Bias in genome-wide association test statistics due to omitted interactions

Burak Yelmen[1,*], Merve Nur Güler[1], Estonian Biobank Research Team[1], Tõnu Kollo[2], Märt Möls[2], Guillaume Charpiat[3], Flora Jay[3]

1 Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia
2 Institute of Mathematics and Statistics, University of Tartu, Tartu, Estonia
3 CNRS, INRIA, LISN, Paris-Saclay University, Orsay, France

*burak.yelmen@ut.ee

Abstract

Over the past two decades, genome-wide association studies (GWAS) enabled the discovery of thousands of variants associated with many complex human traits. However, conventional GWAS are still widely performed with linear models with the assumption that the genetic effects are predominantly additive. In this work, we investigate the test statistic behavior when linear models are used to obtain significant genotype-phenotype associations without accounting for epistasis. We first algebraically derive mean and variance shift in the null statistic due to the omitted interaction term, and define the boundary between conservative (i.e., deflated statistic tail) and anti-conservative (i.e., inflated statistic tail) regimes for the common GWAS significance threshold. We then perform phenotype simulation analyses using the Estonian Biobank genotypes and validate the mathematical model. We demonstrate that the anti-conservative regime is plausible under realistic parameter settings and models omitting interaction terms can produce spurious significance. Our findings suggest caution when interpreting statistically significant signals reported in the literature based on linear models, especially for large-scale GWAS.

Keywords: genome-wide association studies, epistasis, complex traits, linear mixed models

Introduction

Biological systems consist of multiple interrelated components, which makes them difficult to model faithfully using linear functions. It is reasonable to assume that the relationship between genetics and complex traits exhibits similar complexity. Indeed, multiple studies have proposed epistasis and gene-environment interactions as fundamental causal mechanisms of phenotypic variance [1–3]. Genome-wide association studies (GWAS), despite allowing the discovery of thousands of genotype-phenotype links with functional importance, are still widely performed with linear models after almost two decades [4]. This prevailing methodological trend might be attributed to multiple factors such as the ease of interpretability, attainable computational requirements, tradition and studies suggesting that the genetic component of phenotypic variance can mainly be explained by additive models [5–7].

This assumption of mainly additive effects, however, may hinder the efforts to understand the causal genetic mechanisms since the true associations might be nonlinear, despite being captured without nonlinear modeling [8]. Even if we accept the mainly additive model for association discovery, recent methods point to large variability of non-additive variance (i.e., the left-over variance after accounting for additive variance) among phenotypes [9]. In this light, various works have either tried to approach GWAS with nonlinear models such as neural networks [10, 11] or explicitly infer/integrate epistasis [12–14].

There are some hints in the literature on the consequences of omitting interaction terms in GWAS. Omitted variable bias, which is the bias in model-estimated coefficients due to the exclusion of variables relevant to the response variable, is a well known statistical concept [15] and it has been used to demonstrate that GWAS summary statistics might tag non-additive variance whenever there is nonzero correlation between genotypes and the interaction terms [9]. Furthermore, some evidence was presented in a recent study for "synthetic associations" (i.e., associations due to combined effects of multiple other variants [16, 17]) due to epistasis [18]. Interestingly and perhaps alarmingly, it was suggested that 3-5% of GWAS Catalog peaks might be synthetic, but the findings were obtained from indirect evidence based on unusual linkage disequilibrium (LD) patterns and did not necessarily point to spurious associations due to epistasis. Here we use "spurious" to refer to associations arising purely due to epistasis which does not include the target SNP. To the best of our knowledge, the effects of omitting interaction terms in GWAS have not been modeled to understand whether this omission can result in spuriously significant associations under realistic settings.

In this work, we derive a mathematical model to assess the effects of omitted interaction terms on GWAS test statistics. We initially define a true data generating process (DGP) for a phenotype with a fixed interaction component. Our mathematical setup is that a standard linear mixed model (LMM) is fitted for this phenotype to estimate the target SNP coefficient α. We preprocess all vectors from the true DGP and the fitted model to whiten and remove covariate effects. Based on the equality of the phenotype value for both models after preprocessing, we derive ordinary least squares (OLS) estimates for the expected value of the estimated $\hat{\alpha}$ and its variance, under the null hypothesis $\alpha = 0$. These allow us to derive the mean and variance of the true null statistic as opposed to the nominal null $\mathcal{N}(0, 1)$. We then define the $R(x)$ conservativeness ratio of true to nominal p-value to assess anti-conservative and conservative regimes for a given critical value x. To ensure that both the SNP and interaction effects are clearly defined, we introduce an extended null concept, strict no-path null, in which not only the target SNP coefficient is 0, but also the target SNP cannot be part of the interaction features. Intuitively, this makes sure that any identified bias of the test statistics is purely due to the realized interaction component and its correlation with the target SNP. Using this strict null concept, we simulate two sets of 10,000 phenotypes based on the Estonian Biobank genotype dataset [19] with varying non-additive variance fraction estimates from the literature, use REGENIE [20] to obtain summary statistics for target SNPs and confirm the mathematical model. Both the mathematical model and the simulation results demonstrate that even under minimalistic non-additive variance estimates and low correlation between the target SNP and the interaction component, anti-conservative regime is plausible and multiple non-associated SNPs are identified as significant due to the omitted interaction term. The key interpretation of our findings is that widely used linear models which do not compensate for epistatic effects can be prone to identifying spurious associations, especially when the sample size is large.

Materials and Methods

True data-generating process (DGP). For individuals $i = 1, \ldots, n$, let $y \in \mathbb{R}^n$ denote the quantitative trait, $X \in \mathbb{R}^{n \times q}$ the observed covariates (e.g., age, sex, principal components etc.) and $g \in \mathbb{R}^n$ the target SNP. The true DGP is

$$y \;=\; X\beta \;+\; \alpha\, g \;+\; u \;+\; \varepsilon, \qquad \varepsilon \sim \mathcal{N}\big(0,\; \sigma_e^2 I + \sigma_\eta^2 K\big), \qquad \varepsilon \;\perp\!\!\!\perp\; (X, g, u), \tag{1}$$

where $u \in \mathbb{R}^n$ is the realized (fixed) mean contribution of possibly high-dimensional interaction terms (i.e., $u = Z\beta_u$ for some matrix $Z \in \mathbb{R}^{n \times m}$ whose columns encode SNP–SNP interaction terms) and K is the genetic relatedness matrix (GRM). We treat (X, g, u) as fixed and ε (containing environmental and additive polygenic effects) as stochastic.

Fitted (misspecified) model. We fit the standard linear mixed model (LMM)

$$y \;=\; X\beta \;+\; \alpha\, g \;+\; \varepsilon, \qquad \varepsilon \sim \mathcal{N}\big(0,\; \sigma_e^2 I + \sigma_\eta^2 K\big), \tag{2}$$

and examine the test $H_0 : \alpha = 0$ performed under (2) when the true DGP is (1).

Preprocessing. Define a single linear operator T that whitens and removes covariates:

$$V \equiv \sigma_\eta^2 K + \sigma_e^2 I, \qquad L \in \mathbb{R}^{n \times n} \text{ with } L^\top L = V^{-1}, \qquad M \equiv I - LX\big(X^\top V^{-1} X\big)^{-1} X^\top L^\top,$$
$$T \equiv ML,$$

where L is the whitening matrix and M is the orthogonal projector, in the whitened space, that removes the subspace containing the covariate component X, that is, $TX = 0$. Note that $M^2 = M$. We assume V is (ideally) known/estimated correctly in a generalized least squares (GLS) framework and apply T to every vector in (1) and (2). From now on, we overwrite the symbols

$$y \leftarrow Ty, \qquad g \leftarrow Tg, \qquad u \leftarrow Tu, \qquad \varepsilon \leftarrow T\varepsilon,$$

and work entirely with these transformed variables. In this space, the noise is spherical on the working subspace: $\varepsilon := T\,\mathcal{N}(0, V) = M\mathcal{N}(0, I)$ and consequently $\mathrm{Cov}_\varepsilon(\varepsilon) = M^T M = M$. The fitted (misspecified) model (Eq. 2) thus becomes:

$$y \;=\; \alpha\, g \;+\; \varepsilon, \qquad \varepsilon \sim \mathcal{N}\big(0,\; M\big)\;. \tag{3}$$

Moreover, this noise is almost unit-normed on average:

$$\frac{1}{n}\, \mathop{\mathbb{E}}_{\varepsilon} \big\|\varepsilon\big\|^2 = \frac{\mathrm{tr}(M)}{n} = \frac{n-q}{n},$$

which is close to 1 for large n, as $q \ll n$ in typical set-ups. Note that under $H_0 : \alpha = 0$, the (true) working model reduces to

$$y \;=\; u \;+\; \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, M). \tag{4}$$

Moments. Let us consider the empirical variance of g_i and u_i with respect to individuals i:

$$\mathop{\mathrm{Var}}_i(g) = \frac{1}{n}\|g\|^2, \quad \mathop{\mathrm{Var}}_i(u) = \frac{1}{n}\|u\|^2, \quad \mathop{\mathrm{Cov}}_i(g, u) = \frac{1}{n} g^\top u, \quad \rho \equiv \mathop{\mathrm{Corr}}_i(g, u) = \frac{\mathop{\mathrm{Cov}}_i(g, u)}{\sqrt{\mathop{\mathrm{Var}}_i(g)\; \mathop{\mathrm{Var}}_i(u)}},$$

supposing centered variables. All moments are computed in the preprocessed space.

Expected value of the coefficient. The no-intercept OLS estimate of α from the misspecified model (Eq. 3) is

$$\hat{\alpha} \;=\; \frac{g^\top y}{g^\top g} = \frac{g^\top (u+\varepsilon)}{g^\top g} = \frac{g^\top u}{g^\top g} \;+\; \frac{g^\top \varepsilon}{g^\top g}.$$

Hence, expected value of $\hat{\alpha}$ over ε is

$$\boxed{\mathop{\mathbb{E}}_{\varepsilon}[\hat{\alpha}] \;=\; \frac{\mathop{\mathrm{Cov}}_{i}(g,u)}{\mathop{\mathrm{Var}}_{i}(g)}}. \tag{5}$$

Residual variance. The OLS residuals between the observed y (after preprocessing) and their explained component $\hat{\alpha} g$ from the linear regression with the target SNP g are:

$$\tilde{\varepsilon}_{OLS} \;=\; y - \hat{\alpha} g = (u+\varepsilon) - \left(\tfrac{g^\top u}{g^\top g} + \tfrac{g^\top \varepsilon}{g^\top g}\right) g = \left(u - \tfrac{g^\top u}{g^\top g}\, g\right) + \left(\varepsilon - \tfrac{g^\top \varepsilon}{g^\top g}\, g\right).$$

The two summands are uncorrelated on average over ε, yet spuriously slightly empirically correlated. Thus the residual empirical variance is the sum of their variances, plus a small covariance that vanishes with sample size:

(a) Contribution from u:

$$\mathop{\mathrm{Var}}_{i}\!\left(u - \tfrac{g^\top u}{g^\top g}\, g\right) = \mathop{\mathrm{Var}}_{i}(u) - \frac{\mathop{\mathrm{Cov}}_{i}(g,u)^2}{\mathop{\mathrm{Var}}_{i}(g)} = \mathop{\mathrm{Var}}_{i}(u)\,\left(1-\rho^2\right).$$

(b) Contribution from ε: the empirical variance of the second term is

$$\mathop{\mathrm{Var}}_{i}\left(\varepsilon - \tfrac{g^\top \varepsilon}{g^\top g}\, g\right) \;=\; \frac{1}{n}\left\|\varepsilon - \frac{g^\top \varepsilon}{g^\top g}\, g\right\|^2 \;=\; \frac{1}{n}\left(\varepsilon^\top \varepsilon - 2\,\tfrac{g^\top \varepsilon}{g^\top g}\,\varepsilon^\top g + \left(\tfrac{g^\top \varepsilon}{g^\top g}\right)^2 g^\top g\right) \;=\; \frac{1}{n}\left(\|\varepsilon\|^2 - \tfrac{(g^\top \varepsilon)^2}{g^\top g}\right).$$

As $\mathop{\mathbb{E}}_{\varepsilon}\left[\|\varepsilon\|^2\right] = \mathrm{tr}(M) = n-q$, and $\mathop{\mathbb{E}}_{\varepsilon}\left[(g^\top \varepsilon)^2\right] = \mathop{\mathbb{E}}_{\varepsilon}\left[g^\top(\varepsilon\varepsilon^\top)g\right] = g^\top M g = g^\top g$ (since g is obtained by applying $T = ML$ as preprocessing and $M^2 = M$), we obtain that on average, this variance over ε is:

$$\mathop{\mathbb{E}}_{\varepsilon}\left[\; \mathop{\mathrm{Var}}_{i}\left(\varepsilon - \tfrac{g^\top \varepsilon}{g^\top g}\, g\right)\right] \;=\; \frac{n-q-1}{n} \;\xrightarrow[n\to\infty]{}\; 1.$$

Moreover, when the number n of individuals is large, for any given realization of ε, the empirical variance (b) is likely to be very close to this average, because the variance w.r.t. ε vanishes (see proof in Appendix A):

$$\mathop{\mathrm{Var}}_{\varepsilon}\left[\; \mathop{\mathrm{Var}}_{i}\left(\varepsilon - \tfrac{g^\top \varepsilon}{g^\top g}\, g\right)\right] \;=\; \frac{2n+q^2-1}{n^2} \;=\; \underset{n\to+\infty}{O}\left(\frac{1}{n}\right) \;\xrightarrow[n\to\infty]{}\; 0.$$

(c) The spurious cross-contribution from u and ε:

$$\mathop{\mathrm{Cov}}_{i}\left(u - \tfrac{g^\top u}{g^\top g}\, g,\; \varepsilon - \tfrac{g^\top \varepsilon}{g^\top g}\, g\right) \;=\; \frac{1}{n}\, u^T \left(I - \frac{gg^T}{g^T g}\right)^2 \varepsilon$$

similarly vanishes with high sample size n. Indeed, its average over ε is 0, as ε is independent from u and g, and its variance w.r.t. ε also tends to be 0, being equal to $\frac{1}{n^2}\, u^T \left(I - \frac{gg^T}{g^T g}\right) M \left(I - \frac{gg^T}{g^T g}\right) u \;\leqslant\; \frac{1}{n^2}\, u^T \left(I - \frac{gg^T}{g^T g}\right) u \;=\; \frac{1-\rho^2}{n}\, \mathop{\mathrm{Var}}_{i}(u)$, using the fact that the operators $I - \frac{gg^T}{g^T g}$ and M are projectors.

Hence the (asymptotic) total residual variance is

$$\boxed{\sigma^2_{\mathrm{res}} \;=\; 1 \;+\; \mathop{\mathrm{Var}}_{i}(u)\,\left(1-\rho^2\right)}. \tag{6}$$

Standard error. For no-intercept single-regressor OLS, asymptotically

$$\boxed{\mathrm{SE}^2(\hat{\alpha}) = \frac{\sigma^2_{\mathrm{res}}}{g^\top g} = \frac{1 + \operatorname*{Var}_i(u)\big(1-\rho^2\big)}{n \operatorname*{Var}_i(g)}}. \tag{7}$$

The t-statistic. The null hypothesis $\alpha = 0$ is tested with the t-statistic $t = \hat{\alpha}/\mathrm{SE}(\hat{\alpha})$. Under the working model, randomness arises only from ε, while u and g are fixed, so $\hat{\alpha}$ OLS estimator can be written as

$$\hat{\alpha} \;=\; \operatorname*{\mathbb{E}}_{\varepsilon}[\hat{\alpha}] + \frac{g^\top \varepsilon}{g^\top g}, \qquad \operatorname*{\mathbb{E}}_{\varepsilon}[\hat{\alpha}] = \frac{g^\top u}{g^\top g}.$$

Therefore, conditionally on (g, u), the sampling variance of $\hat{\alpha}$ is

$$\operatorname*{Var}_{\varepsilon}(\hat{\alpha}) = \operatorname*{Var}_{\varepsilon}\left(\frac{g^\top \varepsilon}{g^\top g}\right) = \frac{\operatorname*{\mathbb{E}}_{\varepsilon}[(g^\top \varepsilon)^2]}{(g^\top g)^2} = \frac{g^\top \operatorname*{\mathbb{E}}_{\varepsilon}[\varepsilon\varepsilon^\top] g}{(g^\top g)^2} = \frac{g^\top M g}{(g^\top g)^2} = \frac{1}{g^\top g}.$$

The OLS standard error, however, is based on the residual variance σ^2_{res} from (6) as derived in (7). Therefore, the variance of the t-statistic is asymptotically

$$\boxed{\operatorname*{Var}_{\varepsilon}(t) \;=\; \frac{\operatorname*{Var}_{\varepsilon}(\hat{\alpha})}{\mathrm{SE}^2(\hat{\alpha})} \;=\; \frac{1}{\sigma^2_{\mathrm{res}}} \;=\; \frac{1}{1 + \operatorname*{Var}_i(u)\big(1-\rho^2\big)}}. \tag{8}$$

Consequently, the t-statistic is asymptotically $\mathcal{N}\big(\mu,\ 1/\sigma^2_{\mathrm{res}}\big)$ with

$$\boxed{\mu \;=\; \frac{\operatorname*{\mathbb{E}}_{\varepsilon}[\hat{\alpha}]}{\mathrm{SE}(\hat{\alpha})} \;=\; \sqrt{n}\,\frac{\rho\sqrt{\operatorname*{Var}_i(u)}}{\sqrt{1 + \operatorname*{Var}_i(u)\big(1-\rho^2\big)}}}. \tag{9}$$

Equivalent parameterization. Since the average coordinate-wise variance from ε is asymptotically 1 after preprocessing as in (6), we can define a parameter $\lambda \equiv \operatorname*{Var}_i(u)/(1 + \operatorname*{Var}_i(u)) \in [0, 1)$, which is the proportion of variance explained by the interaction term. Then

$$\boxed{\mu \;=\; \frac{\rho\sqrt{\lambda n}}{\sqrt{1-\lambda\rho^2}}, \qquad \sigma^2_{\mathrm{res}} = \frac{1-\lambda\rho^2}{1-\lambda}}. \tag{10}$$

Conservativeness. Let Φ denote the cumulative distribution function (CDF) of the standard normal distribution. Under the nominal null distribution $\mathcal{N}(0, 1)$, the two-sided p-value for an observed statistic x is

$$p_{\mathrm{nom}}(x) = 2\,\Phi\big(-|x|\big).$$

Under the true null distribution $\mathcal{N}\big(\mu,\ 1/\sigma^2_{\mathrm{res}}\big)$, the actual two-sided tail probability is

$$p_{\mathrm{true}}(x) = \mathbb{P}\big(|t| \geq |x|\big) = \Phi\big(-\sigma_{\mathrm{res}}(|x| - \mu)\big) + \Phi\big(-\sigma_{\mathrm{res}}(|x| + \mu)\big).$$

Then define the conservativeness measure

$$\boxed{R(x) \;=\; \frac{p_{\mathrm{true}}(x)}{p_{\mathrm{nom}}(x)} \;\lesseqgtr\; 1}, \tag{11}$$

where $R(x) > 1$ indicates anti-conservativeness due to the shift in null mean and variance, and $R(x) < 1$ indicates conservativeness for a given statistic value x. In other words, for a critical value x (e.g., ± 5.45 corresponding to common GWAS significance threshold), $R(x) > 1$ means that observing x is more likely under the true null distribution, resulting in spurious significance.

Strict no-path null. After preprocessing, decompose u as $u = u_{\neg g} + u_g$, where $u_{\neg g}$ aggregates interaction components not involving the target SNP and u_g those that do. Define the strict no-path null

$$H_0^{\mathrm{NP}}: \quad \alpha = 0 \ \text{ and } \ u_g \equiv 0.$$

Under H_0^{NP} we have $u = u_{\neg g}$, so all expressions remain valid with u replaced by $u_{\neg g}$. Thus, any bias in the null t-statistic arises purely from alignment between g and omitted interaction signal that does not explicitly include g and does not involve any pathway in which g participates. Because g can be in LD with loci that contribute to $u_{\neg g}$, the correlation $\rho_{\neg g}$ can be nonzero even under this strict no-path null. Conversely, if $u_g \neq 0$ (interactions that include g), then the shift includes a genuine path from g to y through nonlinear terms; in that case the single-variant test detects an association but its coefficient targets a best linear approximation rather than the additive parameter α. Therefore, in simulation analyses, we used this strict no-path null concept by excluding the target SNPs from the interaction features, to demonstrate only the effects of pure bias due to the omitted interaction term.

Upper bound on ρ. We can provide plausible ranges for sample size n and the variance ratio λ from the literature, but ρ is difficult to estimate because the true interaction component for a given phenotype is unknowable. We can nevertheless obtain an upper bound without specifying interaction coefficients. Let $Z \in \mathbb{R}^{n\times m}$ collect the allowed interaction features (e.g., two-way or three-way products), excluding any feature that explicitly uses the target SNP g (strict no-path). All columns are already in the preprocessed space (i.e., $Z \leftarrow TZ$ and $g \leftarrow Tg$ were applied upstream). Define the column space $\mathcal{H} \equiv \mathrm{col}(Z)$; under strict no-path we posit $u \in \mathcal{H}$. We seek the largest possible correlation between g and any interaction signal u consistent with the specification,

$$\rho_{\max} \ \equiv \ \sup\big\{\, |\operatorname*{Corr}_i(g,u)| \ : \ u \in \mathcal{H},\ u \neq 0 \,\big\}.$$

Let $P_{\mathcal{H}}$ be the orthogonal projector onto $\mathcal{H}$ (in the Euclidean inner product of the preprocessed space). Then

$$\boxed{\rho_{\max} \ = \ \frac{\|P_{\mathcal{H}} g\|}{\|g\|} \ = \ \sqrt{\frac{g^\top P_{\mathcal{H}} g}{g^\top g}}\ .} \tag{12}$$

Proof. Correlation is scale-free, so set $\|u\| = 1$. Then

$$|\operatorname*{Corr}_i(g,u)| \ = \ \frac{|g^\top u|}{\|g\|\,\|u\|} \ = \ \frac{|g^\top u|}{\|g\|}, \qquad u \in \mathcal{H},\ \|u\| = 1.$$

Decompose g as $g = P_{\mathcal{H}} g + (I - P_{\mathcal{H}})g$, with $P_{\mathcal{H}} g \in \mathcal{H}$ and $(I - P_{\mathcal{H}})g \perp \mathcal{H}$. For any $u \in \mathcal{H}$,

$$g^\top u \ = \ (P_{\mathcal{H}} g)^\top u \ + \ \underbrace{((I - P_{\mathcal{H}})g)^\top u}_{=\,0} \ = \ (P_{\mathcal{H}} g)^\top u.$$

By Cauchy–Schwarz inequality, $|(P_{\mathcal{H}} g)^\top u| \leq \|P_{\mathcal{H}} g\|\,\|u\| = \|P_{\mathcal{H}} g\|$, with equality for $u = \pm P_{\mathcal{H}} g / \|P_{\mathcal{H}} g\|$ whenever $P_{\mathcal{H}} g \neq 0$. Hence

$$\sup_{\substack{u' = u/\|u\| \\ u \in \mathcal{H}}} \frac{|g^\top u'|}{\|g\|} \ = \ \frac{\|P_{\mathcal{H}} g\|}{\|g\|}.$$

Practical computation of ρ_{max}. We compiled a dataset consisting of 210,145 samples with chromosomes 21 and 22 (total 8170 bi-allelic SNPs from GSA SNP array framework) from the Estonian Biobank [19] and mean-centered all SNPs. We used a list of unrelated individuals obtained with King v2.2.7 [21] to subsample to 100,144 individuals (removing up to second degree relatives). Using this filtered dataset, we can assume V has no structure. Therefore, we only residualized genotypes and Z to remove covariates (age, sex, first 5 principal components) and work approximately in the preprocessed space without the

need for V estimation. We designed two different types of Z interaction matrices: one where target SNP and the interaction features are on the same chromosome and one where target SNP and interaction features are on different chromosomes. In both cases, the interaction feature matrix Z was generated randomly for 100 times with 100 two-way interaction terms, with the constraint that the target SNPs g are not part of these features (resulting in 383400 and 413500 total tested target SNPs, respectively, for the two designs over 100 iterations). To compute $\rho_{\max}$, we then solved the least-squares problem

$$\hat{b} \;=\; \arg\min_{b} \; \|Zb - g\|^2, \qquad \text{so that} \qquad P_{\mathcal{H}} g = Z\hat{b},$$

and obtained

$$\rho_{\max} \;=\; \frac{\|Z\hat{b}\|}{\|g\|}.$$

This allowed us to get $\rho_{\max}$ for a given interaction feature matrix Z without specifying the coefficients, hence without simulating phenotypes. Since $u = Z\beta_u$ and β_u ranges over an unbounded continuum, deriving $\rho_{\max}$ via u would otherwise require infinite simulations.

Simulation analysis. Using the same genotype dataset of 210,145 samples, we simulated phenotypes as

$$y \;=\; G\theta_g \;+\; Z\theta_z \;+\; \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I),$$

where G is the genotype matrix with 1000 SNPs on chromosome 22, Z is the two-way interaction matrix with 100 features (200 SNPs) formed with SNPs on chromosome 21, θ_g and θ_z are sampled from $\mathcal{N}(0,1)$ independently. To be able to imitate the previously defined variance fraction term λ, the terms in the simulation model were scaled so that the variance of the combined genetic and noise component would be 1. We varied the λ from 0.001 to 0.171, with 10,000 evenly spaced values, and for each value we drew a random genetic architecture as defined previously (i.e., total 10,000 simulated phenotypes). Furthermore, using these same steps, we obtained another set of 10,000 simulated phenotypes with the randomly subsampled sample size of 100,000. The upper limit for λ was obtained from Hivert et al. [7] estimation of upper limit for non-additive variance.

We then used a state-of-the-art LMM-based tool REGENIE v3.4 [20] to obtain test statistics over target SNPs, which were chosen to be on the same chromosome as the interaction SNPs while not being part of the interaction features (resulting in 3834 target SNPs for each simulated phenotype). This close proximity of target and interaction SNPs yielded multiple high-value realizations for the correlation term ρ with reasonable number of iterations. We provided age, sex and the first 5 principal components as covariates to REGENIE.

Data availability

Code for reproducing the findings is available at `https://github.com/genodeco/GWAS-Epistasis-Bias`. Access to the Estonian Biobank dataset is subject to an application procedure, which can be found at `https://genomics.ut.ee/en/content/estonian-biobank`.

Results

We first examined the distributions of ρ_{max} obtained only based on the interaction subspace and identified descriptive statistics such as the maximum $\rho_{max} = 0.849$ (when interaction SNPs and the target SNP are on the same chromosome) and $\rho_{max} = 0.042$ (when interaction SNPs and the target SNP are on different chromosomes) (Fig 1). Based on the mean, minimum and maximum values obtained from these distributions, we computed the conservativeness measure R from the expression (11), setting the t-statistic value to $x = \pm 5.45$, which corresponds to the common GWAS significance p-value = 5e-8. For ρ values from 0.021 to 0.849, R was consistently observed in the anti-conservative regime (i.e., $R > 1$) over all λ and n values, and the regime where $p_{\text{true}}(x = \pm 5.45) > 0.5$ (i.e., an observation as extreme

as $|x| = 5.45$ occurs at least 50% of the time, indicating spuriously significant hits at least $\sim 50\%$ of the time) was achieved with sufficient n sample size (Fig 2). We further examined n versus ρ plots for $\lambda = 0.055$, $\lambda = 0.119$ and $\lambda = 0.171$ (estimated mean and maximum non-additive variance fractions from [7]), limiting ρ to $\rho_{max} = 0.032$ (Fig 3). R was again mostly observed in the anti-conservative regime. Furthermore, $p_{\text{true}}(x = \pm 5.45) > 0.5$ could be attained even with low ρ and λ values, and realistic sample sizes.

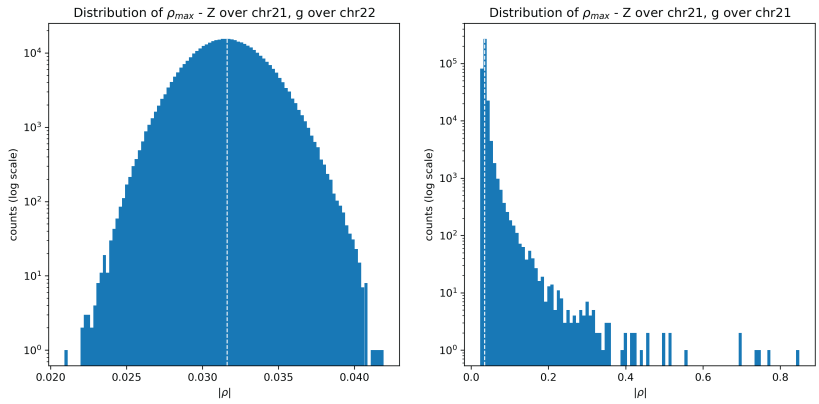

Figure 1. Distribution of ρ_{max} estimates obtained from the design where the interaction SNPs and target SNPs are on different chromosomes (left) and for the design where interaction SNPs and target SNPs are on the same chromosome (right). Vertical dashed lines indicate mean value. For the different chromosome design, min $\rho_{max} = 0.021$, mean $\rho_{max} = 0.032$, max $\rho_{max} = 0.042$. For the same chromosome design, min $\rho_{max} = 0.024$, mean $\rho_{max} = 0.035$, max $\rho_{max} = 0.849$.

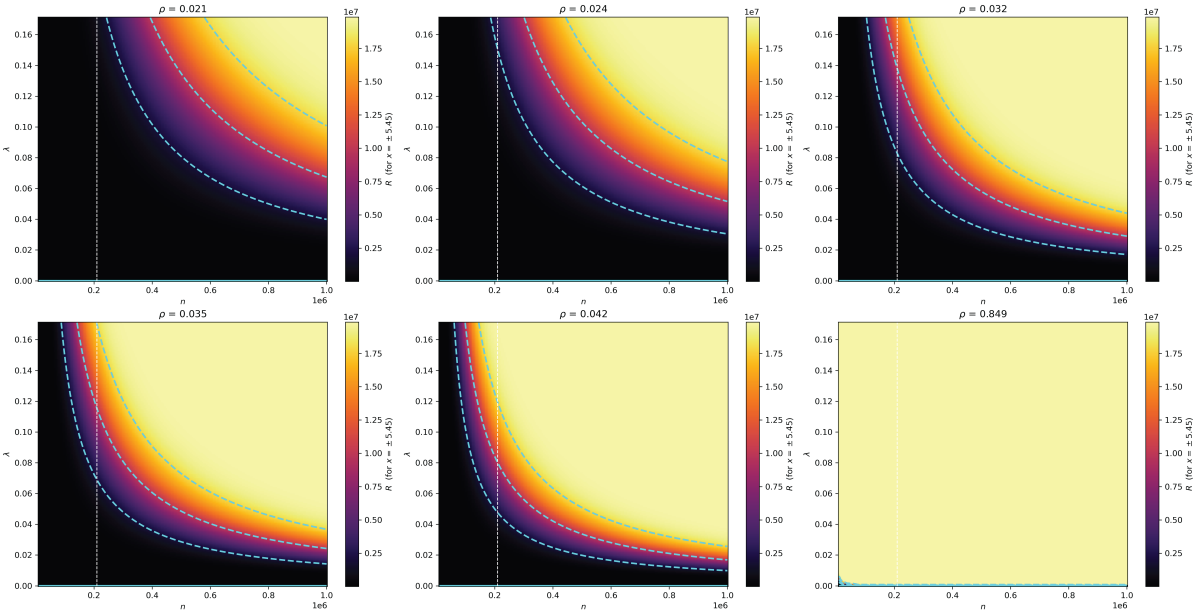

Figure 2. Heatmap plots of $R = p_{\text{true}}(x)/p_{\text{nom}}(x)$ under the mathematical model for $x = \pm 5.45$ with corresponding n (x-axis) and λ (y-axis) values, and different fixed ρ values. Cyan lines indicate $R = 1$ contour. Dashed cyan lines from center outwards indicate $p_{\text{true}}(x) = 0.1$, 0.5 and 0.9, respectively. Vertical dashed line corresponds to sample size $n = 210{,}145$.

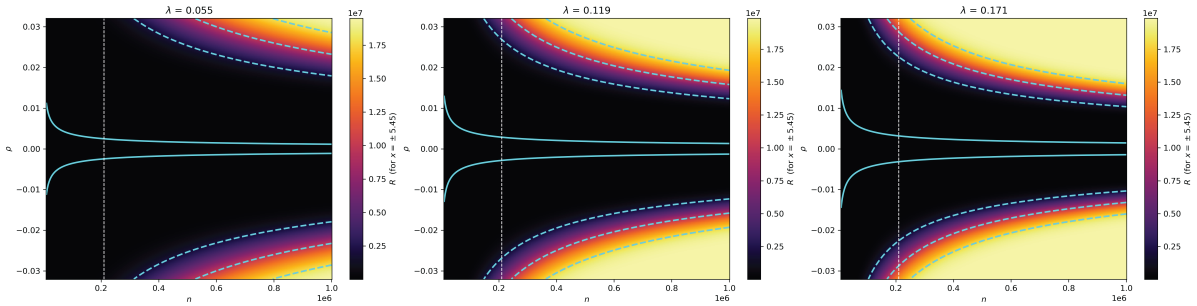

Figure 3. Heatmap plots of $R = p_{\text{true}}(x)/p_{\text{nom}}(x)$ under the mathematical model for $x = \pm 5.45$ with corresponding n (x-axis) and ρ (y-axis) values, and different fixed λ values. Cyan lines indicate $R = 1$ contour. Dashed cyan lines from center outwards indicate $p_{\text{true}}(x) = 0.1$, 0.5 and 0.9, respectively. Vertical dashed line corresponds to sample size $n = 210{,}145$.

As the next step, we compared the results from REGENIE applied to the simulated phenotypes for $n = 210{,}415$ and $n = 100{,}000$ (by randomly subsampling the main dataset) with the mathematical model of conservativeness. There was high concordance between simulation results and the algebraic model for both n values (Fig 4). The expected spuriously significant hits were clearly observed in the simulations; mean $|z|$ score was above 5.45 in the region corresponding to $p_{\text{true}}(x = \pm 5.45) > 0.5$ regime in the mathematical model (Fig 4a-b). Despite that the simulations did not produce some extreme value regions (such as high-ρ - low-λ) due to limited iterations, we were able to find clear evidence for $R > 1$ anti-conservative regime as we observed $|z| > 5.45$ for even extremely small ρ values (Fig 4c).

Discussion

In this work, we derived a mathematical model characterizing the test statistic bias in GWAS if the true genetic architecture involves epistasis which is not taken into account in the fitted linear model. The inflationary bias at the extreme tail is mainly driven by the positive shift in the mean of the null statistic, whereas the downward shift in the variance drives deflation, yet the dominant tendency for the critical t-statistic value ± 5.45 (corresponding to the common GWAS significance p-value 5e-8) is inflation (hence, increased spurious significance) under a vast range of realistic parameter values (Fig 2, 3). We were able to define the mean and variance shift in terms of three terms: (i) λ, the variance fraction of the interaction component, (ii) ρ, the correlation between the target SNP and the realized interaction term, and (iii) n, the sample size. For the range of λ, we used the estimations from literature [7]. For the range of ρ, we first derived a maximum threshold algebraically without needing the interaction term coefficients. We also performed phenotype simulations with the full interaction architecture. As expected, even under the strict no-path null, absolute value of ρ can be high due to LD between the target SNP and the interaction SNPs. Interestingly, even when the target SNP and the interaction SNPs are on different chromosomes, ρ can reach sufficient non-zero values to produce spurious significance (Fig 1, 2).

Another interesting observation was that the problem of spurious significance is positively correlated with the sample size. Average sample size for GWAS in 2022 was reported to be up to 140,000 [4]. It is now also not uncommon to utilize more than a million sample-sized datasets thanks to large biobanks [22]. As the sample size reaches a million, even minimal estimates for the interaction variance fraction λ (e.g., ~ 0.03) and correlation ρ (e.g., ~ 0.03) can result in $p_{\text{true}}(x = \pm 5.45) \gtrapprox 0.5$ regime where we expect approximately half of the statistically significant hits to be spurious (Fig 2).

One of the limitations of our model is the initial assumption of perfectly estimated V excluding u, which we mainly proposed to be able to equate true and misspecified model's phenotype vectors. This is not realistic because the structured part of the estimated $\hat{V}$ could potentially capture some variation from the interaction term. However, modern LMMs generally estimate kinship matrix K and variance of the polygenic effect σ_η^2 with the leave-one-chromosome-out (LOCO) method in which the chromosome of the target SNP is removed from the estimation to avoid proximal contamination. This means that the

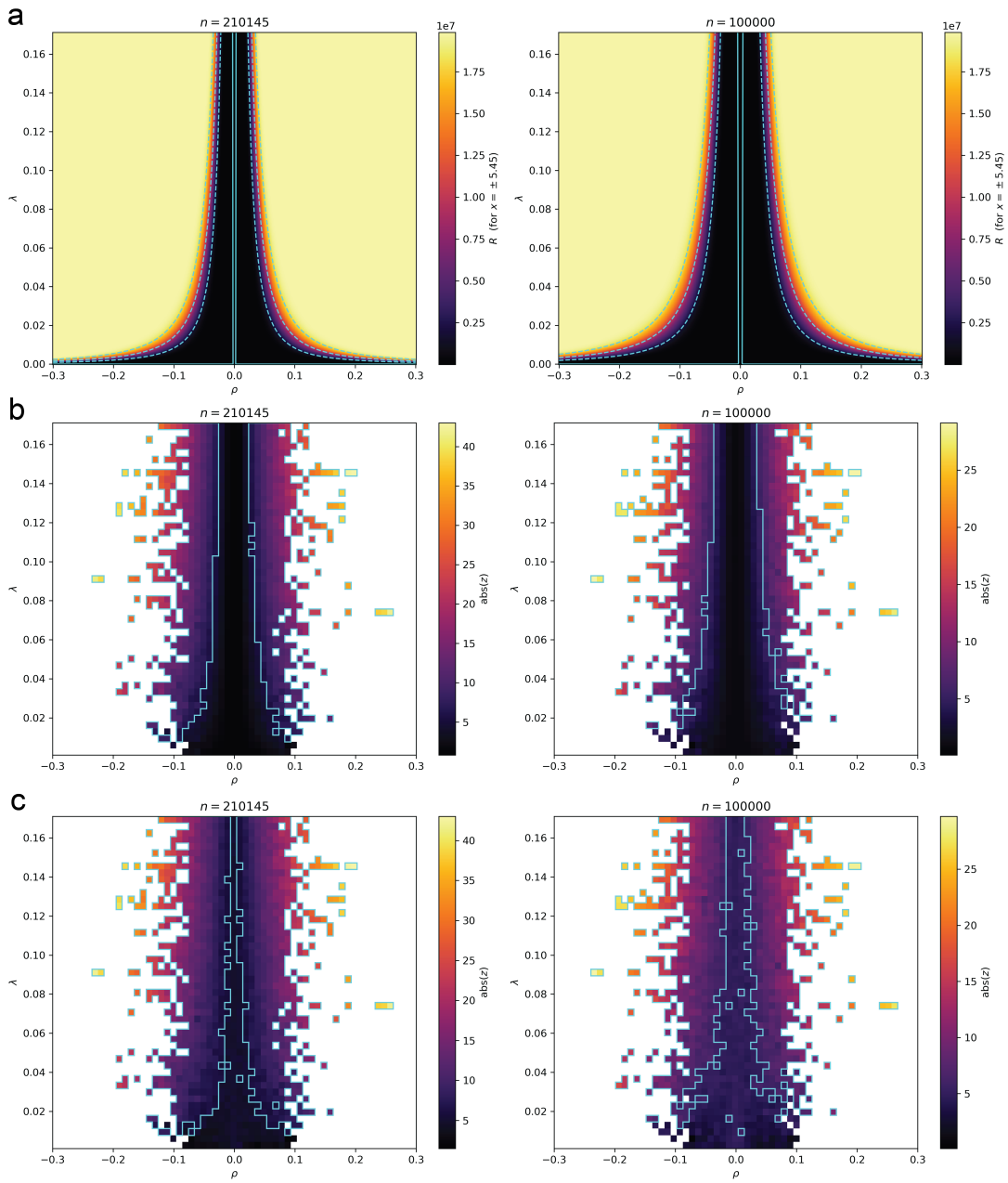

Figure 4. Comparison of mathematical model to simulated results for $n = 100,000$ (right) and $n = 210,145$ (left). **a**. $R = p_{\text{true}}(x)/p_{\text{nom}}(x)$ ratio from the mathematical model with $x = \pm 5.45$. Cyan lines indicate $R = 1$ contour. Dashed cyan lines from center outwards indicate $p_{\text{true}}(x) = 0.1$, 0.5 and 0.9, respectively. **b**. Simulation-based heatmap plot of $|z|$ scores obtained by mapping the values to a 50x50 grid and calculating the mean score for each square. White regions indicate squares where simulations did not have specific λ - ρ values (since λ could be set while simulating the phenotypes but ρ was only calculated depending on the realized interaction term). Cyan lines indicate $|z| = 5.45$ contour, corresponding to the common GWAS significance threshold. **c**. Simulation-based heatmap plot of $|z|$ scores obtained by mapping the values to a 50x50 grid and calculating the maximum score for each square.

most severe bias where the target SNP and interaction SNPs are on the same chromosome most likely persists even when the u interaction space partially aligns with K. Indeed, we were able to confirm via LMM-based GWAS on simulated phenotypes that our mathematical model is consistent (Fig 4). As a precautionary check for spurious significance without the omitted term, for each simulated phenotype, we also simulated a version for which the only difference is the subtraction of the realized interaction component from the final phenotype. We detected 4 and 2 spurious associations (i.e., p-val >5e-8) over 38,340,000 tested SNPs in 10,000 simulated phenotypes for $n = 100,000$ and $n = 210,145$ datasets, respectively. In contrast, for phenotypes with the interaction term, 1851 and 6076 spuriously significant SNPs were detected for $n = 100,000$ and $n = 210,145$ datasets, respectively.

GWAS are crucial for providing the basis of functional studies which are required to understand the causal mechanisms behind complex traits. In this work, we demonstrated how GWAS with linear models can be susceptible to producing spurious associations when the true genetic model has epistasis, especially with increasing sample sizes and when the SNPs involved in epistasis are on the same chromosome. Our findings suggest that future studies should pursue the ongoing efforts of explicit modeling of interactions or assumption-free (or minimum-assumption) models to characterize genotype-phenotype associations [23].

Contributions

Conceptualisation: BY. Data curation: BY, MNG. Methodology: BY. Investigation: BY, MNG. Formal analysis: BY, GC, FJ. Visualisation: BY. Supervision: BY, MM, GC, FJ. Original draft: BY. Review & editing: BY, TK, FJ, GC.

Acknowledgements

Thanks to the High Performance Computing Center (HPC) of the University of Tartu for providing computational resources. Thanks to the Estonian Biobank Research Team (Andres Metspalu, Lili Milani, Tõnu Esko, Reedik Mägi, Mari Nelis, Georgi Hudjashov) for the genotype and phenotype datasets used in this work.

Ethics statement

The activities of the EstBB are regulated by the Human Genes Research Act, which was adopted in 2000 specifically for the operations of the EstBB. Individual level data analysis was carried out under ethical approval 1.1-12/624 (issued 24.03.2020), 1.1-12/2618 (issued 04.08.2022) and 1.1-12/3454 (issued 20.10.2022) from the Estonian Committee on Bioethics and Human Research (Estonian Ministry of Social Affairs).

References

[1] Pankhuri Singhal, Shefali Setia Verma, and Marylyn D. Ritchie. "Gene Interactions in Human Disease Studies—Evidence Is Mounting". en. In: *Annual Review of Biomedical Data Science* 6.Volume 6, 2023 (2023). Publisher: Annual Reviews, pp. 377–395. ISSN: 2574-3414. DOI: 10.1146/annurev-biodatasci-102022-120818.

[2] Esther Herrera-Luis et al. "Gene–environment interactions in human health". en. In: *Nature Reviews Genetics* 25.11 (2024). Publisher: Nature Publishing Group, pp. 768–784. ISSN: 1471-0064. DOI: 10.1038/s41576-024-00731-z.

[3] Patrick C. Phillips. "Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems". en. In: *Nature Reviews Genetics* 9.11 (2008). Publisher: Nature Publishing Group, pp. 855–867. ISSN: 1471-0064. DOI: 10.1038/nrg2452.

[4] Abdel Abdellaoui et al. "15 years of GWAS discovery: Realizing the promise". In: *The American Journal of Human Genetics* 110.2 (2023), pp. 179–194. ISSN: 0002-9297. DOI: 10.1016/j.ajhg.2022.12.011.

[5] William G. Hill, Michael E. Goddard, and Peter M. Visscher. "Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits". en. In: *PLOS Genetics* 4.2 (2008). Publisher: Public Library of Science, e1000008. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1000008.

[6] Guosheng Su et al. "Estimating Additive and Non-Additive Genetic Variances and Predicting Genetic Merits Using Genome-Wide Dense Single Nucleotide Polymorphism Markers". en. In: *PLOS ONE* 7.9 (2012). Publisher: Public Library of Science, e45293. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0045293.

[7] Valentin Hivert et al. "Estimation of non-additive genetic variance in human complex traits from a large sample of unrelated individuals". In: *The American Journal of Human Genetics* 108.5 (2021), pp. 786–798. ISSN: 0002-9297. DOI: 10.1016/j.ajhg.2021.02.014.

[8] Gibran Hemani, Sara Knott, and Chris Haley. "An Evolutionary Perspective on Epistasis and the Missing Heritability". en. In: *PLOS Genetics* 9.2 (2013). Publisher: Public Library of Science, e1003295. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1003295.

[9] Samuel Pattillo Smith et al. "Discovering non-additive heritability using additive GWAS summary statistics". In: *eLife* 13 (2024). Ed. by George H Perry. Publisher: eLife Sciences Publications, Ltd, e90459. ISSN: 2050-084X. DOI: 10.7554/eLife.90459.

[10] Bettina Mieth et al. "DeepCOMBI: explainable artificial intelligence for the analysis and discovery in genome-wide association studies". In: *NAR Genomics and Bioinformatics* 3.3 (2021), lqab065. ISSN: 2631-9268. DOI: 10.1093/nargab/lqab065.

[11] Burak Yelmen et al. *Interpreting artificial neural networks to detect genome-wide association signals for complex traits*. arXiv:2407.18811 [cs, q-bio]. 2024.

[12] Lotfi Slim et al. "Novel methods for epistasis detection in genome-wide association studies". en. In: *PLOS ONE* 15.11 (2020). Publisher: Public Library of Science, e0242927. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0242927.

[13] Yu-Chuan Chang et al. "GenEpi: gene-based epistasis discovery using machine learning". en. In: *BMC Bioinformatics* 21.1 (2020), p. 68. ISSN: 1471-2105. DOI: 10.1186/s12859-020-3368-2.

[14] Lorin Crawford et al. "Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits". en. In: *PLOS Genetics* 13.7 (2017). Publisher: Public Library of Science, e1006869. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1006869.

[15] Jeffrey M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data, second edition*. en. Google-Books-ID: hSs3AgAAQBAJ. MIT Press, 2010. ISBN: 978-0-262-29679-3.

[16] Alexander Platt, Bjarni J. Vilhjálmsson, and Magnus Nordborg. "Conditions Under Which Genome-Wide Association Studies Will be Positively Misleading". In: *Genetics* 186.3 (2010), pp. 1045–1052. ISSN: 0016-6731. DOI: 10.1534/genetics.110.121665.

[17] Samuel P. Dickson et al. "Rare Variants Create Synthetic Genome-Wide Associations". en. In: *PLOS Biology* 8.1 (2010). Publisher: Public Library of Science, e1000294. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1000294.

[18] Hai-Jun Liu et al. "Potential synthetic associations created by epistasis". In: *Genome Biology* 26.1 (2025), p. 336. ISSN: 1474-760X. DOI: 10.1186/s13059-025-03807-z.

[19] Lili Milani et al. "The Estonian Biobank's journey from biobanking to personalized medicine". en. In: *Nature Communications* 16.1 (2025). Publisher: Nature Publishing Group, p. 3270. ISSN: 2041-1723. DOI: 10.1038/s41467-025-58465-3.

[20] Joelle Mbatchou et al. "Computationally efficient whole-genome regression for quantitative and binary traits". en. In: *Nature Genetics* 53.7 (2021). Publisher: Nature Publishing Group, pp. 1097–1103. ISSN: 1546-1718. DOI: 10.1038/s41588-021-00870-7.

[21] Ani Manichaikul et al. "Robust relationship inference in genome-wide association studies". In: *Bioinformatics* 26.22 (2010), pp. 2867–2873. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq559.

[22] Loïc Yengo et al. "A saturated map of common genetic variants associated with human height". en. In: *Nature* 610.7933 (2022). Publisher: Nature Publishing Group, pp. 704–712. ISSN: 1476-4687. DOI: 10.1038/s41586-022-05275-y.

[23] Marleen Balvert et al. "Considerations in the search for epistasis". In: *Genome Biology* 25.1 (2024), p. 296. ISSN: 1474-760X. DOI: 10.1186/s13059-024-03427-z.

Appendix

A Variance variance

We prove here that

$$\operatorname*{Var}_{\varepsilon}\left[\operatorname*{Var}_{i}\left(\varepsilon - \tfrac{g^\top \varepsilon}{g^\top g}\, g\right)\right] \;=\; \frac{2n+q^2-1}{n^2}\,.$$

By developing:

$$\operatorname*{Var}_{\varepsilon}\left[\operatorname*{Var}_{i}\left(\varepsilon - \tfrac{g^\top \varepsilon}{g^\top g}\, g\right)\right] \;=\; \mathop{\mathbb{E}}_{\varepsilon}\left[\left(\frac{1}{n}\left(\|\varepsilon\|^2 - \tfrac{(g^\top\varepsilon)^2}{g^\top g}\right) - 1\right)^2\right]$$

$$= \mathop{\mathbb{E}}_{\varepsilon}\left[\frac{1}{n^2}\,\|\varepsilon\|^4 + \frac{1}{n^2}\tfrac{(g^\top\varepsilon)^4}{(g^\top g)^2} + 1 - \frac{2}{n^2}\tfrac{\varepsilon^2(g^\top\varepsilon)^2}{g^\top g} - \frac{2}{n}\|\varepsilon\|^2 + \frac{2}{n}\tfrac{(g^\top\varepsilon)^2}{g^\top g}\right]$$

The two last terms were already computed (for the estimation of the average). For the other terms, which are moments of order 4 in ε, we will rely on the following formulas for multivariate Gaussian distributions, that is when $X \sim \mathcal{N}(0,\Sigma)$:

$$\mathbb{E}\left[X_i^2 X_j^2\right] \;=\; \sigma_{ii}\sigma_{jj} + 2\sigma_{ij}^2$$

$$\mathbb{E}\left[X_i^2 X_j X_k\right] \;=\; \sigma_{ii}\sigma_{jk} + 2\sigma_{ij}\sigma_{ik}$$

Hence, regarding the first term:

$$\mathop{\mathbb{E}}_{\varepsilon}[\varepsilon^4] = \mathop{\mathbb{E}}_{\varepsilon}\left[\left(\sum_i \varepsilon_i^2\right)\left(\sum_j \varepsilon_j^2\right)\right] = \mathop{\mathbb{E}}_{\varepsilon}\left[\sum_{i,j}\varepsilon_i^2\varepsilon_j^2\right] = \sum_{i,j}\mathop{\mathbb{E}}_{\varepsilon}\left[\varepsilon_i^2\varepsilon_j^2\right] = \sum_{i,j} M_{ii}M_{jj} + 2M_{ij}^2$$

$$= \operatorname{Tr}(M)^2 + 2\operatorname{Tr}(M^TM) = (n-q)^2 + 2(n-q)$$

using $M^TM = M$.

Regarding the second term, we use the same formula, but using $M' = gg^T$ instead of M (note that M' is symmetric and that $M'^2 = \|g\|^2 M'$):

$$\mathop{\mathbb{E}}_{\varepsilon}\left[(g^T\varepsilon)^4\right] = \mathop{\mathbb{E}}_{\varepsilon}\left[\left\|\frac{g}{\|g\|}g^T\varepsilon\right\|^4\right] = \mathop{\mathbb{E}}_{\varepsilon'\sim\mathcal{N}(0,gg^T)}[\varepsilon'^4] = \operatorname{Tr}(gg^T)^2 + 2\|g\|^2\operatorname{Tr}(gg^T) = 3\|g\|^4$$

as $\frac{g}{\|g\|}$ is a unit-norm vector and $g^T\varepsilon$ is a scalar.

Finally, regarding the third term:

$$\mathop{\mathbb{E}}_{\varepsilon}\left[\varepsilon^2(g^T\varepsilon)^2\right] = \mathop{\mathbb{E}}_{\varepsilon}\left[\operatorname{Tr}\left(\varepsilon^2\varepsilon^Tgg^T\varepsilon\right)\right] = \mathop{\mathbb{E}}_{\varepsilon}\left[\operatorname{Tr}\left(\varepsilon^2\varepsilon\varepsilon^Tgg^T\right)\right] = \operatorname{Tr}\left(\mathop{\mathbb{E}}_{\varepsilon}\left[(\varepsilon^2)\varepsilon\varepsilon^T\right]gg^T\right)$$

Then $\mathop{\mathbb{E}}_{\varepsilon}\left[(\varepsilon^2)\varepsilon\varepsilon^T\right]$ is the following matrix:

$$\left(\mathop{\mathbb{E}}_{\varepsilon}\left[(\varepsilon^2)\varepsilon\varepsilon^T\right]\right)_{jk} = \mathop{\mathbb{E}}_{\varepsilon}\left[(\sum_i \varepsilon_i^2)\varepsilon_j\varepsilon_k\right] = \sum_i \mathop{\mathbb{E}}_{\varepsilon}\left[\varepsilon_i^2\varepsilon_j\varepsilon_k\right] = \sum_i M_{ii}M_{jk} + 2M_{ij}M_{ik}$$

hence

$$\mathop{\mathbb{E}}_{\varepsilon}\left[(\varepsilon^2)\varepsilon\varepsilon^T\right] = \mathrm{Tr}(M)M + 2M^TM = (\mathrm{Tr}(M)+2)M = (n-q+2)M$$

and the third term rewrites:

$$\mathop{\mathbb{E}}_{\varepsilon}\left[\varepsilon^2(g^T\varepsilon)^2\right] = (n-q+2)\,\mathrm{Tr}(Mgg^T) = (n-q+2)\|g\|^2$$

as $Mg = g$.

Summing all terms, we have:

$$\mathop{\mathrm{Var}}_{\varepsilon}\left[\ \mathop{\mathrm{Var}}_{i}\left(\varepsilon - \tfrac{g^\top\varepsilon}{g^\top g}\,g\right)\ \right] \ = \ \frac{(n-q)^2+2(n-q)}{n^2} + \frac{3}{n^2} + 1 - \frac{2(n-q+2)}{n^2} - \frac{2}{n}(n-q) + \frac{2}{n} \ = \ \frac{2n+q^2-1}{n^2}\ .$$