression of melanoma in mice. *Cancer biology & therapy* **6**, 1222-1228.

TEMPELMAN, R. AND GIANOLA, D. (1993). Marginal maximum likelihood estimation of variance components in Poisson mixed models using Laplacian integration. *Genetics Selection Evolution* **25**, 305-319.

THOMAS, D. (2010). Gene–environment-wide association studies: emerging approaches. *Nature reviews genetics* **11**, 259-272.

TZENG, J.-Y., ZHANG, D., PONGPANICH, M., SMITH, C., MCCARTHY, M. I., SALE, M. M., WORRALL, B. B., HSU, F.-C., THOMAS, D. C. AND SULLIVAN, P. F. (2011). Studying gene and gene-environment effects of uncommon and common variants on continuous traits: a marker-set approach using gene-trait similarity regression. *The American Journal of Human Genetics* **89**, 277-288.

VELÁZQUEZ-MARRERO, C., BURGOS, A., GARCÍA, J. O., PALACIO, S., MARRERO, H. G., BERNARDO, A., PÉREZ-LASPIUR, J., RIVERA-OLIVER, M., SEALE, G. AND TREISTMAN, S. N. (2016). Alcohol regulates BK surface expression via Wnt/β-catenin signaling. *Journal of Neuroscience* **36**, 10625-10639.

WANG, X., LIM, E., LIU, C. T., SUNG, Y. J., RAO, D. C., MORRISON, A. C., BOERWINKLE, E., MANNING, A. K. AND CHEN, H. (2020). Efficient gene–environment interaction tests for large biobank-scale sequencing studies. *Genetic Epidemiology* **44**, 908-923.

WILLIAMS, C. K. AND BARBER, D. (1998). Bayesian classification with Gaussian processes. *IEEE Transactions on pattern analysis and machine intelligence* **20**, 1342-1351.

WU, M. C., LEE, S., CAI, T., LI, Y., BOEHNKE, M. AND LIN, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics* **89**, 82-93.

XIAO, L., XU, B., ZHOU, L., TAN, R. J., ZHOU, D., FU, H., LI, A., HOU, F. F. AND LIU, Y. (2019). Wnt/β-catenin regulates blood pressure and kidney injury in rats. *Biochimica Et Biophysica Acta (BBA)-Molecular Basis of Disease* **1865**, 1313-1322.

ZACHARY, I. AND GLIKI, G. (2001). Signaling transduction mechanisms mediating biological actions of the vascular endothelial growth factor family. *Cardiovascular research* **49**, 568-581.

ZHOU, L. AND LIU, Y. (2016). Wnt/β-catenin signaling and renin-angiotensin system in chronic kidney disease. *Current opinion in nephrology and hypertension* **25**, 100.

ZHOU, X. (2017). A unified framework for variance component estimation with summary statistics in genome-wide association studies. *The annals of applied statistics* **11**, 2027.

ZHOU, X. AND LIU, J. (2014). Role of Notch signaling in the mammalian heart. *Brazilian Journal of Medical and Biological Research* **47**, 1-10.

Table 1. Empirical type I error of MAGEIT_RAN and MAGEIT_FIX, based on 10^6 replicates

Test	**Nominal Level**	**Continuous**			**Binary**		
		Scenario 1	**Scenario 2**	**Scenario 3**	**Scenario 1**	**Scenario 2**	**Scenario 3**
MAGEIT_RAN	0.01	**9.66×10^{-3}**	**8.85×10^{-3}**	**8.62×10^{-3}**	**9.51×10^{-3}**	**7.77×10^{-3}**	**8.47×10^{-3}**
	0.001	**8.17×10^{-4}**	**6.09×10^{-4}**	**5.30×10^{-4}**	**7.90×10^{-4}**	**4.92×10^{-4}**	**5.04×10^{-4}**
	0.0001	**6.70×10^{-5}**	**2.90×10^{-5}**	**3.40×10^{-5}**	**6.20×10^{-5}**	**2.80×10^{-5}**	**2.20×10^{-5}**
MAGEIT_FIX	0.01	9.87×10^{-3}	9.98×10^{-3}	1.02×10^{-2}	**9.68×10^{-3}**	**9.67×10^{-3}**	**9.70×10^{-3}**
	0.001	9.89×10^{-4}	9.99×10^{-4}	9.57×10^{-4}	9.38×10^{-4}	9.58×10^{-4}	**8.99×10^{-4}**
	0.0001	1.01×10^{-4}	9.80×10^{-5}	8.80×10^{-5}	9.00×10^{-5}	8.70×10^{-5}	9.20×10^{-5}

The 95% confidence interval of a nominal level α was calculated as $\alpha \pm 1.96\sqrt{\alpha(1-\alpha)/10^6}$. Specifically, the 95% confidence intervals are $(9.80 \times 10^{-3}, 1.02 \times 10^{-2})$ for $\alpha = 0.01$, $(9.38 \times 10^{-4}, 1.06 \times 10^{-3})$ for $\alpha = 0.001$, and $(8.04 \times 10^{-5}, 1.20 \times 10^{-4})$ for $\alpha = 0.0001$. Rates outside of the 95% confidence interval are in bold.

Table 2. Genes with p-value < 10^{-4} in at least one of the tests in the MESA data

Chr	Gene	# SNP	Region	MAGEIT_RAN	MAGEIT_FIX	GESAT-W	aMiSTi	ADABF
15	*CCNDBP1*	237	15q15.2	**2.80×10^{-5}**	2.03×10^{-3}	6.28×10^{-3}	3.32×10^{-2}	4.90×10^{-2}
	EPB42	269	15q15.2	**5.98×10^{-5}**	2.05×10^{-3}	1.12×10^{-2}	3.97×10^{-2}	1.00×10^{-1}

The smallest p-values among the five tests at the given genes are in bold.

Table 3. Pathways with FDR < 0.05 in the MESA data

Pathway	P-value	FDR
Signal transduction_ERK1/2 signaling pathway	9.66×10^{-5}	1.08×10^{-2}
Signal transduction_Adenosine A1 receptor signaling pathway	1.06×10^{-4}	1.08×10^{-2}
Signal transduction_Adenosine A3 receptor signaling pathway	4.76×10^{-4}	2.16×10^{-2}
Development_Thromboxane A2 signaling pathway	5.57×10^{-4}	2.17×10^{-2}
Translation_Translation regulation by Alpha-1 adrenergic receptors	6.96×10^{-4}	2.47×10^{-2}
Signal transduction_S1P2 receptor inhibitory signaling	8.58×10^{-4}	2.92×10^{-2}
Signal transduction_Angiotensin II signaling via Beta-arrestin	9.17×10^{-4}	3.00×10^{-2}
Regulation of CFTR activity (normal and CF)	1.26×10^{-3}	3.21×10^{-2}
Development_Positive regulation of WNT/Beta-catenin signaling in the nucleus	1.87×10^{-3}	3.92×10^{-2}
Signal transduction_S1P1 receptor signaling	2.13×10^{-3}	3.92×10^{-2}
Development_WNT and Notch signaling in early cardiac myogenesis	2.35×10^{-3}	4.00×10^{-2}
Nociception_Nociceptin receptor signaling	2.67×10^{-3}	4.19×10^{-2}
Development_VEGF signaling and activation	3.97×10^{-3}	4.50×10^{-2}
Development_Estrogen-independent activation of ESR1 and ESR2	4.53×10^{-3}	4.80×10^{-2}
Development_Activation of ERK by Alpha-1 adrenergic receptors	4.82×10^{-3}	4.92×10^{-2}
Development_EPO-induced MAPK pathway	4.82×10^{-3}	4.92×10^{-2}

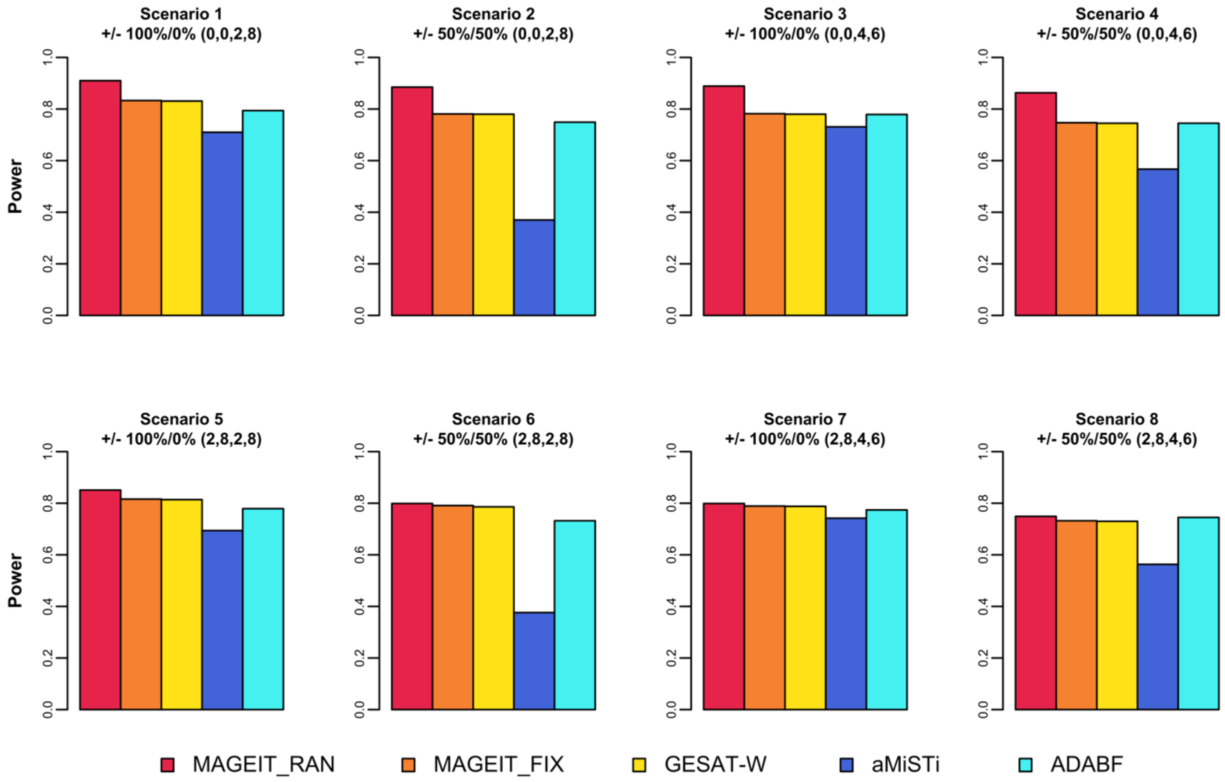

Figure 1. Empirical power of MAGIT_RAN, MAGIT_FIX, GESAT-W, aMiSTi and ADABF for a continuous phenotype.

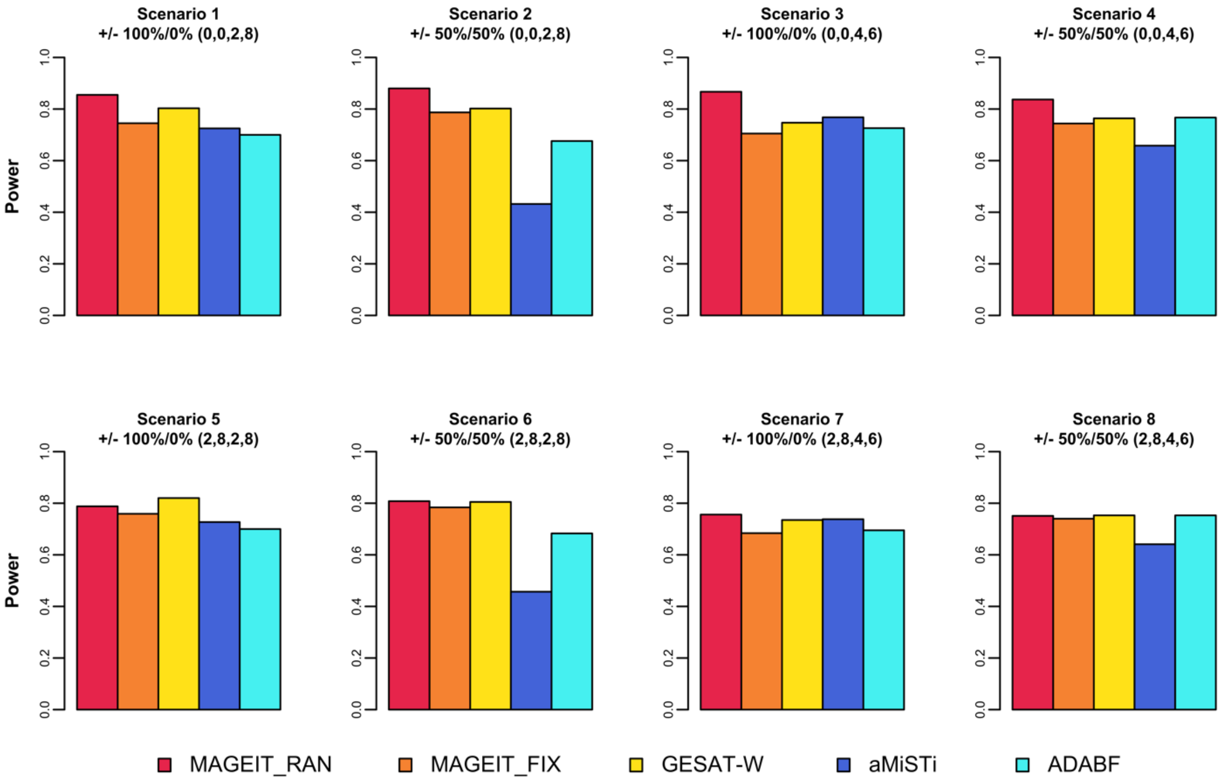

Figure 2. Empirical power of MAGIT_RAN, MAGIT_FIX, GESAT-W, aMiSTi and ADABF for a binary phenotype.