5 Appendix

5.1 Testing for latent genetic interactions

To review the regression model from the Results section, suppose Y_{jk} depends on a biallelic locus with genotype X_j, an unobserved (or latent) environmental variable M_j, and a latent genotype-by-environment (GxE) interaction $X_j M_j$ for $j = 1, 2, \ldots, n$ unrelated individuals with $k = 1, 2, \ldots r$ measurable traits. The regression model is expressed as

$$Y_{jk} = \beta_k X_j + \phi_k M_j + \gamma_k X_j M_j + \epsilon_{jk}, \tag{S1}$$

The left side of the equation are the trait values which are observable random variables. The right side contains four components: the observable genotype X_j with effect size β_k; an unobservable variable M_j with effect size ϕ_k; an unobservable interaction $X_j M_j$ with effect size γ_k; and an unobservable random error ϵ_{jk} with mean zero and variance σ_k^2. Without loss of generality, we assume that M_j is mean zero with unit variance. Our inference goal is it to test whether $\gamma_k = 0$ for $k = 1, 2, \ldots, r$ without having to observe the latent environmental variable M_j.

The following sections are outlined as follows. We first show that a latent genetic interaction induces trait variance and covariance patterns under the above model assumptions. We then review the distributional theory behind the individual-level trait central cross moments. Using these results, we briefly show how latent interactive effects can be detected within a regression model framework.

5.1.1 Latent interactions induce differential variance and covariance patterns

We show in the main text that a latent interaction can be detected based on calculating the individual-specific trait variances (ITV) and covariances (ITC). To construct these quantities, let $e_{jk} = Y_{jk} - \beta_k X_j$ denote the trait residuals after removing the additive genetic effect. For simplicity, assume the effect sizes are known. For the jth individual, given the genotype X_j, the $r \times r$ individual-specific trait covariance matrix is

$$\mathbf{\Sigma}_j \mid X_j = \begin{bmatrix} \mathrm{E}\left[e_{j1}^2 \,\middle|\, X_j\right] & \mathrm{E}[e_{j1}e_{j2} \mid X_j] & \cdots & \mathrm{E}[e_{j1}e_{jr} \mid X_j] \\ \mathrm{E}[e_{j2}e_{j1} \mid X_j] & \mathrm{E}\left[e_{j2}^2 \,\middle|\, X_j\right] & \cdots & \mathrm{E}[e_{j2}e_{jr} \mid X_j] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[e_{jr}e_{j1} \mid X_j] & \mathrm{E}[e_{jr}e_{j2} \mid X_j] & \cdots & \mathrm{E}\left[e_{jr}^2 \,\middle|\, X_j\right] \end{bmatrix},$$

where the ITV are the r diagonal elements and ITC are the $s = \binom{r}{2}$ off-diagonal elements.

The presence of a latent interaction shared by multiple traits induces differential ITV and ITC patterns as a function of genotype. More specifically, given our model assumptions, the ITC between the

kth and k'th trait is

$$\begin{aligned}
\mathrm{Cov}\big[Y_{jk}, Y_{jk'} \bigm| X_j\big] &= \mathrm{E}\big[e_{jk}e_{jk'} \bigm| X_j\big] \\
&= \mathrm{E}\big[(\phi_k M_j + \gamma_k X_j M_j + \epsilon_{jk})(\phi_{k'} M_j + \gamma_{k'} X_j M_j + \epsilon_{jk'}) \bigm| X_j\big] \\
&= \mathrm{E}\big[\phi_k\phi_{k'} M_j^2 + (\phi_{k'}\gamma_k + \phi_k\gamma_{k'})X_j M_j^2 + \gamma_{k'}\gamma_k X_j^2 M_j^2 \bigm| X_j\big] \\
&+ \mathrm{E}\big[\phi_k M_j \epsilon_{jk'} + \gamma_k X_j M_j \epsilon_{jk'} + \phi_{k'} M_j \epsilon_{jk} + \gamma_{k'} X_j M_j \epsilon_{jk} + \epsilon_{jk}\epsilon_{jk'} \bigm| X_j\big] \\
&= \mathrm{E}\big[\phi_k\phi_{k'} M_j^2 + (\phi_{k'}\gamma_k + \phi_k\gamma_{k'})X_j M_j^2 + \gamma_{k'}\gamma_k X_j^2 M_j^2 \bigm| X_j\big] \\
&= \big(\phi_k\phi_{k'} + (\phi_{k'}\gamma_k + \phi_k\gamma_{k'})X_j + \gamma_{k'}\gamma_k X_j^2\big)\,\mathrm{E}\big[M_j^2 \bigm| X_j\big] \\
&= \tilde{a}_{kk'} + \tilde{b}_{kk'} X_j + \tilde{c}_{kk'} X_j^2,
\end{aligned} \tag{S2}$$

where $\tilde{a}_{kk'} = \phi_k\phi_{k'}$, $\tilde{b}_{kk'} = \phi_k\gamma_{k'} + \phi_{k'}\gamma_k$, and $\tilde{c}_{kk'} = \gamma_k\gamma_{k'}$. Note that the fourth line follows from our assumption that the random errors of each trait are independent of each other, the genotype, and the environmental variable, and so $\mathrm{E}\big[M_j\epsilon_{jk'} \bigm| X_j\big] = \mathrm{E}[M_j\epsilon_{jk} \mid X_j] = \mathrm{E}\big[\epsilon_{jk}\epsilon_{jk'} \bigm| X_j\big] = 0$. The fifth line follows from the assumption that the environmental variable M_j is mean zero with unit variance and independent of the genotype, and so $\mathrm{E}[M_j \mid X_j] = \mathrm{E}[M_j] = 0$ implying that $\mathrm{E}\Big[M_j^2 \Bigm| X_j\Big] = \mathrm{Var}[M_j \mid X_j] + \mathrm{E}[M_j \mid X_j]^2 = \mathrm{Var}[M_j \mid X_j] = \mathrm{Var}[M_j] = 1$. Following similar steps as above, the ITV is

$$\begin{aligned}
\mathrm{Var}[Y_{jk} \mid X_j] &= \mathrm{E}\big[e_{jk}^2 \bigm| X_j\big] \\
&= a_k + b_k X_j + c_k X_j^2,
\end{aligned} \tag{S3}$$

where $a_k = \phi_k^2 + \sigma_k^2$, $b_k = 2\phi_k\gamma_k$, and $c_k = \gamma_k^2$. Thus, we have shown that a latent GxE interaction will create differential trait variance and covariance patterns that depend on genotype. In particular, a latent GxE interaction in trait k ($\gamma_k \neq 0$) will induce a variance pattern that depends on genotype (Equation S3), and also induce a covariance pattern between traits k and k' when there is a shared interaction ($\gamma_{k'} \neq 0$) or a shared interacting variable ($\phi_{k'} \neq 0$; Equation S2).

Even though we limit our discussion to a single latent environmental effect and genotype, our results hold more generally under the polygenic trait model. Furthermore, while we consider a simple interaction effect, it is straightforward to show that other complex latent signals involving the genotype induce differential variance and covariance patterns. Although, the exact functional form may be more complicated than above.

5.1.2 Distribution of the cross products

Following the above discussion, we describe the distribution for the cross product of two random variables that follow a Normal distribution. We then use this result to describe the sampling variability of the cross product and squared residual terms within a regression model framework in the next section.

To simplify notation, let $Y_1 \equiv Y_{j1}$ and $Y_2 \equiv Y_{j2}$ denote the first two traits of the jth individual. Without loss of generality, suppose these traits are normally distributed with mean zero, unit variance, and correlation coefficient ρ. The cross product term is denoted by $Z = Y_1 Y_2$.

The relationship between traits can be expressed as

$$Y_2 = \rho Y_1 + \sqrt{1-\rho^2} U, \tag{S4}$$

where $U \sim \mathrm{N}(0,1)$. The cross product term is then

$$\begin{aligned} Z &= Y_1(\rho Y_1 + \sqrt{1-\rho^2} U) \\ &= \rho Y_1^2 + \sqrt{1-\rho^2} Y_1 U, \end{aligned} \tag{S5}$$

where $Y_1^2 \sim \chi_1^2$ and $Y_1 U \sim \mathrm{B}_0$ where B_0 is the modified Bessel distribution of the second kind of order zero. For perfectly correlated variables, Z is distributed as a Chi-squared distribution with one degree of freedom. Alternatively, for uncorrelated variables, Z follows a modified Bessel distribution of the second kind of order zero. See ref. [69, 70] for the distribution of the product of two normal random variables.

The first two moments are

$$\begin{aligned} \mathrm{E}[Z] &= \rho \\ \mathrm{Var}[Z] &= 1 + \rho^2, \end{aligned} \tag{S6}$$

and, more generally, for mean centered traits with variances (σ_1^2, σ_2^2), the first two moments are

$$\begin{aligned} \mathrm{E}[Z] &= \sigma_1 \sigma_2 \rho \\ \mathrm{Var}[Z] &= \sigma_1^2 \sigma_2^2 (1 + \rho^2). \end{aligned} \tag{S7}$$

We use this result in the next section to describe the heteroskedasticity in a regression model that treats the cross products or squared residuals as outcome variables.

5.1.3 Regression model for the cross products and squared residuals

Using the central moments result, we first describe the regression model for the cross product terms. Let $P = \{(1,2), (1,3), \ldots, (2,3), (2,4), \ldots, (r-1, r)\}$ denote the set of cross product pairs such that $|P| = s$. The first and second element of the qth cross product is P_{q1} and P_{q2}, respectively, and the cross product between traits is $Z_{jq}^{\mathrm{CP}} = e_{j,P_{q1}} e_{j,P_{q2}}$. The regression model is

$$\begin{aligned} Z_{jq}^{\mathrm{CP}} \mid X_j &= \mathrm{E}\big[Z_{jq}^{\mathrm{CP}} \mid X_j\big] + \epsilon_{jq} \\ Z_{jq}^{\mathrm{CP}} \mid X_j &= \tilde{a}_q + \tilde{b}_q X_j + \tilde{c}_q X_j^2 + \epsilon_{jq}, \end{aligned} \tag{S8}$$

where $\mathrm{E}\Big[Z_{jq}^{\mathrm{CP}} \;\Big|\; X_j\Big] = \mathrm{Cov}[e_{j,P_{q1}}, e_{j,P_{q2}} \mid X_j]$ is expressed in Equation S2. The results in Section 5.1.2 can be used to describe the random error in the model: The error term ϵ_{jq} is independent for $j =$

$1, 2, \ldots, n$ observations, but in general, is not normally distributed or identically distributed. Under the null hypothesis of no interactive effects, the errors are identically distributed.

We note that the above regression model differs from typical regression models in two ways. First, the random error does not follow a Normal distribution, although for typical large GWAS sample sizes, this should not impact inference. Second, under the alternative hypothesis where interactions exists, heteroskedasticity arises in the model. To see why, using the results from the previous section, the variance of the error term can be expressed as

$$\mathrm{Var}[\epsilon_{jq} \mid X_j] = \sigma^2_{j,Y_{P_{q1}}|X_j}\sigma^2_{j,Y_{P_{q2}}|X_j} + \mathrm{E}\big[Z^{\mathrm{CP}}_{jq} \mid X_j\big]^2 \tag{S9}$$

where $\sigma^2_{Y_{j,P_{q1}}|X_j} = (\phi_{P_{q1}} + \gamma_{P_{q1}}X_j)^2 + \sigma^2_{P_{q1}}$ and $\sigma^2_{Y_{j,P_{q2}}|X_j} = (\phi_{P_{q2}} + \gamma_{P_{q2}}X_j)^2 + \sigma^2_{P_{q2}}$. Under the null hypothesis, if the heteroskedasticity is uncorrelated with the explanatory variables then there is type I error rate control. Therefore, controlling for sources of variation such as population structure and nearby SNPs with strong additive effects is important to avoid an inflated type I error rate. Finally, in addition to these sources of variation, an incorrect trait scaling will likely induce heteroskedasticity and also impact type I error rate control.

We briefly state the regression model using the ITV. For the ITV, we are modeling the change in variance of trait k as a function of X_j:

$$\begin{aligned} Z^{\mathrm{SQ}}_{jk} \mid X_j &= \mathrm{E}\left[Z^{\mathrm{SQ}}_{jk} \;\middle|\; X_j\right] + \epsilon'_{jk} \\ Z^{\mathrm{SQ}}_{jk} \mid X_j &= a_k + b_k X_j + c_k X_j^2 + \epsilon'_{jk}, \end{aligned} \tag{S10}$$

where $\mathrm{Var}\left[\epsilon'_{jk} \;\middle|\; X_j\right] = 2\sigma^4_{Y_{jk}|X_j}$. The ITVs are a special case of the ITCs when $\rho = 1$.

Thus far, we assumed that the effect sizes of the additive genetic term is known to simplify the theory. However, in practice, we use the residuals so the above theory does not exactly hold: while the studentized residuals are unbiased estimates, they follow a t-distribution and so the squared residuals follow an F-distribution (similar adjustments with the cross products). This nuance did not impact any inferences in our simulation study.

There are a few important details with the above regression model approach. First, a test for differential ITV patterns is related to the Breusch-Pagan test [21]. In addition, a regression model on the correlation scale has been discussed elsewhere (see, e.g., [71]) and, more recently, is related to one studied by Lea et al. (2019) [30]. Second, the quadratic relationship between the cross products (or squared residuals) and genotypes only holds for simple interactions, and the underlying (and unknown) functional form is expected to be more complicated. Regardless, for GWAS data where interactions are difficult to detect, c_q (or c_k) is likely much smaller than b_q (or b_k) and so it is reasonable to assume that the linear term will dominate the signal compared to higher order terms.

5.2 Supplementary figures

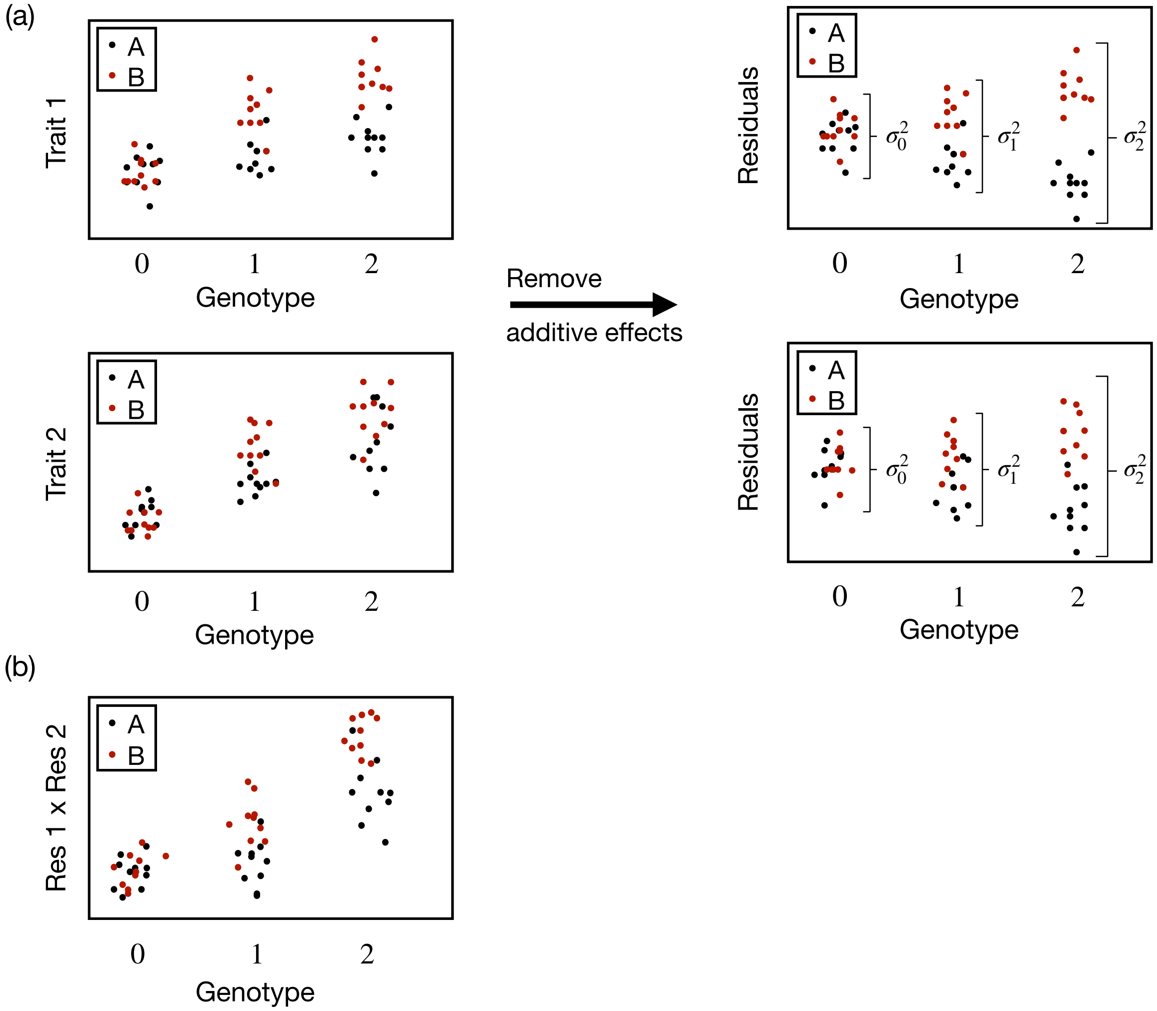

Figure S1: General strategy to detect latent genetic interactions when there are two unobserved environments denoted by ‘A’ and ‘B.’ (a) The additive genetic effect is removed and any heteroskedasticity correlated with genotype implies a latent genetic interaction. (b) When there are two traits measured, the pairwise products between the residuals (cross products) can be used to test for latent genetic effects.

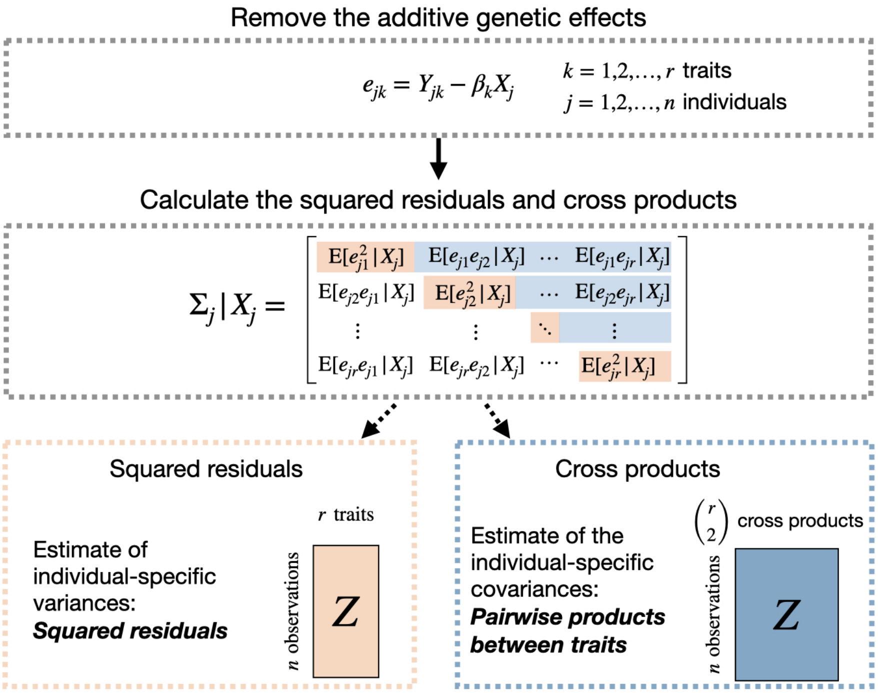
Remove the additive genetic effects
$e_{jk} = Y_{jk} - \beta_k X_j$
$k = 1,2,\ldots,r$ traits
$j = 1,2,\ldots,n$ individuals
Calculate the squared residuals and cross products
$\Sigma_j | X_j = \begin{bmatrix} E[e_{j1}^2 | X_j] & E[e_{j1}e_{j2} | X_j] & \cdots & E[e_{j1}e_{jr} | X_j] \\ E[e_{j2}e_{j1} | X_j] & E[e_{j2}^2 | X_j] & \cdots & E[e_{j2}e_{jr} | X_j] \\ \vdots & \vdots & \ddots & \vdots \\ E[e_{jr}e_{j1} | X_j] & E[e_{jr}e_{j2} | X_j] & \cdots & E[e_{jr}^2 | X_j] \end{bmatrix}$
Squared residuals
Estimate of individual-specific variances: Squared residuals
r traits
n observations
Z
Cross products
Estimate of the individual-specific covariances: Pairwise products between traits
$\binom{r}{2}$ cross products
n observations
Z

Figure S2: Revealing latent interactive effects using multiple traits. The first step is to remove the additive genetic signal to ensure that the covariance between traits is not caused by the main (additive) effects of the SNP. The individual-specific covariance matrix can then be estimated by calculating the corresponding squared residuals (estimate of the diagonal elements) and the cross products (estimate of the off-diagonal elements). These quantities can be used to infer latent interactive effects.

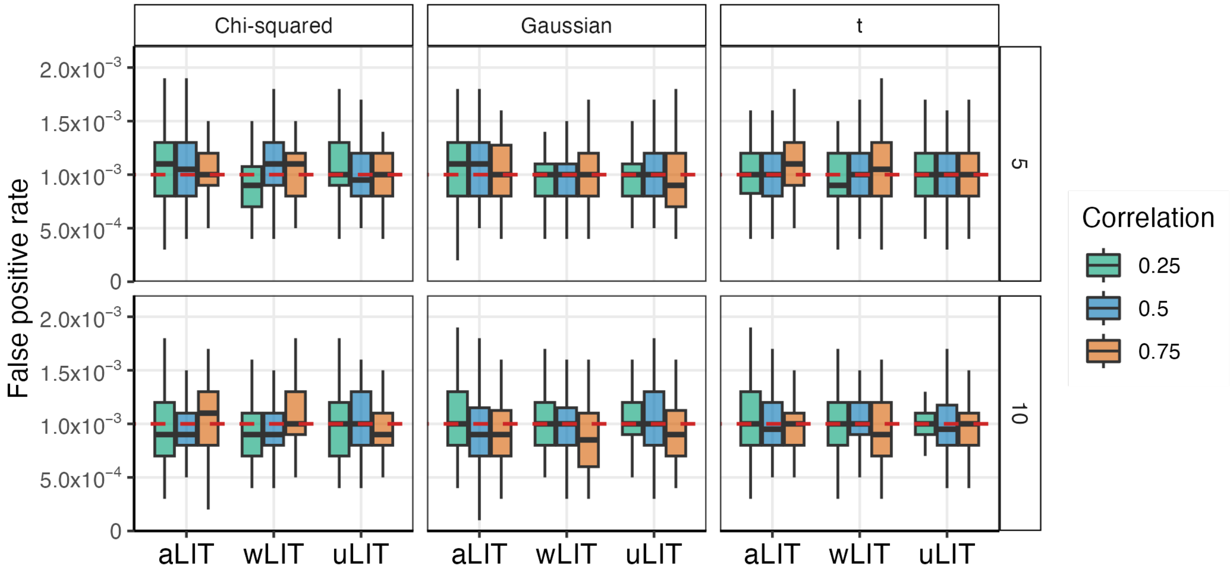

Figure S3: False positive rate of the LIT implementations under the null hypothesis of no interaction. Our simulation study varied the number of traits (rows), baseline trait correlation (0.25 (green), 0.50 (blue), and 0.75 (orange)), and error distribution (columns). For each configuration, there are 50 replicates at a sample size of $300{,}000$. The empirical false positive rate at a type I error rate of 1×10^{-3} (red dashed line).

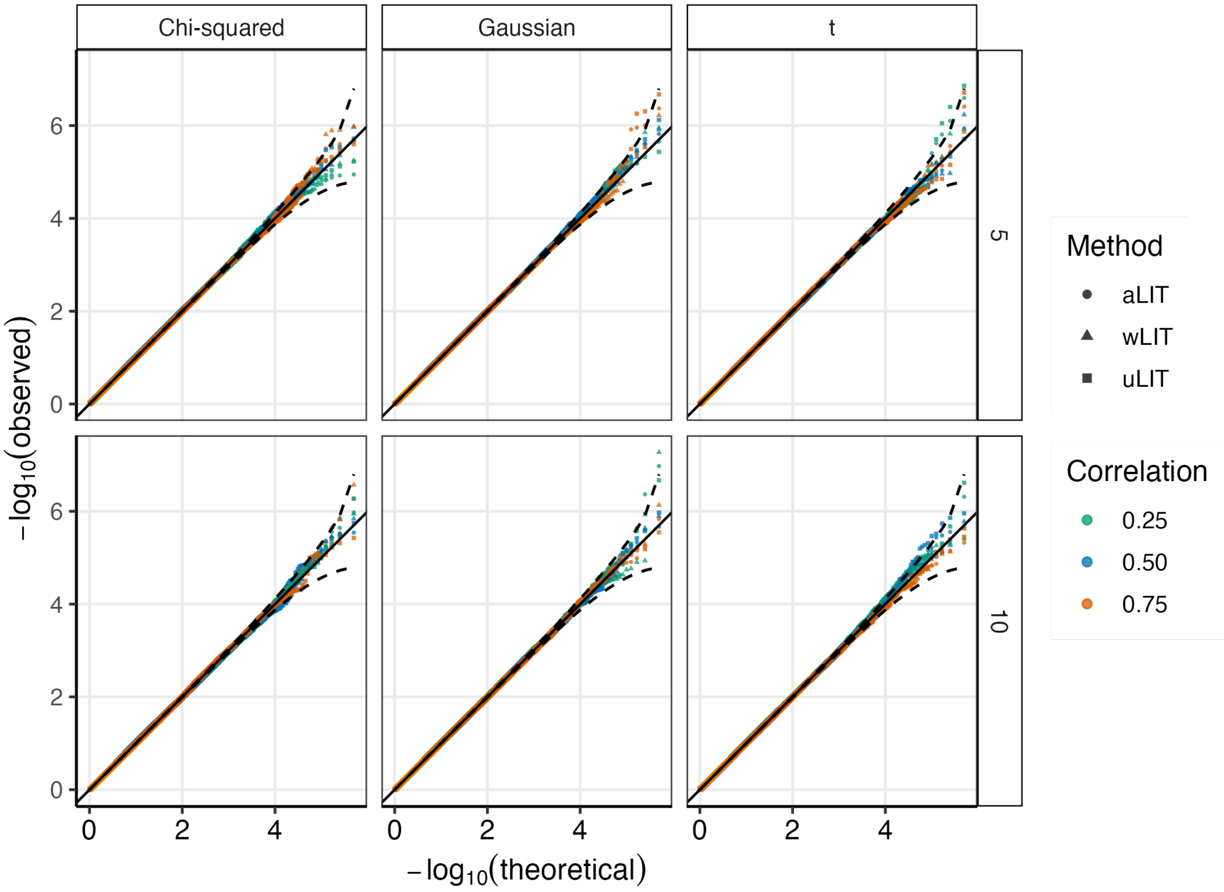

Figure S4: Q-Q plot of the LIT implementations under the null hypothesis of no interaction. Similar to Figure S3, our simulation study varied the number of traits (rows), baseline trait correlation (0.25 (green), 0.50 (blue), and 0.75 (orange)), and error distribution (columns). At each configuration, we simulated 50 datasets of $10{,}000$ SNPs and then combined the p-values for a total of $500{,}000$ p-values per configuration.

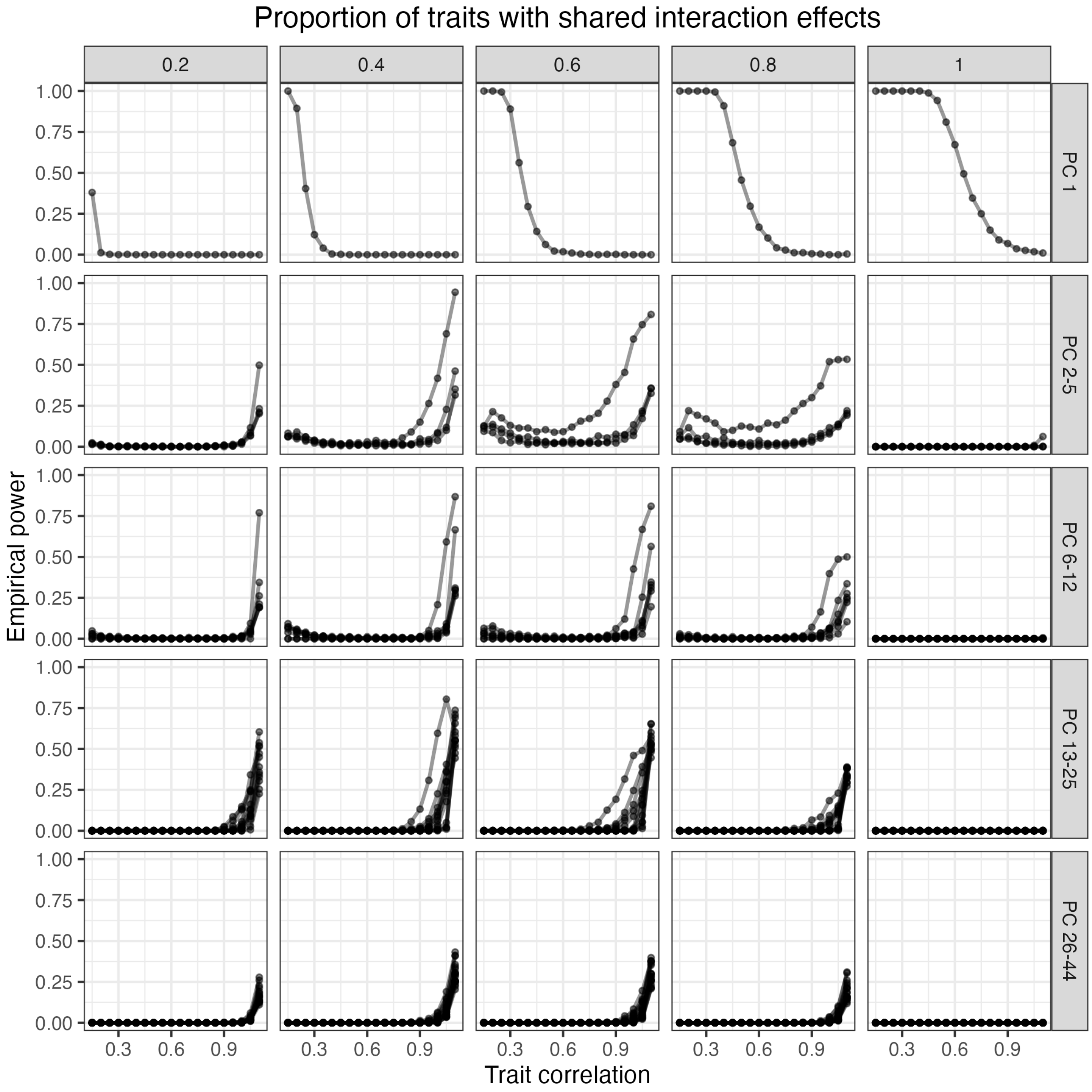

Figure S5: The empirical power of the principal components (rows) for the squared residual and cross product matrix at various baseline correlations (x-axis). In total, there was 10 traits simulated and the proportion of traits with shared interaction effects (columns) was varied. Each point represents the average power across 500 simulations at a significance threshold of 5×10^{-8}.

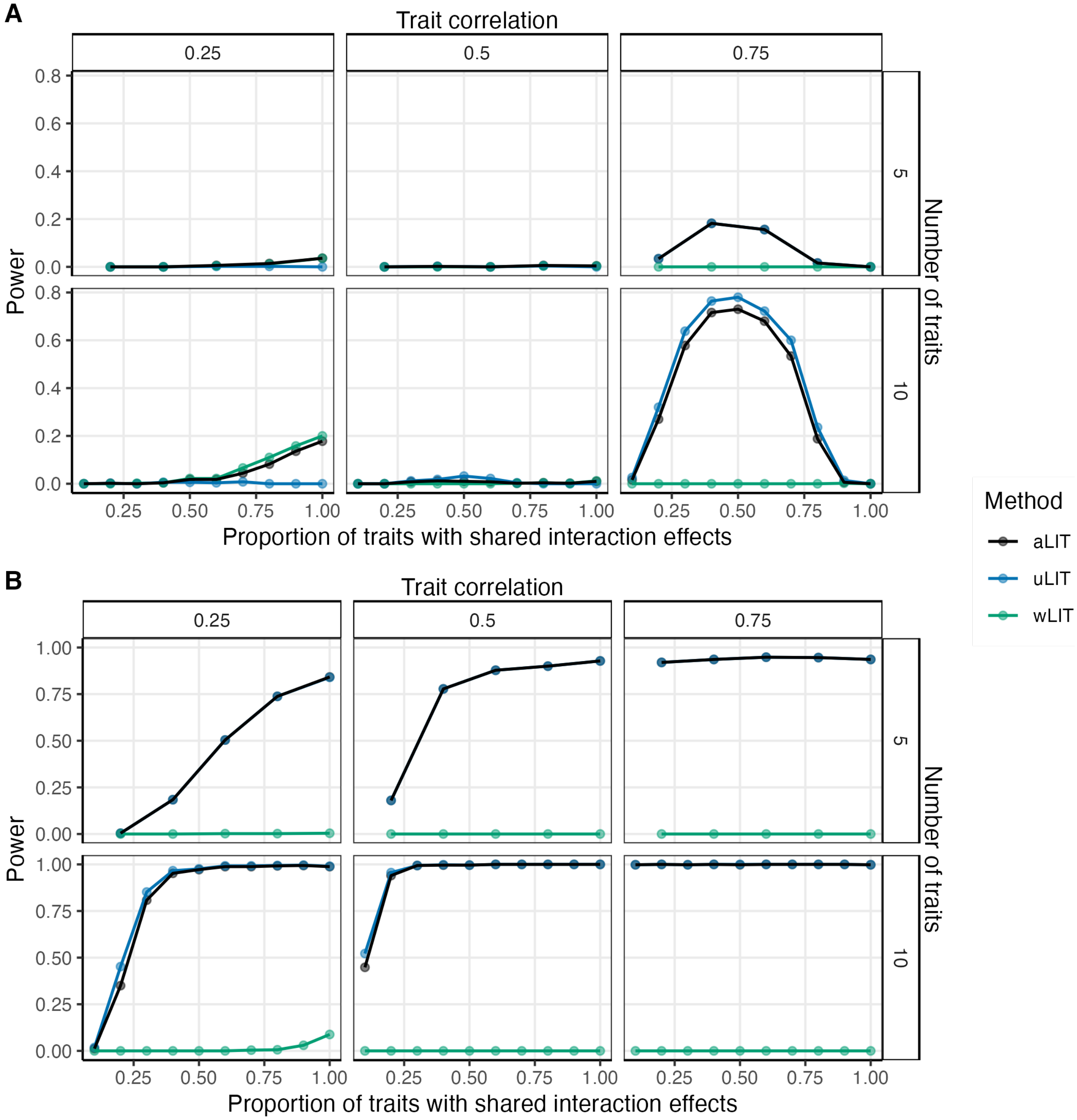

Figure S6: A similar simulation setting to Figure 2 with the direction of the effect size for the interaction term is opposite of the interacting environmental variable under (A) positive pleiotropy and (B) a mixture of positive and negative pleiotropy.

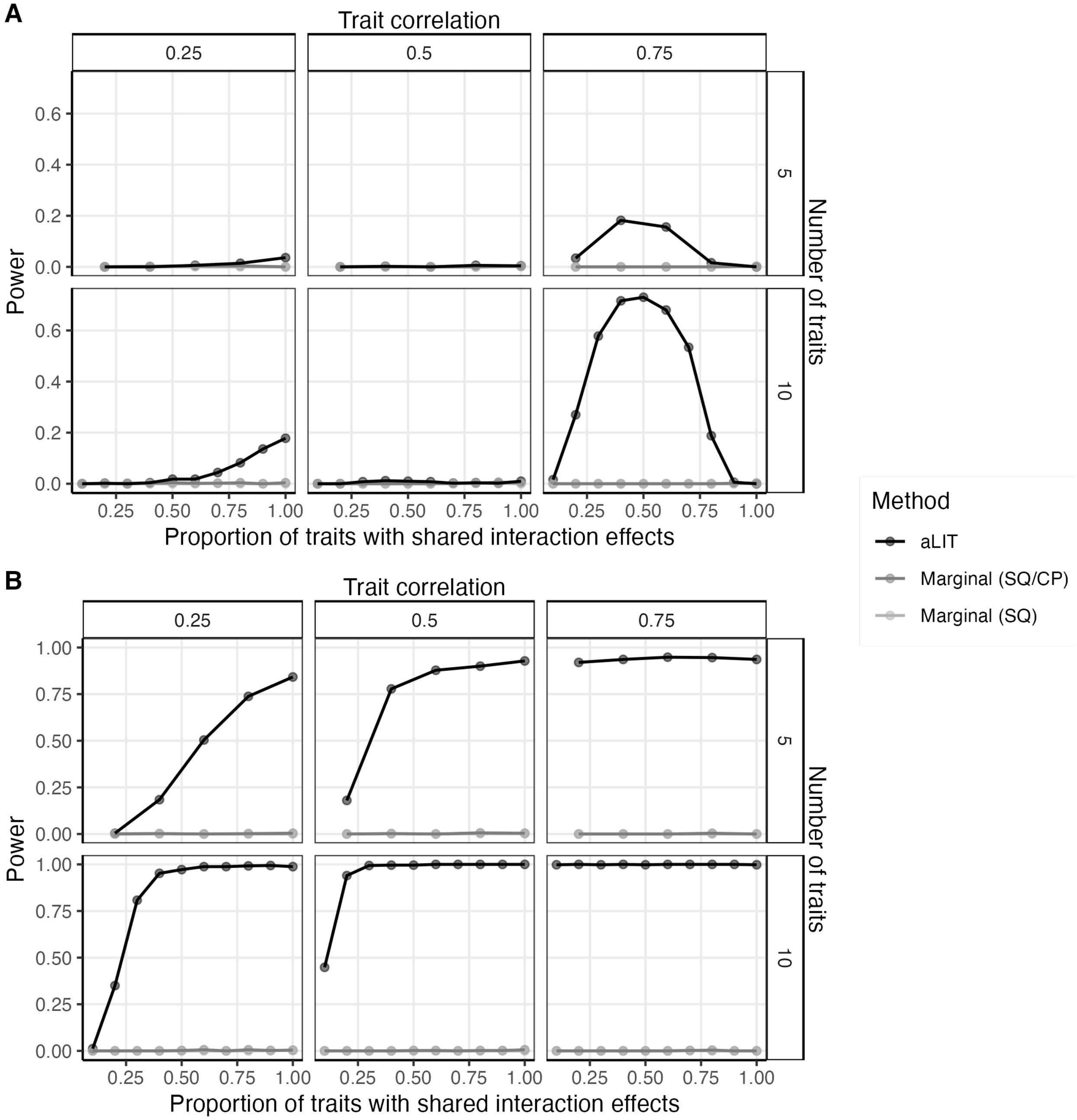

Figure S7: A similar simulation setting to Figure 3 with the direction of the effect size for the interaction term is opposite of the interacting environmental variable under (A) positive pleiotropy and (B) a mixture of positive and negative pleiotropy.

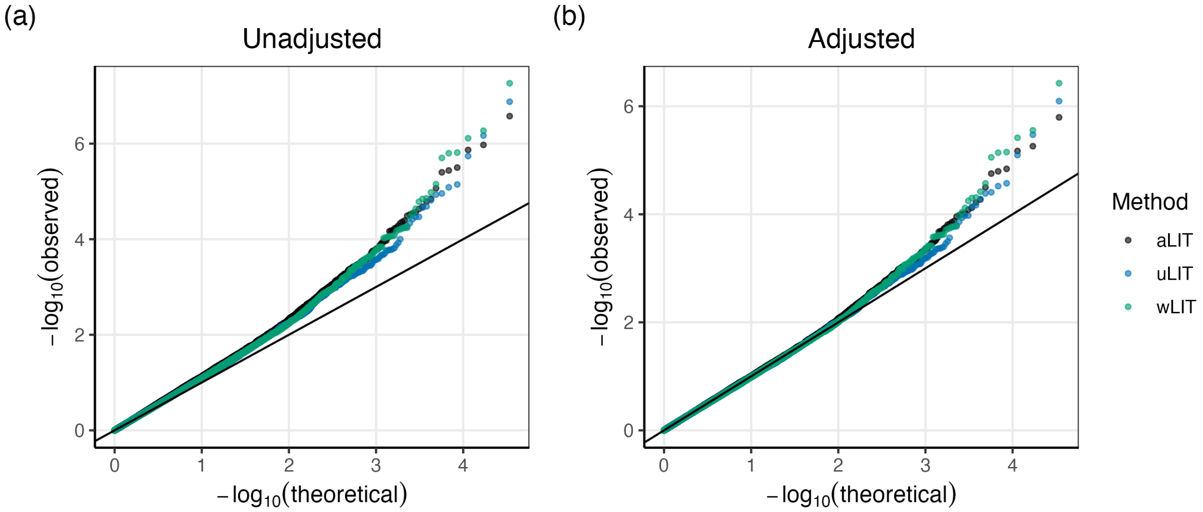

Figure S8: Quantile-Quantile plot of the uLIT, wLIT, and aLIT p-values from the UK Biobank. (a) The unadjusted p-values and (b) adjusted p-values using the genomic inflation factor. The figure removes significant p-values and those in strong linkage disequilibrium.

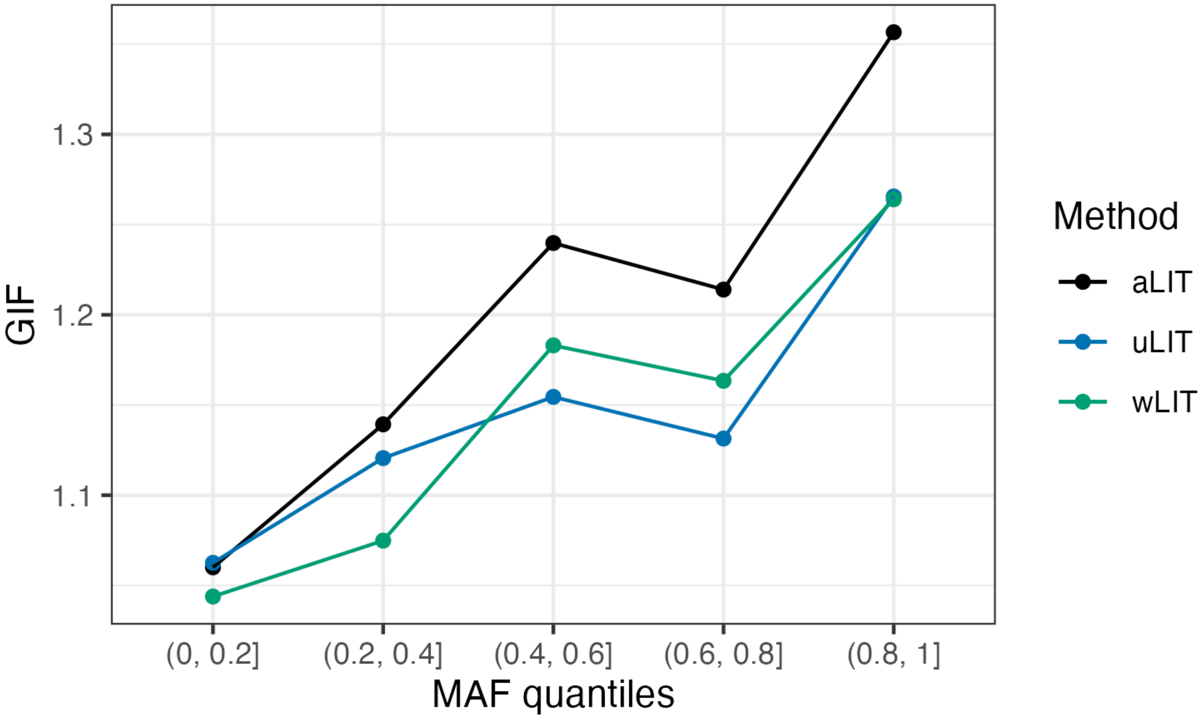

Figure S9: The genomic inflation factor from the UK Biobank analysis using uLIT, wLIT, and aLIT at different minor allele frequency quantiles.

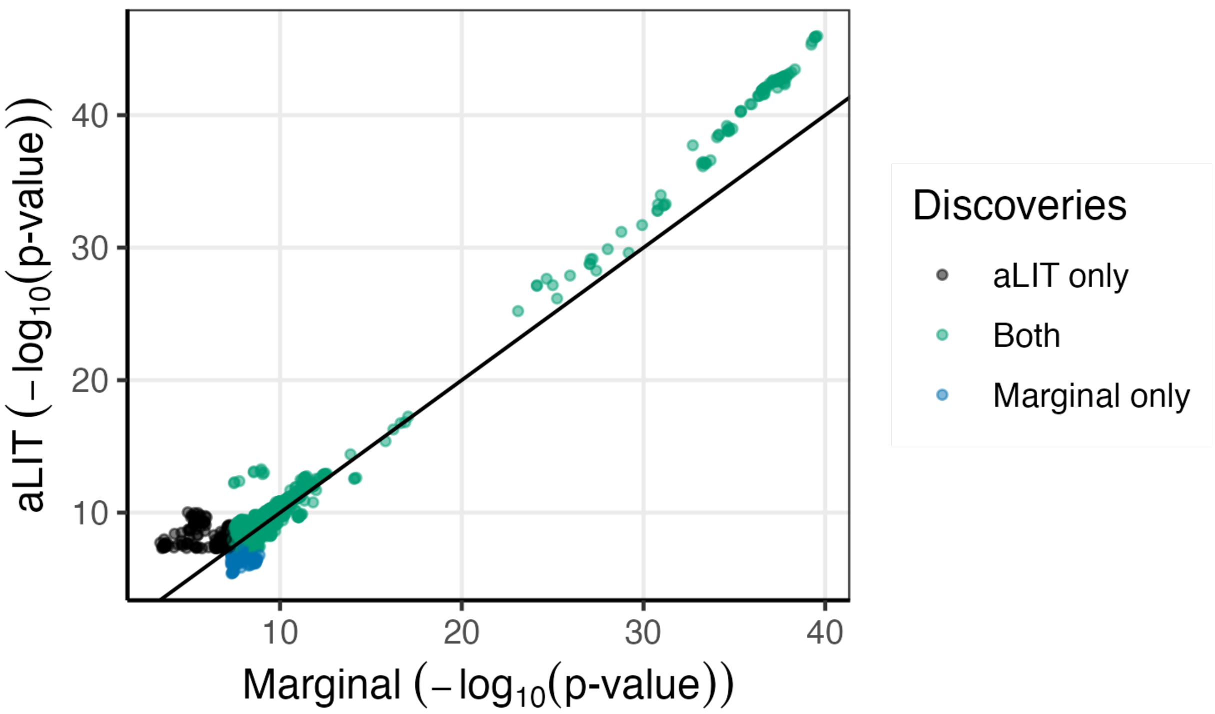

Figure S10: Comparison of the significance results using the marginal testing procedure and aLIT. The genome-wide significance threshold is 5×10^{-8}.

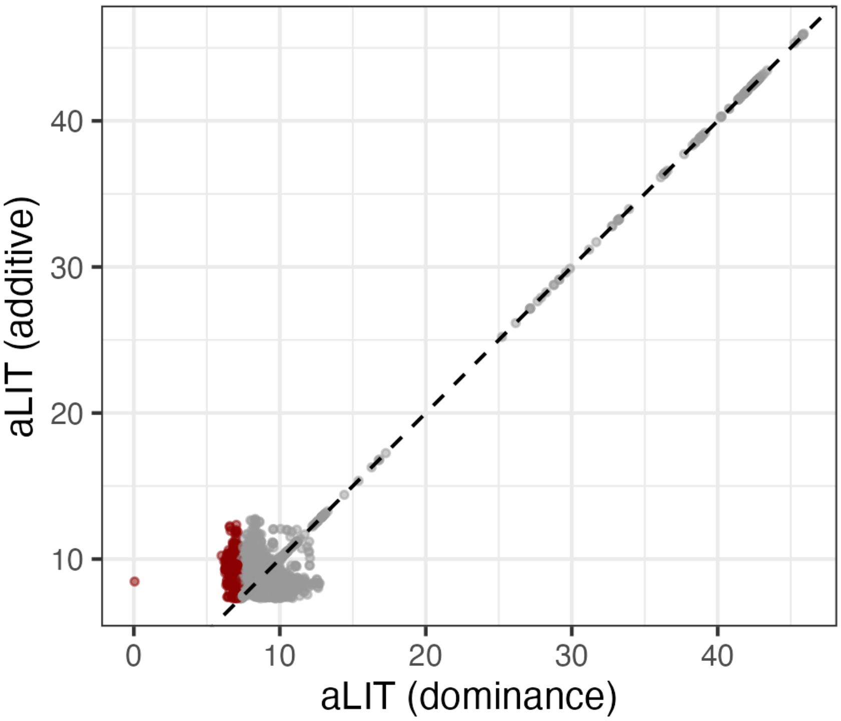

Figure S11: Comparison of aLIT p-values after adjusting for additive genetic effects (y-axis) and dominance/scaling effects (x-axis). The dark red points are SNPs that are above the genome-wide significance threshold of 5×10^{-8}. The p-values are transformed to be on a logarithmic scale similar to Figure S10.

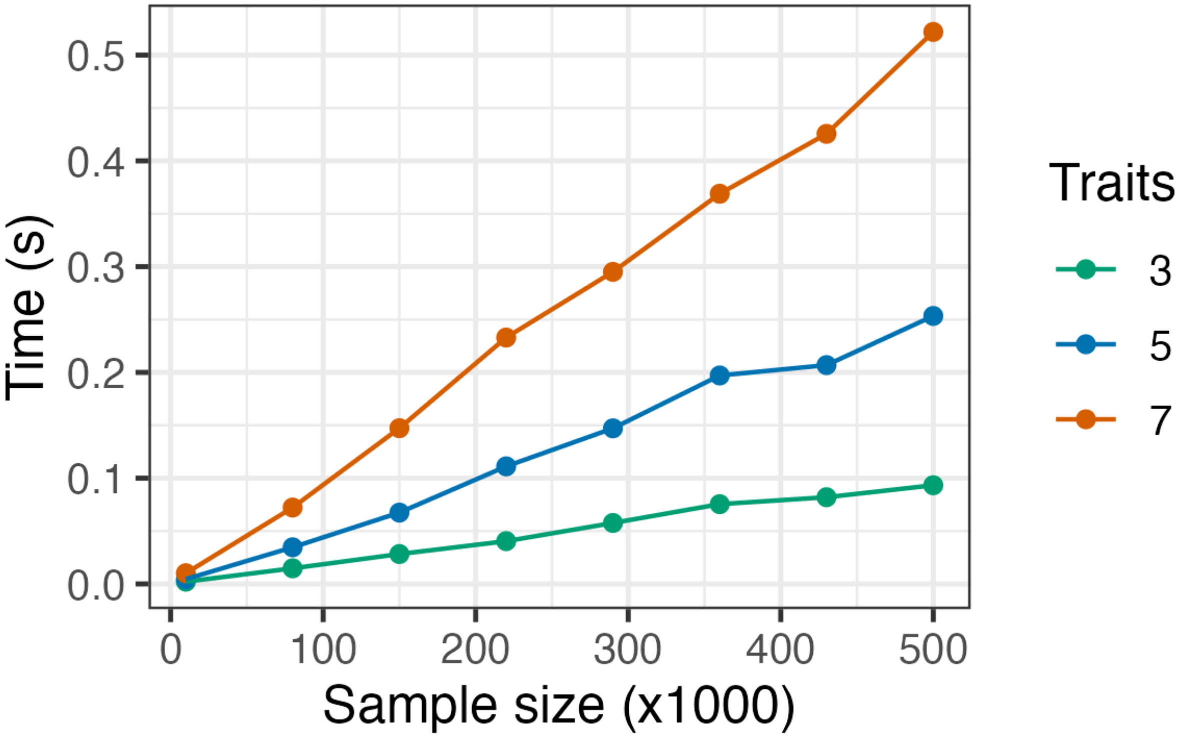

Figure S12: The average computational time to run aLIT on a SNP as a function of sample size and number of traits. Data were simulated the same way in the simulation study and each point is the average time across 500 replicates. Note that only a single core is used and that aLIT can distribute across multiple cores to substantially reduce the computational time.